EVOLUTION AND FUNCTION OF FLAVIN-BASED ELECTRON BIFURCATION

by

Saroj Poudel

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

of

Doctor of Philosophy

in

Microbiology

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2018

DEDICATION

To my parents Mr. Lok Nath Sharma and Mrs. Bishnu Kumari Sharma for their unconditional love, support and encouragement and who have been my inspiration throughout my life.

ACKNOWLEDGEMENTS

Firstly, I would like to extend my sincere gratitude to my advisor Dr. Eric Boyd for all his time, patience and immense knowledge in supporting and guiding me through my Ph.D. studies. I consider myself very lucky and privileged to have worked and learned under his guidance. I also would like to express my thanks to my co-advisor Dr. John Peters for supporting and guiding me throughout the project. I am also lucky to have fun, smart and outgoing labmates: Dan, John, Max, Melody, Eric D., Erik A., Libby and past members who always supported and encouraged me in and outside of the project. Thanks to the various co-authors who contributed to the chapters of my dissertation and to my other committee members, Dr. Ross Carlson and Dr. Seth Walk, who helped me grow as a scientist. I also owe big thanks to Dr. Sunil Ojha who introduced me to Dr. Boyd's research and for his constant support and guidance throughout my study. I would also like to thank to all the former and current staff of the Microbiology and Immunology Department for helping me keep on track with my progress and any assistance I required. I am also thankful to all colleagues in the Microbiology department and my former labmates in the algorithm lab of the Computer Science department—Sean, Alan, and Dan—for their support during my study in the University. I also owe big thanks to Dr. Olga K. Kamneva, who has been an inspiration and role model to me and has always supported and guided me since my undergraduate days. I am also grateful to the small Nepalese community at MSU for their constant support and encouragement. Finally, thank you to my amazing parents, without whom I wouldn't even have come to the USA, let alone earn an opportunity to continue my education here.

TABLE OF CONTENTS

TABLE OF CONTENTS CONTINUED

TABLE OF CONTENTS CONTINUED

TABLE OF CONTENTS CONTINUED

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

ABSTRACT

Anaerobic microorganisms live in energy limited environments with low nutrient fluxes. Thus, selection has likely acted on these cells to innovate mechanisms that improve the efficiency of anaerobic energy metabolism. In 2008, the process of flavin-based electron bifurcation (FBEB) was discovered and has since been shown to be a critical process that allows anaerobic cells to overcome thermodynamic barriers and to improve metabolic efficiency. FBEB enzymes catalyze the coupling of exergonic and endergonic oxidation–reduction reactions with the same electron donor to circumvent thermodynamic barriers and minimize free energy loss. To date, a total of 12 FBEB enzymes have been discovered that share common features that include the presence of protein-bound flavin, the proposed site of bifurcation, and the electron carrier ferredoxin. Due to its recent discovery, a comprehensive description of the natural history of bifurcating enzymes is lacking. In this thesis, we report the taxonomic and ecological distribution, functional diversity, and evolutionary history of bifurcating enzyme homologs in available complete genomes and environmental metagenomes. Moreover, we investigated the functional and ecological constraints that led to the emergence of FBEB enzymes. Bioinformatics analyses revealed that FBEB enzyme homologs were primarily detected in the genomes of anaerobes, including those of sulfate-reducers, acetogens, fermenters, and methanogens. Phylogenetic analyses of these enzyme homologs suggest that they were not a property of the Last Universal Common Ancestor of Archaea and Bacteria indicating that they are a more recent evolutionary innovation. Consistent with the role of these enzymes in the energy metabolism of anaerobes, FBEB homologs were enriched in metagenomes from subsurface environments relative to those from surface environments. In fact, the earliest evolving homologs of most bifurcating enzymes were detected in subsurface environments, including fluids from subsurface rock fractures and hydrothermal systems. Together, these data highlight the central role that FBEB played and continued to play in the energy metabolism of anaerobic microbial cells inhabiting subsurface environments.

CHAPTER ONE

GENERAL INTRODUCTION

Life requires energy and acquires it from environmental sources that include light (e.g., photosynthesis) and/or chemicals (e.g., chemosynthesis) and conserves it in the form of electrochemical gradients or chemical bonds (e.g. ATP) using one of several mechanisms. Organisms that use oxygen (i.e., aerobes) rely on distinct energy conserving mechanisms compared to the organisms that do not use oxygen (i.e., anaerobes) (Lipmann, 1941; Keltch et al, 1950; Mitchell, 1961). While aerobes inhabit highly oxidized and energy rich environments, anaerobes inhabit highly reduced and energy limited environments that include hydrothermal vents and the deep subsurface that likely exhibit characteristics of early Earth (Martin, 2012; Martin et al., 2017, Colman et al., 2017). Hence, the energy metabolism of anaerobes has been extensively studied since mid-1900s (Thauer et al., 1977) to understand how life may have survived in low energy environments thought to be reminiscent of those present on early Earth. Of particular interest to studies of early life are autotrophic organisms that are dependent on hydrogen as an electron donor (e.g., methanogens, acetogens) and which are widely believed to have retained their primitive physiology (Martin, 2012; Martin et al., 2017). Interestingly, it was not known how these organism's energy metabolisms fully worked until 2008 when flavin-based electron bifurcation was discovered. Studies since then have shown that flavin-based electron bifurcation is a key process that methanogens, acetogens, and other anaerobic organisms rely on to carry out various anaerobic metabolisms.

Electron bifurcation (EB) couples exergonic (i.e. releasing energy ($\Delta G < 0$)) and endergonic (i.e. requiring energy ($\Delta G > 0$)) reactions such that thermodynamically unfavorable reactions can be driven by thermodynamically favorable reactions (**Fig. 1**) (Herrmann *et al*, 2008; Li *et al*, 2008). Typically, the thermodynamic barriers presented by endergonic reactions are overcome with hydrolysis of a high-energy bond such as the phosphoester in ATP. However, EB does not rely on hydrolysis of ATP or other high-energy bond molecules (e.g., acetyl-CoA), but rather takes advantage of the reaction with significant negative free energy change (i.e., exergonic reaction) to drive the reaction with significant positive free energy change (i.e., endergonic reaction) thereby minimizing the energy loss in the form of heat. EB in association with other canonical energy conserving mechanisms [i.e., substrate level phosphorylation (SLP) and oxidative phosphorylation] has been found to generate more ATP compared to SLP or oxidative phosphorylation alone (Schuchmann and Müller, 2012; Schut and Adams, 2009).

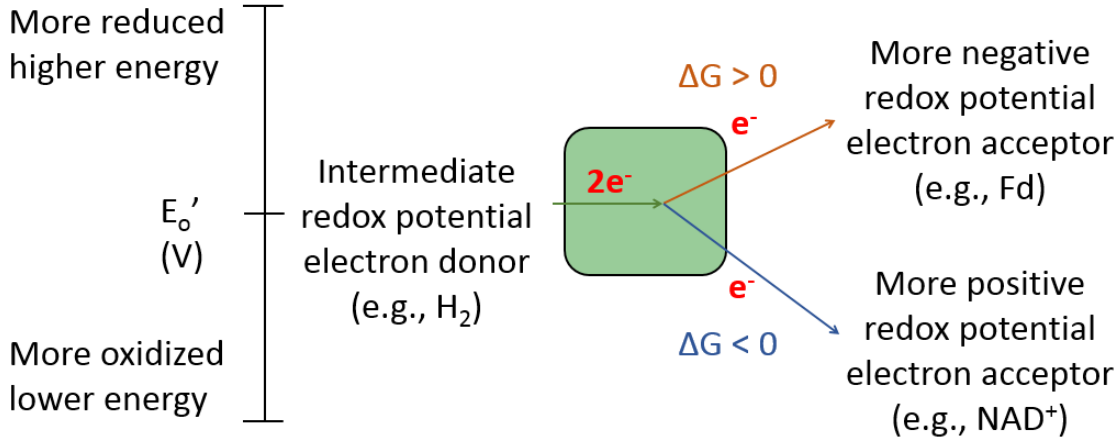


Fig. 1. Proposed mechanism of electron bifurcation from a single intermediate potential electron donor to two acceptors that have different redox potentials.

For example, glucose oxidation in *T. maritima* produces a total of three ATP along with a mixture of products (e.g., ethanol and acetate) when only SLP is operating (Schut and Adams, 2009). Likewise, in *Pyrococcus furiosus*, SLP and oxidative phosphorylation can generate 3.2 mol of ATP with only acetate as its product (Sapra *et al*, 2003). However, glucose oxidation to only acetate yields a net of four ATP when both SLP and EB are operational in *T. maritima* (Schut and Adams, 2009). The authors hypothesized that by-products, such as CO_2, that are produced during the reaction re-enter the acetyl-CoA cycle, thereby completely oxidizing the glucose to acetate and resulting in increased ATP production. The bifurcating enzyme ([FeFe]-hydrogenase in this case) present in the cell takes the electron or hydride from the reduced Fd (Fd^-) and NADH, respectively, that is generated during the reaction to yield H_2. This frees Fd and NAD^+ as electron/hydride acceptors that allows the reaction to further continue and for the oxidation of the byproduct of glucose oxidation (i.e., acetyl-coA) to acetate which generates more ATP (Schut and Adams, 2009).

Historical Perspectives on EB

EB was first observed in the Q-cycle, which occurs in aerobic respiratory chains, where coenzyme Q: cytochrome c oxidoreductase (Complex III) catalyzes the EB reaction (Mitchell, 1975). This mechanism uses quinone to bifurcate electrons and hence is termed quinone-based EB. Quinone-based EB has only been observed in aerobes (Dick and Shock, 2011) and therefore it occurs at high potentials where all the cofactors

involved in electron transfer generally have more positive potentials ranging from -300 mV to +300 mV (Zhang *et al*, 2008).

In contrast, flavin-based EB mechanism uses either flavin adenine dinucleotide (FAD) or flavin mononucleotide (FMN) as the proposed site of bifurcation (Buckel and Thauer, 2013; Peters et al., 2016; Peters et al., 2018). This new mechanism is therefore called flavin-based electron bifurcation (FBEB). Besides using flavin as their bifurcating site, all the FBEB enzymes share several common characteristics: they are cytoplasmically located, typically contain several iron-sulfur (FeS) clusters predicted to play roles in electron transfer, and require at least two redox partners in addition to Fd (e.g. NADH or NADPH) (Wang *et al*, 2013b; Yan *et al*, 2017). Compared to quinone-based EB, FBEB has primarily been identified in anaerobes and has been shown to occur at more negative potentials ranging from -300 mV to -800 mV [**Fig. 2**; (Peters *et al*, 2016)]. This indicates that while quinone-based EB might be important to aerobes, FBEB appears to be crucial for anaerobes that live in a highly reduced environment.

To date, a total of twelve bifurcating enzymes have been discovered with roles in a variety of metabolic processes (Buckel and Thauer, 2013; Peters *et al*, 2016). These include hydrogen metabolism (e.g., [FeFe]-hydrogenase (Schuchmann and Müller, 2012; Schut and Adams, 2009) and [NiFe]-hydrogenase (Kaster *et al*, 2011)), carbon metabolism (e.g., butyryl-CoA dehydrogenase (Herrmann *et al*, 2008), lactate dehydrogenase (Weghoff *et al*, 2015)), and nitrogen fixation (e.g., Fix (Ledbetter *et al*, 2017)). Homologs of these enzymes have been detected in both Archaea and Bacteria but have not been discovered in Eukarya. Among Archaea and Bacteria, the diversity of

bifurcating enzymes is not well understood. Likewise, the evolutionary history of these enzymes is unknown, despite the fundamental role that they have in the metabolisms of deeply rooted, anaerobic organisms.

Proposed Mechanism of FBEB

There exists solved crystal structures of four unique bifurcating enzymes: NADH-dependent reduced ferredoxin:$NADP^+$ (Nfn) from *Thermotoga maritima* (Demmer et. al., 2015) and *Pyrococcus furiosus* (Lubner et. al., 2017), butyryl-coenzyme A dehydrogenase/electron transferring flavoprotein (Bcd-Etf) from *Clostridium difficile* (Demmer et. al., 2017), heterodisulfide reductase-linked [NiFe]-hydrogenase (Mvh) from *Methanothermococcus thermolithotrophicus* (Wagner et. al., 2017), and caffeyl-coA reductase (Car) from *Acetobacterium woodii* (Demmer et. al., 2018). These crystal structures permit a more holistic understanding of enzyme structure function relationships, as they pertain to the mechanism of FBEB.

Biochemical analysis of Nfn

Nfn is a dimeric enzyme that couples the oxidation of NADPH with the reduction of oxidized Fd (Fd^+) and NAD^+ (Demmer *et al*, 2015; Lubner *et al*, 2017). It comprises two subunits: a small subunit (Nfn-S), containing one flavin adenine dinucleotide (FAD) and one [2Fe-2S] cluster, and a large subunit (Nfn-L) that contains one FAD and two [4Fe-4S] clusters (Berry *et al*, 2017; Demmer *et al*, 2015; Lubner *et al*, 2017). Three isoforms of Nfn have been identified: NfnI, NfnII and NfnIII (Nguyen *et al*, 2017). Of the three isoforms, the crystal structures of only NfnI from *Thermotoga maritima* and

Pyrococcus furiosus have been solved (Demmer *et al*, 2015; Lubner *et al*, 2017). A recent structural study (Lubner *et al*, 2017) conducted on NfnI from *P. furiosus* showed that flavin (i.e., L-FAD in NfnI) is the bifurcation site (Lubner *et al*, 2017) and that electrons first travel in the thermodynamically favorable direction (i.e., along the exergonic arm) and then the thermodynamically unfavorable direction (i.e. along the endergonic arm).

This study further provided new mechanistic insights into FBEB. Therein, it was shown that the doubly reduced L-FAD (i.e., HQ) from NADPH is oxidized twice. First, it is oxidized to reduce NAD^+ (which is an exergonic reaction), creating an unstable flavin semiquinone (SQ) that can easily reduce low potential Fd^+ (which is an endergonic reaction) with the help of FeS clusters present in both subunits. In other words, the thermodynamically favorable single electron oxidation of flavin hydroquinone (HQ) provides enough energy to reduce low potential Fd.

This is in stark contrast to what we know of how flavin is reduced sequentially in other instances. Typically, flavin undergoes sequential electron reductions such that the flavin with two electrons is in its most reduced state (Deistung and Thorneley, 1986). For example, in *Azotobacter chroococcum* the redox potential of flavin when it is reduced by one electron is -110 mV (i.e., $E^{0'}_{Ox/SQ} = -110$ mV), and when further reduced by a second electron it is -520 mV ($E^{0'}_{SQ/HQ} = -520$ mV). Unlike in sequential reductions, the redox potentials of the flavins in bifurcating enzymes are crossed (i.e. $E^{0'}_{SQ/HQ} > E^{0'}_{Ox/SQ}$) (**Fig. 2**). Therefore, it is much easier to reduce substrates with the second electron than the first electron. Hence, this mechanism allows for the generation of much more reduced species

such as Fd than is otherwise thermodynamically favorable. Interestingly, this same crossed potential can also be observed in quinone-based EB, albeit occurring in a much more positive redox potential range (**Fig. 2**).

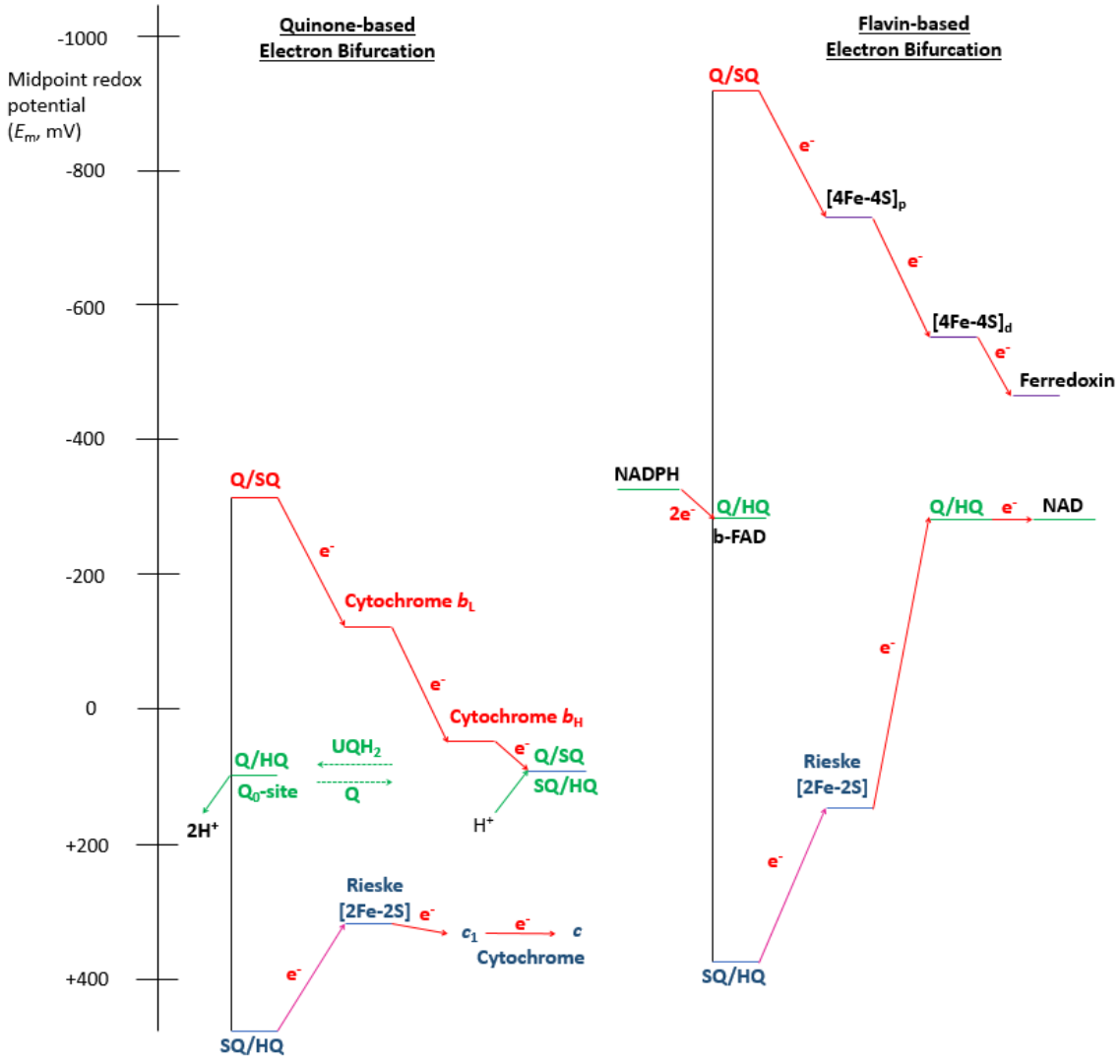


Fig. 2. Electron bifurcation schemes for quinone-based bifurcation in *Rhodobacter capsulatus* and flavin-based bifurcation of Nfn in *Pyrococcus furiosus*. Adapted from (Buckel and Thauer, 2018b).

Biochemical analysis with Bcd-Etf

Bcd-Etf was the first bifurcating enzyme that was shown to carry out FBEB (Hermann et. al., 2008). Bcd-Etf belongs to a larger group of the electron transfer flavoprotein family (Etf), many of which do not have the capability to bifurcate electrons (Garcia Costas *et al*, 2017). Bcd-Etf couples the oxidation of NADH to the simultaneous reduction of Fd^+ and crotonyl-CoA during heterotrophic growth (Chowdhury *et al*, 2014). Biochemical and structural analyses of the Bcd-Etf complexes in *A. fermentans* and *Clostridium difficle* revealed that this complex is comprised of three unique subunits: EtfA, EtfB, and Bcd (Chowdhury *et al*, 2014; Demmer *et al*, 2017). All three subunits contain one flavin adenine dinucleotide (FAD) binding site, and the FAD in EtfB is proposed to be the bifurcating site.

A recent structural study of Bcd-Etf in *Clostridium difficile* revealed that the redox potentials of flavins undergoing double oxidation were indeed crossed, allowing the favorable transfer of a second electron to much more reduced species such as Fd^+ (Demmer *et al*, 2017). It was further shown that the binding of its redox partners (i.e., NADH, Fd) is under allosteric regulation wherein each cause conformational changes upon binding. In fact, these structural changes allow the cofactors (i.e., FeS clusters) involved in electron transfer to come closer together, thereby allowing FBEB.

Both of the studies conducted on Nfn and Etf-Bcd strongly suggest that crossed redox potentials could be one of the unique features of bifurcating enzymes. In addition, it appears that allosteric regulation might play an important role in coordinating the binding of the substrates and influencing their redox potentials, as reported for Nfn

(Berry *et al*, 2017). Therein, specific residues were identified that formed an allosteric communication pathway that coordinated the binding of the three substrates involved in FBEB. Interestingly, a recent theoretical study suggests that although crossed potentials offer obvious advantages in these reactions, they are not required to bifurcate electrons based on thermodynamic or kinetic considerations (Zhang *et al*, 2017). Therefore, crossed potentials may not be observed in all bifurcating reactions and not required for FBEB, requiring further studies to have a detailed understanding of how FBEB functions.

Our understanding of FBEB has increased since its discovery in 2008, but the general mechanism that controls the branched electron pathway, the subunit architectures that control FBEB, and the key variations among the enzymes still remain vague. To generalize FBEB, we need a better understanding of the distribution of bifurcating enzymes taxonomically and ecologically so that we can identify the unique features that all such enzymes share. Understanding their taxonomic and ecological distribution will help to narrow down the unique niches of organisms capable of performing FBEB. That, in turn, will help us identity key environmental factors that promoted the emergence of FBEB enzymes. We can then also carry out comparative genomic analyses and structural comparisons at the level of proteins to identify key determinants that demarcate bifurcating enzymes from their non-bifurcating counterparts.

In summary, FBEB is an important mechanism that allows anaerobic microorganisms to minimize energy loss thereby conserving energy in the form of reduced compounds (e.g., Fd). Furthermore, this appears to be an efficient mechanism to conserve more energy than in cells operating SLP alone. A detailed understanding of

bifurcating enzymes could help us understand more about the lifestyles of these fermentative organisms and could also shed light on the evolutionary origins of these enzymes. In addition, several of these enzymes can generate commercial products such as ethanol, methanol and H_2 as a byproduct. Therefore, understanding the FBEB mechanism could lead to seminal advances in the engineering and control of electron transfer in anaerobic fermentative metabolism for the production of commercial products.

Overarching Goals

Enzymes that carry out FBEB were first discovered in an acetogen (e.g., *Clostridium klyuveri* (Li *et al*, 2008)). Since then, eleven more bifurcating enzymes have been identified (Buckel and Thauer, 2018a; Buckel and Thauer, 2018b). Except for one such enzyme (Fix, the electron transfer flavoprotein involved in nitrogen fixation), all of these were first described in strict to facultative anaerobes that primarily included sulfate reducers, methanogens and acetogens (Buckel and Thauer, 2018a; Buckel and Thauer, 2018b). The presence of bifurcating enzymes in methanogens and acetogens, which are widely believed to retain primitive physiology, has led to the speculation that FBEB is perhaps an ancestral process and may even have evolved before the divergence of Archaea and Bacteria (Martin, 2012; Sousa *et al*, 2016; Sousa *et al*, 2018).

Both methanogens and acetogens rely on the Wood-Ljungdahl pathway, which is the only exergonic pathway to fix carbon-dioxide (CO_2) (Nelson, 2008). The very first step of this pathway involves oxidation of hydrogen (H_2) to reduce Fd^+. Interestingly, under standard state conditions, Fd has much more negative potential than H_2, and

therefore the reduction of Fd^+ using H_2 is not thermodynamically feasible (Nelson, 2008; Thauer *et al*, 2008). The mechanism that allows autotrophs to reduce Fd^+ using H_2 remained a mystery until FBEB was discovered. Therefore, the hypothesis that bifurcating enzymes must be present in the LUCA (Last Common Universal Ancestor) of Archaea and Bacteria has been put forward. Although it is an attractive hypothesis, there exists no evidence that FBEB is an ancestral process or that bifurcating enzymes were present in the LUCA. A recent, *in-vitro* study showed that native transition metals (e.g. iron, nickel, and cobalt) have the capability to reduce CO_2 to acetate and pyruvate without the help of enzymes (Varma *et al*, 2018), suggesting that bifurcating enzymes and, in fact, biological mechanisms in general were not required to carry out CO_2 fixation on early Earth. This would indicate that FBEB need not be an ancestral process. This raises the question of why bifurcating enzymes evolved in the first place if not for CO_2 fixation and why more of them have not been found in aerobes.

To address the above questions, we generated the following hypotheses:

1) FBEB is not an ancestral process, and therefore bifurcating enzymes were not present in the LUCA of Archaea and Bacteria.
2) FBEB likely evolved to balance the Fd/nucleotide pool to improve the efficiency of energy metabolism, which is more crucial for strict anaerobes than for aerobes.
3) Aerobes might encode bifurcating enzymes only if they need to reduce Fd^+ to carry out anaerobic processes such as nitrogen fixation.

To test these hypotheses and to contribute to the understanding of FBEB, we integrated bioinformatics analyses with available biochemical and biogeochemical data. In Chapter 2, we focus on the distribution and evolution of bifurcating enzymes which will address the first two hypotheses. To assess the distribution and study the evolution of bifurcating enzymes, we leveraged available microbial genomic data in public databases such as NCBI (National Center for Biotechnology Information) for collecting available complete genomes and IMG (Integrated Microbial Genomes System) for extracting available environmental metagenomes. This chapter will cover the techniques used to identify the homologs of bifurcating enzymes in available complete genomes and metagenomes to assess their taxonomic and ecological distribution. In addition, phylogenetic analyses were conducted using the homologs identified in complete genomes to study the evolution of all known bifurcating enzymes and predict their taxonomic origin. Finally, phylogenetic analyses were conducted using the homologs identified in metagenomes to predict the environment types that likely promoted the emergence of FBEB enzymes. Besides documenting the distribution of bifurcating enzymes in complete genomes and in the environment, this study will shed light on whether FBEB is an ancestral process and whether any of these enzymes were present in LUCA of Archaea and Bacteria.

In Chapters 3 and 4, we investigate the functional role of bifurcating enzymes in aerobes to test my hypothesis 3. The only bifurcating enzyme found in aerobes is Fix, shown to be involved in nitrogen fixation which, interestingly, is an anaerobic process that has diversified into aerobes (Boyd *et al*, 2011a; Boyd *et al*, 2011b; Boyd and Peters,

2013; Boyd *et al*, 2015). Fix belongs to a large family of electron transfer flavoproteins (Etfs) that have been shown to be involved in diverse metabolisms including lipid and amino acid metabolism (Crane and Beinert, 1956). This family has been shown to form complexes with at least nine different acyl-CoA dehydrogenases, suggesting a versatile role in diverse metabolic reactions (Ghisla and Thorpe, 2004). Besides Fix, there exist three other Etf complexes shown to bifurcate electrons: butyryl-CoA dehydrogenase-Etf (Chowdhury *et al*, 2014), lactate dehydrogenase-Etf (Weghoff *et al*, 2015), and caffeyl-CoA reductase-Etf (Bertsch *et al*, 2013). Therefore, we identified all the organisms that encode Etf and focused primarily on Fix (as they are primarily found in aerobes) to understand their functional role. Since there are bifurcating and non-bifurcating Etfs, we also identified key determinants differentiating bifurcating Etfs. In chapter 3, we analyzed phylogenetic and structural variations among Etf homologs identified in complete genomes to discern the functional variation among Etfs and identify key sequence motifs that demarcate bifurcating from non-bifurcating Etfs. In Chapter 4, we identified all the organisms that can fix nitrogen (diazotrophs) and dissected their nitrogen fixation pathways to identify key enzymes and electron carriers involved in nitrogen fixation pathway. This chapter documents all the diazotrophs that encode a Fix complex and highlights their functional role in nitrogen fixation. Importantly, this chapter also sheds new light onto why the bifurcating Fix complex is required by these organisms to fix atmospheric nitrogen.

Chapter 5 summarizes all the findings discussed in Chapters 2, 3 and 4. In this section, major points found in each of the chapter will be discussed and they will be

connected to our overall theme to discuss whether FBEB is a property of LUCA and why the process of FBEB evolved. I will also talk about potential implications of this work for both academia and industry purposes. Furthermore, I will also include future directions that can further enhance our understanding of FBEB.

REFERENCES

Berry L, Poudel S, Tokmina-Lukaszewska M, Colman DR, Nguyen DM, Schut GJ et al (2017). H/D exchange mass spectrometry and statistical coupling analysis reveal a role for allostery in a ferredoxin-dependent bifurcating transhydrogenase catalytic cycle. Biochimica et Biophysica Acta (BBA)-General Subjects 1862: 9-17.

Bertsch J, Parthasarathy A, Buckel W, Muller V (2013). An electron bifurcating caffeyl-CoA reductase. J Biol Chem 288: 11304-11311.

Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW (2011a). A late methanogen origin for molybdenum-dependent nitrogenase. Geobiology 9: 221-232.

Boyd ES, Hamilton TL, Peters JW (2011b). An alternative path for the evolution of biological nitrogen fixation. Front Microbiol 2.

Boyd ES, Peters JW (2013). New insights into the evolutionary history of biological nitrogen fixation. Front Microbiol 4.

Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW (2015). Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. J Bact 197: 1690-1699.

Buckel W, Thauer RK (2013). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. Bba-Bioenergetics 1827: 94-113.

Buckel W, Thauer RK (2018a). Flavin-based electron bifurcation, ferredoxin, flavodoxin, and anaerobic respiration with protons (Ech) or NAD+ (Rnf) as electron acceptors: A historical review. Front Microbiol 9: 401.

Buckel W, Thauer RK (2018b). Flavin-Based Electron Bifurcation, A New Mechanism of Biological Energy Coupling. Chem Rev.

Chowdhury NP, Mowafy AM, Demmer JK, Upadhyay V, Koelzer S, Jayamani E et al (2014). Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J Biol Chem 289: 5145-5157.

Colman, D.R., Poudel, S., Stamps, B.W., Boyd, E.S., and Spear, J.R. (2017). The deep, hot biosphere: Twenty-five years of retrospection. Proc. Natl. Acad. Sci. USA 114, 6895-6903.

Crane F, Beinert H (1956). On the mechanism of dehydrogenation of fatty acyl derivatives of coenzyme A. J Biol Chem 218: 701-716.

Deistung J, Thorneley RNF (1986). Electron transfer to nitrogenase. Characterization of flavodoxin from *Azotobacter chroococcum* and comparison of its redoc potentials with those of flavodoxins from *Azotobacter vinelandii* and *Klebsiella pneumoniae* (*nifF*-gene product). Biochemical Journal 239: 69-75.

Demmer JK, Huang H, Wang SN, Demmer U, Thauer RK, Ermler U (2015). Insights into Flavin-based Electron Bifurcation via the NADH-dependent reduced Ferredoxin:NADP oxidoreductase structure. J Biol Chem 290: 21985-21995.

Demmer JK, Chowdhury NP, Selmer T, Ermler U, Buckel W (2017). The semiquinone swing in the bifurcating electron transferring flavoprotein/butyryl-CoA dehydrogenase complex from *Clostridium difficile*. Nat Commun 8.

Dick JM, Shock EL (2011). Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring. Plos One 6.

Edgren T, Nordlund S (2004). The *fixABCX* genes in *Rhodospirillum rubrum* encode a putative membrane complex participating in electron transfer to nitrogenase. J Bacteriol 186: 2052-2060.

Garcia Costas AM, Poudel S, Miller A-F, Schut GJ, Ledbetter RN, Fixen KR et al (2017). Defining electron bifurcation in the electron transferring flavoprotein family. J Bacteriol: JB. 00440-00417.

Ghisla S, Thorpe C (2004). Acyl-CoA dehydrogenases. The FEBS Journal 271: 494-508.

Herrmann G, Jayamani E, Mai G, Buckel W (2008). Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. J Bacteriol 190: 784-791.

Kaminski PA, Norel F, Desnoues N, Kush A, Salzano G, Elmerich C (1988). Characterization of the *fixABC* region of *Azorhizobium caulinodans ors571* and identification of a new nitrogen fixation gene. Mol Gen Genet 214: 496-502.

Kaster AK, Moll J, Parey K, Thauer RK (2011). Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. P Natl Acad Sci USA 108: 2981-2986.

Keltch AK, Strittmatter CF, Walters CP, Clowes GHA (1950). Oxidative phosphorylation by a cell-free particulate system from unfertilized Arbacia eggs. J Gen Physiol 33: 547-553.

Ledbetter R, Garcia-Costas AM, Lubner CE, Mulder DW, Tokmina-Lukaszewska M, Artz JH et al (2017). The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: Generation of low-potential reducing equivalents for nitrogenase catalysis. Biochemistry 56: 4177-4190.

Li F, Hinderberger J, Seedorf H, Zhang J, Buckel W, Thauer RK (2008). Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. J Bacteriol 190: 843-850.

Lipmann F (1941). Metabolic generation and utilization of phosphate bond energy. Adv Enzymol Rel S Bi 1: 99-162.

Lubner CE, Jennings DP, Mulder DW, Schut GJ, Zadvornyy OA, Hoben JP et al (2017). Mechanistic insights into energy conservation by flavin-based electron bifurcation. Nat Chem Biol 13: 655-659.

Martin WF (2012). Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. Febs Lett 586: 485-493.

Martin, W.F., Bryant, D.A., and Beatty, J.T. (2017). A physiological perspective on the origin and evolution of photosynthesis. FEMS Microbiol. Rev.

Mitchell P (1961). Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. Nature 191: 144-148.

Mitchell P (1975). The protonmotive Q cycle: a general formulation. Febs Lett 59: 137-139.

Nelson DL, Michael M. Cox (2008). Lehninger principles of biochemistry Fifth.

Nguyen DMN, Schut GJ, Zadvornyy OA, Tokmina-Lukaszewska M, Poudel S, Lipscomb GL et al (2017). Two functionally distinct NADP(+)-dependent ferredoxin oxidoreductases maintain the primary redox balance of *Pyrococcus furiosus*. J Biol Chem 292: 14603-14616.

Peters JW, Miller AF, Jones AK, King PW, Adams MWW (2016). Electron bifurcation. Curr Opin Chem Biol 31: 146-152.

Peters, JW, Beratan, DN, Schut, GJ, & Adams, MW (2018). On the nature of organic and inorganic centers that bifurcate electrons, coupling exergonic and endergonic oxidation–reduction reactions. Chemical Communications, 54(33), 4091-4099.

Reyntjens B, Jollie DR, Stephens PJ, GaoSheridan HS, Burgess BK (1997). Purification and characterization of a *fixABCX*-linked 2[4Fe-4S] ferredoxin from *Azotobacter vinelandii*. J Biol Inorg Chem 2: 595-602.

Sapra R, Bagramyan K, Adams MWW (2003). A simple energy-conserving system: Proton reduction coupled to proton translocation. P Natl Acad Sci USA 100: 7545-7550.

Schuchmann K, Müller V (2012). A bacterial electron-bifurcating hydrogenase. J Biol Chem 287: 31165-31171.

Schut GJ, Adams MWW (2009). The iron-hydrogenase of *Thermotoga maritima* utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. J Bacteriol 191: 4451-4457.

Sousa FL, Nelson-Sathi S, Martin WF (2016). One step beyond a ribosome: The ancient anaerobic core. Biochimica et Biophysica Acta (BBA)-Bioenergetics 1857: 1027-1038.

Sousa FL, Preiner M, Martin WF (2018). Native metals, electron bifurcation, and CO2 reduction in early biochemical evolution. Curr Opin Microbiol 43: 77-83.

Thauer RK, Kaster AK, Seedorf H, Buckel W, Hedderich R (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. Nat Rev Microbiol 6: 579-591.

Thauer, R.K., Jungermann, K. and Decker, K., 1977. Energy conservation in chemotrophic anaerobic bacteria. Bacteriological reviews, 41(1), p.100.

Varma SJ, Muchowska KB, Chatelain P, Moran J (2018). Native iron reduces CO2 to intermediates and end-products of the acetyl-CoA pathway. Nature ecology & evolution: 1.

Wang S, Huang H, Kahnt J, Thauer RK (2013). *Clostridium acidurici* electron-bifurcating formate dehydrogenase. Appl Environ Microb 79: 6176-6179.

Weghoff MC, Bertsch J, Muller V (2015). A novel mode of lactate metabolism in strictly anaerobic bacteria. Environmental Microbiology 17: 670-677.

Yan Z, Wang MY, Ferry JG (2017). A ferredoxin- and F420H2-dependent, electron bifurcating, heterodisulfide reductase with homologs in the domains Bacteria and Archaea. Mbio 8.

Zhang HB, Chobot SE, Osyczka A, Wraight CA, Dutton PL, Moser CC (2008). Quinone and non-quinone redox couples in Complex III. Journal of Bioenergetics and Biomembranes 40: 493-499.

Zhang P, Yuly JL, Lubner CE, Mulder DW, King PW, Peters JW et al (2017). Electron Bifurcation: Thermodynamics and kinetics of two-electron brokering in biological redox chemistry. Accounts Chem Res 50: 2410-2417.

CHAPTER TWO

ORIGIN AND EVOLUTION OF FLAVIN-BASED ELECTRON BIFURCATING ENZYMES

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Saroj Poudel

Contributions: Designed and conducted the experiment. Contributed to the writing of the manuscript.

Co-Author: Eric C. Dunham

Contributions: Downloaded the metagenomes and contributed to the writing of the manuscript.

Co-Author: Melody R. Lindsay

Contributions: Downloaded the metagenomes and contributed to the writing of the manuscript.

Co-Author: Maximiliano J. Amenabar

Contributions: Downloaded the metagenomes and contributed to the writing of the manuscript.

Co-Author: Elizabeth M. Fones

Contributions: Downloaded the metagenomes and contributed to the writing of the manuscript.

Co-Author: Daniel R. Colman

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Eric S. Boyd

Contributions: Provided feedback on the study design. Provided statistical advice and contributed to the writing of the manuscript.

Manuscript Information Page

Saroj Poudel, Eric C. Dunham, Melody R. Lindsay, Maximiliano J. Amenabar, Elizabeth M. Fones, Daniel R. Colman, and Eric S. Boyd

Frontiers in Microbiology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-review journal
____ Accepted by a peer-reviewed journal
__x_ Published in a peer-reviewed journal

Published: August 2018, doi: 10.3389/fmicb.2018.01762.

Origin and Evolution of Flavin-Based Electron Bifurcating Enzymes

Saroj Poudel, Eric C. Dunham, Melody R. Lindsay, Maximiliano J. Amenabar, Elizabeth M. Fones, Daniel R. Colman, and Eric S. Boyd*

Department of Microbiology and Immunology, Montana State University, Bozeman, Montana

*Author of correspondence: Eric S. Boyd (*eboyd@montana.edu*)
Department of Microbiology & Immunology
109 Lewis Hall
Montana State University
Bozeman, MT 59717
Phone: (406) 994-7046
Fax: (406) 994-4926

ABSTRACT

Twelve evolutionarily unrelated oxidoreductases form enzyme complexes that catalyze the simultaneous coupling of exergonic and endergonic oxidation–reduction reactions to circumvent thermodynamic barriers and minimize free energy loss in a process known as flavin-based electron bifurcation. Common to these twelve bifurcating (Bf) enzymes are protein-bound flavin, the proposed site of bifurcation, and the electron carrier ferredoxin. Despite the documented role of Bf enzymes in balancing the redox state of intracellular electron carriers and in improving the efficiency of cellular metabolism, a comprehensive description of the diversity and evolutionary history of Bf enzymes is lacking. Here, we report the taxonomic distribution, functional diversity, and evolutionary history of Bf enzyme homologs in 4,588 archaeal, bacterial, and eukaryal genomes and 3,136 community metagenomes. Bf homologs were primarily detected in the genomes of anaerobes, including those of sulfate-reducers, acetogens, fermenters, and methanogens. Phylogenetic analyses of Bf enzyme catalytic subunits (oxidoreductases) suggest they were not a property of the Last Universal Common Ancestor of Archaea and Bacteria, which is consistent with the limited and unique taxonomic distributions of enzyme homologs among genomes. Further, phylogenetic analyses of oxidoreductase subunits reveal that non-Bf homologs predate Bf homologs. These observations indicate that multiple independent recruitments of flavoproteins to existing oxidoreductases enabled coupling of numerous new electron Bf reactions. Consistent with the role of these enzymes in the energy metabolism of anaerobes, homologs of Bf enzymes were enriched in metagenomes from subsurface environments relative to those from surface

environments. Phylogenetic analyses of homologs from metagenomes reveal that the earliest evolving homologs of most Bf enzymes are from subsurface environments, including fluids from subsurface rock fractures and hydrothermal systems. Collectively, these data suggest strong selective pressures drove the emergence of Bf enzyme complexes via recruitment of flavoproteins that allowed for an increase in the efficiency of cellular metabolism and improvement in energy capture in anaerobes inhabiting a variety of subsurface anoxic habitats where the energy yield of oxidation-reduction reactions is generally low

Keywords: Electron bifurcation, flavin, ferredoxin, anoxic, subsurface, oxidoreductase, LUCA, metagenomes

Short Title: Evolution of Flavin-Based Electron Bifurcation

INTRODUCTION

Flavin-based electron bifurcation (FBEB) involves the simultaneous reduction of two electron acceptors using a single electron donor in an enzyme complex, whereby a thermodynamically favorable exergonic reaction drives a thermodynamically unfavorable endergonic reaction (Buckel and Thauer, 2013; Peters et al., 2016). A total of twelve enzymes have been shown to catalyze FBEB to date (**Table 1**) and each of these comprise multiple protein subunits that form a complex (**Figs. 1 & 2**). A common theme among bifurcating (Bf) enzymes is the involvement of ferredoxin (Fd) as a substrate as well as coordination of at least one flavin, the proposed site of bifurcation (Buckel and Thauer, 2013; Peters et al., 2016). Studies conducted on NAD(H)-dependent reduced Fd:NADP(H) oxidoreductase (Nfn), a protein complex that catalyzes the simultaneous endergonic reduction of oxidized Fd (Fd^+) and exergonic reduction of NAD^+ via oxidation of NADPH (Wang et al., 2010), reveals that the fully reduced flavin (hydroquinone) undergoes one electron oxidation by NAD^+ to generate an unstable flavin anionic semiquinone that is then further oxidized by one electron by Fd^+ (Lubner et al., 2017). Thus, an exergonic electron transfer from the flavin hydroquinone to NAD^+ 'pays for' the endergonic reduction of Fd^+ via the unstable semiquinone intermediate. In this way, Bf Nfn functions to reversibly reduce Fd^+ that can then be used to drive low potential electron transfer reactions. Nfn also functions to balance the ratio of oxidized to reduced NAD(H) and NADP(H) and is thus a key regulator modulating the favorability of catabolic and anabolic reactions (Wang et al., 2010; Demmer et al., 2015; Lubner et al., 2017).

In addition to allowing for the reduction of the low potential electron carrier Fd^+ and balancing the redox state of the pyridine nucleotide pool, FBEB has been suggested to improve the efficiency of cellular metabolism by allowing for more complete capture of energy released during substrate oxidation (Herrmann et al., 2008; Buckel and Thauer, 2013). This phenomenon appears to be particularly relevant for anaerobes that inhabit highly reduced environments where metabolic intermediates often accumulate due to difficulty in regenerating endogenous oxidants (i.e., Fd^+ or NAD^+) for those compounds. For example, glucose fermentation in *Clostridium pasteurianum* generates 3.0 mol ATP per glucose when non-Bf butyryl-CoA dehydrogenase (Bcd) is involved whereas 3.3 mol ATP per glucose is generated when Bf Bcd is involved, an increase of ~ 11% (Jungermann et al., 1973; Herrmann et al., 2008; Buckel and Thauer, 2013). During oxidation of 1.5 mol of glucose in the Embden-Meyerhof-Parnas pathway, a total of three NADH and three reduced Fd (Fd^-) are generated. All three Fd^- are shuttled toward hydrogen (H_2) generation, while all three NADH are used to produce butyryl-CoA when non-Bf Bcd is involved. However, when Bf Bcd is involved, one NADH is used to reduce Fd^+ that eventually yields an additional H_2, one NADH is used to reduce two acetyl-CoA to crotonyl-CoA, and the other NADH is used to further reduce crotonyl-CoA to butyryl-CoA, which generates additional ATP. The reader is referred to several recent reviews (Herrmann et al., 2008; Buckel and Thauer, 2013; Peters et al., 2016; Buckel and Thauer, 2018b; Müller et al., 2018; Peters et al., 2018) for additional discussion and examples of enhanced energy capture in the presence of Bf systems.

Despite the multiple, yet interrelated, roles of FBEB in the energy metabolism of anaerobes, little is known of the taxonomic distribution and functional diversity of Bf enzymes among cultivars or in the natural environment. Moreover, despite suggestions that FBEB was integrated into the metabolism of primitive lifeforms on early Earth (Martin, 2012; Nitschke and Russell, 2012; Martin et al., 2017; Baymann et al., 2018; Sousa et al., 2018) little is known of the evolutionary history of these enzymes or of the characteristics of environments that might have led to their evolutionary origin(s). In the present study, we hypothesized that Bf enzymes would be enriched in the genomes of anaerobes relative to aerobes and in communities that inhabit subsurface environments that tend to be reduced relative to those that inhabit surface environments that tend to be oxidized (Colman et al., 2017). Moreover, we hypothesized that Bf enzymes first evolved in anaerobes that inhabit subsurface environments. Finally, given that Bf enzymes are multi-subunit complexes [as summarized in (Buckel and Thauer, 2013; Peters et al., 2016; Buckel and Thauer, 2018b)] and often exhibit limited taxonomic distributions (Poudel et al., 2016; Berry et al., 2017; Garcia Costas et al., 2017; Nguyen et al., 2017), we hypothesized that Bf enzyme complexes emerged after the divergence of Archaea and Bacteria from the Last Universal Common Ancestor (LUCA). To address these interrelated hypotheses, we used bioinformatics techniques informed by biochemical data to identify homologs of the twelve known and biochemically characterized Bf enzyme complexes in 4,588 complete genome sequences. The distribution, diversity and evolutionary history of Bf enzyme homologs were analyzed in the context of the physiology of host organisms based on prior characterizations. Phylogenetic analyses

were conducted on each of the twelve Bf enzyme homolog datasets to identify microbial lineages that host the earliest evolving enzyme systems. To better define the characteristics of environments that select for microbial cells with electron Bf capability, and that may have precipitated their emergence, we also characterized the distribution and diversity of homologs of the twelve biochemically characterized Bf enzyme complexes in 3,136 available community metagenomes from a range of environments and cross-referenced this information with available metadata for these environments. Phylogenetic reconstructions were then performed to identify characteristics of environments that host early evolving homologs of Bf enzymes. Results are discussed in the context of the physiological and geochemical settings that enabled the multiple independent and recent origins of FBEB enzymes in biological systems and the role of environmental variation in driving the diversification of these enzymes.

MATERIALS AND METHODS

Generation of a genomic and metagenomic sequence database. All complete genomes (n=4,588) available in the National Center for Biotechnology Information (NCBI) database as of March 2016 were compiled. This compiled database included complete genomes of Archaea (n=230), Bacteria (n=4,343), and Eukarya (n=15). In addition, the protein sequences encoded in each of the environmental metagenomes (n = 3,136) in the Department of Energy's Integrated Microbial Genomes and Microbiomes (DOE-IMG) (Markowitz et al., 2011) as of April, 2017 were compiled. The DOE-IMG database was selected since it includes standardized metadata for metagenomes that are not always included in submissions to other databases such as NCBI or MG-RAST.

Identification and compilation of electron Bf enzyme homologs in genomes and metagenomic datasets. ***Overview of approach.*** The minimum number of subunits that comprise a Bf enzyme complex (**Table 1**) was identified based on literature surveys and empirical bioinformatics analyses, as described below for each Bf system. Representative sequences of each subunit were used as bait sequences to extract all the homologs of each subunit for each enzyme complex from the complete genome database (described above) using the phmmer program via the HMMER (ver. 3) software package (Eddy, 2015). Compiled homologs of each subunit for each enzyme complex were aligned with Clustal Omega (Sievers et al., 2011) and were further filtered to remove homologs that did not exhibit conservation in key active site motifs if they have been defined, as described for each enzyme system below. As a further screen, we then examined the protein encoding genes that flank the gene coding for the catalytic subunit for each homolog for the presence of the minimum subunits required to constitute a Bf complex and removed those candidate Bf homologs that did not meet our specified criteria. These steps are similar to the approaches that were taken to demarcate homologs of Bf [FeFe]-hydrogenase (Poudel et al., 2016), Bf transhydrogenase (Berry et al., 2017; Nguyen et al., 2017), and Bf electron transfer flavoproteins (Garcia Costas et al., 2017) from non-Bf paralogs. Using this information, a specific e-value was empirically determined to demarcate Bf homologs from their closest paralogs in metagenomic datasets (**Table 1**). Importantly, since assembled genomes from natural environments are often incomplete (only partial genomes or contigs are available), and since our criteria required the presence of genes encoding the minimum complement of proteins that

constitute a given Bf enzyme complex to be present for it to be counted as a homolog, the distribution of Bf enzyme homologs reported here for metagenomic sequences is likely to be a conservative estimate of their actual abundance.

For each Bf enzyme homolog, companion subunits for the catalytic subunit were identified in flanking gene regions based on the expected gene distributions from empirical analyses (see below for descriptions for each Bf enzyme as well as **Table 1**). A custom python script was used to assess the presence of companion subunits in each case, and the number of open reading frames that were surveyed for companion subunits are provided for each catalytic subunit in **Table 1**. If the genes encoding subunits necessary for bifurcation capacity were identified within these open reading frames, the enzyme homolog was considered as a putative Bf enzyme, whereas putative non-Bf enzyme homologs were identified by the absence of the specified subunits in the flanking gene regions. We also screened metagenomes for homologs of genes coding for putative Bf enzymes using the same approaches outlined briefly above and as described for each system below in more detail.

[FeFe]-hydrogenase. Our previous bioinformatics work classified putative Bf [FeFe]-hydrogenases (Hyd) as multimeric (**Fig. 1A**) and non-Bf [FeFe]-hydrogenases as monomeric or dimeric (Poudel et al., 2016). Putatively Bf trimeric Hyd complexes, first identified in *Thermotoga maritima*, include the catalytic subunit HydA, HydB that contains a flavin binding site, and HydC that contains ligands for iron-sulfur (FeS) cluster(s) (**Fig. 2A**) (Schut and Adams, 2009; Schuchmann and Muller, 2012). In addition to HydABC, tetrameric Hyd complexes comprise HydD that includes numerous

conserved cysteine residues that are putatively involved in coordinating FeS cluster(s) (Poudel et al., 2016). Therefore, the minimum gene complement that encodes for a putative Bf Hyd is *hydABC* (minimum among both trimeric and tetrameric forms) and these genes tend to be co-localized in genomes (Poudel et al., 2016). Because HydA often includes N- and C-terminal domains with homology to various proteins and FeS cluster binding domains such as is the case for HydA from *T. maritima* (AAD36496, **Table 1**), we first screened our databases using HydA from *Chlamydomonas reinhardtii* (AAL23572) as a query since it does not encode N- and C-terminal FeS cluster binding motifs (Mulder et al., 2010). Extracted HydA homologs (Bf and non-Bf) were then aligned with Clustal Omega (Sievers et al., 2011) and were further filtered to remove homologs that did not exhibit conservation in key active site L1, L2, and L3 motifs, as we and others have previously described (Meyer, 2007; Poudel et al., 2016).

[NiFe]-hydrogenase. Previous bioinformatics analyses have classified [NiFe]-hydrogenase into four phylogenetically and functionally coherent groups, with the "c" subgroup of group 3 (group 3c or Mvh) harboring homologs of Bf enzymes (Vignais et al., 2001; Boyd et al., 2014; Greening et al., 2016). Mvh comprise MvhA (the catalytic subunit where H_2 oxidation occurs), MvhG, and MvhD; both MvhG and MvhD harbor FeS cluster binding motifs (Peters et al., 2015; Greening et al., 2016). Mvh associates with heterodisulfide reductase (Hdr) forming a hexameric complex. This complex includes HdrA where the proposed Bf flavin is coordinated, and HdrB and HdrC, both of which have FeS cluster binding motifs (**Figs. 1B** and **2B**) (Kaster et al., 2011; Buckel and Thauer, 2013). Therefore, the minimum number of subunits required for a Bf Mvh

includes MvhADG and HdrABC. All homologs of the large subunit of [NiFe]-hydrogenase were compiled (see **Table 1**), aligned and screened for distal and vicinal cysteine pairs that delineate [NiFe]-hydrogenase from paralogs such as MbxL, FuoD, and NuoD (Schut et al., 2015). MvhA harbors L1 (ICGxCxxxH) and L2 (AYDPCccCATH) sequence motifs that delineate it from other non-Bf [NiFe]-hydrogenase large subunit groups (i.e., groups 1, 2, 3a, 3b, 3d, and 4) (Vignais et al., 2001; Greening et al., 2016). Therefore, the extracted MvhA homologs were further filtered to only include sequences that contained those motifs.

The HdrABC proteins that form a complex with MvhAGD are often not co-localized with MvhAGD [as is the case for *Methanothermobacter marburgensis* where Mvh-Hdr was first described (Kaster et al., 2011)]. Therefore, we relaxed our script to allow for genes that encode for HdrABC to be located anywhere in genomes that also encode MvhAGD. However, this approach cannot be used to scan for the subunits of HdrABC in metagenomes since the possibility exists that these genes could be from a different genome than homologs of MvhAGD. To identify the Mvh homologs in metagenomes we only screened for MvhAGD subunits, since, to date they have only been shown to associate with HdrABC and are co-localized (Kaster et al., 2011; Greening et al., 2016). Thus, if MvhAGD was identified, we assumed that HdrABC were also likely to be encoded in the same genome in our metagenomic screens.

Heterodisulfide-linked formate dehydrogenase (Fdh). Formate dehydrogenase typically functions as a monomeric unit (Ferry, 1990) but can associate with Hdr (i.e., HdrABC) to form a Bf Fdh complex. This enzyme complex was first identified in

Methanococcus maripaludis [**Fig. 1C**; (Costa et al., 2010; Costa et al., 2013)]. In fact, formate dehydrogenase was shown to associate with the same protein domain in MvhD that MvhAG associates with during the Mvh-Hdr bifurcation reaction (Costa et al., 2013). Hence, the MvhAG complex has been suggested to compete with formate dehydrogenase for MvhD (Costa et al., 2010; Costa et al., 2013). The Bf Fdh complex comprises two subunits, FdhA and FdhB, in addition to HdrABC subunits. FdhA exhibits a unique cysteine signature (i.e., $CxxCxxCx_{26}C$) that distinguishes it from its paralog FdhF2; FdhF2 is involved in another Bf complex that is described below (Wang et al., 2013b). Therefore, FdhA sequences were aligned, and only sequences that contained the specified cysteine motifs were retained for downstream analysis. A similar approach to that described above for Mvh was taken to identify homologs of Fdh in metagenomes. Since genes encoding for HdrABC are often not co-localized with those coding for FdhAB, we assumed that if FdhAB were detected that HdrABC were also likely to be encoded in that same genome in our metagenomic screens.

NADP(H)-dependent formate dehydrogenase (Hyt). The NADP(H)-dependent formate dehydrogenase (Hyt) complex, first identified in *Clostridium autoethanogenum*, comprises seven subunits (**Fig. 1D**) that include FdhA that contains a conserved [4Fe-4S] binding motif and HytA that is the catalytic site for H_2 oxidation (Wang et al., 2013a). In addition, the Hyt complex includes HytB, which is thought to coordinate a flavin, as well as HytC, HytD, HytE1, and HytE2, all of which comprise FeS cluster binding motifs (**Fig. 2D**). HytA is homologous to HydA and contains key active site motifs that are largely the same as those in HydA, including the L1 (TSCCPxW), L2 (MPCxxKxxE),

and L3 (ExMACxxGCxxGGGxP) motifs. Likewise, HytB and HytC are homologous to HydB and HydC, respectively. Homologs of HytA were aligned and were screened for the L1, L2, and L3 motifs, as defined above; those sequences that did not contain these motifs were discarded. HytE1, HytE2 and HytABCD are co-localized in the genome (**Supp. Table 2**), whereas FdhA tends to be located within four genes upstream of HylA (data not shown). These criteria were used to identify Bf Hyt in genomes and metagenomes (**Table 1**).

NAD(H)-dependent reduced ferredoxin:NADP oxidoreductase (Nfn). Nfn, which was first identified in *Clostridium klyuveri* comprises two subunits: the small subunit (NfnS) that contains one FAD and one [2Fe-2S] cluster and the large subunit (NfnL) that contains one FAD and two [4Fe-4S] clusters (**Figs. 1E** and **2E**) (Wang et al., 2010; Demmer et al., 2015; Berry et al., 2017; Lubner et al., 2017; Nguyen et al., 2017). Several classes of Nfn (NfnI, NfnII or Xfn and NfnIII) have been identified (Nguyen et al., 2017). NfnI has been shown to bifurcate electrons (Wang et al., 2010; Demmer et al., 2015; Berry et al., 2017; Lubner et al., 2017; Nguyen et al., 2017). NfnII retains a structure and cofactor composition like NfnI and thus has been proposed to bifurcate, although the high potential electron acceptor is likely to be different than NAD^+ (Nguyen et al., 2017). While there exists no structural studies on NfnIII, multiple sequence comparison of NfnIII with NfnI and NfnII revealed conserved motifs that potentially ligate the bifurcating flavin suggesting that NfnIII may also bifurcate (Nguyen et al., 2017). For these reasons, we treated all identified Nfn homologs as Bf. Extracted NfnS and NfnL sequences were aligned individually and then screened for the presence of

conserved motifs, as identified previously (Demmer et al., 2015), in order to further delineate putative Bf NfnSL homologs.

Electron transfer flavoprotein involved in nitrogen fixation (Fix). There are several classes of electron transfer flavoproteins, including those that bifurcate and those that do not, as discussed previously (Garcia Costas et al., 2017). Non-Bf Etf comprise two subunits, EtfA and EtfB, while all known Bf Etf that are involved in nitrogen fixation (i.e., Fix), comprise four subunits: FixA (EtfB), FixB (EtfA), FixC, and FixX (**Fig. 1F**) (Edgren and Nordlund, 2004; Garcia Costas et al., 2017; Ledbetter et al., 2017). Fix from *Azotobacter vinelandii* was the first enzyme shown to bifurcate electrons (Ledbetter et al., 2017). FixA, FixB, and FixC all contain flavin binding domains, but FixA is the proposed site of bifurcation (Ledbetter et al., 2017). FixX contains two [4Fe-4S] cluster binding motifs (**Fig. 2F**).

Butyryl-CoA dehydrogenase/electron transfer flavoprotein (Bf-Bcd). Butyryl-CoA dehydrogenase (Bcd) is typically involved in fatty-acid metabolism (Djordjevic et al., 1995) and has not been shown to bifurcate electrons. However, Bcd is homotetrameric and can associate with EtfAB and form a Bf complex, first identified in *C. kluyveri* [termed Bf-Bcd; (Herrmann et al., 2008)]. A crystal structure of Bf-Bcd revealed that it indeed is in complex with EtfA and EftB (Li et al., 2008; Demmer et al., 2017). All three subunits of this complex contain flavin binding motifs, while the flavin in EtfB has been proposed to be the Bf site (**Figs. 1G** and **2G**).

Caffeyl-CoA reductase/electron transfer flavoprotein (Car). To our knowledge, caffeyl-CoA reductase (CarC) has not yet been reported to function on its own (see caveat below). Rather, characterized enzymes form a complex that consists of three subunits: CarC, EtfA (CarD), and EtfB (CarE). This complex was first identified in *Acetobacterium woodii* (Bertsch et al., 2013). All three subunits contain flavin binding domains and CarE potentially harbors the Bf flavin (**Figs. 1H** and **2H**). In addition to containing a flavin binding domain, CarD also contains two [4Fe-4S] cluster binding motifs (**Fig. 2H**).

NAD(H) dependent formate dehydrogenase (Hyl). The NAD(H)-dependent formate dehydrogenase (Hyl) complex is tetrameric and was first identified in *Clostridium acidurici* (**Fig. 1I**) (Wang et al., 2013b). The tetrameric complex includes HylA, HylB, HylC, and FdhF2 (i.e., a formate dehydrogenase). All of the subunits contain FeS cluster binding motifs (**Fig. 2I**). In addition to putatively binding FeS clusters, HylB also contains a flavin binding motif (**Fig. 2I**). HylA is homologous to HydA but lacks conservation in the three aforementioned active site motifs, L1, L2, and L3 (Wang et al., 2013b). Extracted HylA homologs were aligned and demarcated from HydA and HytA homologs by screening for the absence of conserved L1, L2, and L3 signature motifs.

Lactate dehydrogenase/electron transfer flavoprotein (Bf-Ldh). Lactate dehydrogenase (Ldh) is involved in the reduction of pyruvate to lactate (Garvie, 1980). Ldh can associate with EtfAB and form a trimeric Bf complex [termed Bf-Ldh; **Fig. 1J**;

(Weghoff et al., 2015)]. Bf-Ldh was first identified in *A. woodii* and Ldh and EtfAB contain flavin binding motifs (**Fig. 2J**) (Weghoff et al., 2015). In addition to flavin binding motifs, EtfA also encodes a [4Fe-4S] cluster binding motif, and EtfB potentially houses the Bf flavin. Previous studies have found that Ldh and EtfAB tend to be co-localized in genomes (Weghoff et al., 2015), which guided our flanking gene analyses (**Table 1**).

$F_{420}H_2$-dependent heterodisulfide reductase (Hdr2). HdrABC are components of Bf Fdh and Mvh complexes. A functionally distinct paralog of HdrABC was discovered in *Methanosarcina acetivorans* (Buan and Metcalf, 2010) that was recently shown to function alone by coupling the oxidation of coenzyme $F_{420}H_2$ with the reduction of Fd^+ and heterodisulfide from coenzyme M (CoM) and coenzyme B (CoB) (termed Hdr2) [**Figs. 1K** and **2K**, (Yan et al., 2017)]. Like the Hdr complex identified in the Fdh and Mvh complexes, Hdr2 is comprised of three subunits that include the large Bf subunit Hdr2A that contains motifs to bind FAD, four [4Fe-4S] clusters, and one [2Fe-2S] cluster. Hdr2B and Hdr2C contain motifs for coordinating one [4Fe-4S] and two [4Fe-4S] clusters, respectively (**Fig. 2K**). Like HdrABC, the catalytic subunit of Hdr2A is not co-localized [as is the case for *M. acetovorans* (Yan et al., 2017)]. Furthermore, multiple copies of HdrA have previously been shown to exist in a single genome (Buan and Metcalf, 2010; Yan et al., 2017). Therefore, we classified HdrABC as Hdr2ABC only if the genome contained extra copies of HdrBC that were unaccounted for after considering the presence of other complexes that HdrABC forms associations with (i.e., Mvh, Fdh and Met). Metagenomes typically comprise multiple organisms which makes it difficult

to assess whether the Hdr2A is from the same genome as the Hdr2BC subunits and whether an identified Hdr complex functions alone or in complex with Mvh, Fdh, or Met. For these reasons, we did not assess the distribution of Hdr2 in metagenomes.

Methylene-H_4F reductase/ heterodisulfide reductase (Met). Methylene-tetrahydrofolate (H_4F) reductase, in its simplest form (i.e., as a single subunit), catalyzes the reduction of methylene-H_4F with reducing power from NADH (Guenther et al., 1999). It can also couple with other subunits that include MetV or Rnf where it then functions in a variety of anabolic and catabolic reactions (Bertsch et al., 2015). Furthermore, it has been shown to interact with HdrABC to form a Bf complex (termed Met), first identified in *Moorella thermoacetica* (Mock et al., 2014). In addition to HdrABC, the Bf Met complex comprises the three subunits, MetF, MetV, and MvhD (**Fig. 1L**). MetF contains a flavin binding site whereas MetV and MvhD encode motifs predicted to ligate a [4Fe-4S] cluster and a [2Fe-2S] cluster, respectively (**Fig. 2L**) (Bertsch et al., 2015). Importantly, MetF that forms a Bf complex contains a motif involved in flavin binding that is different from that found in MetF from *E. coli* which does not form a Bf complex (Bertsch et al., 2015). Hence, MetF homologs were aligned and only sequences that encoded the specified conserved residues as described in (Bertsch et al., 2015) were retained for downstream analysis. A similar approach to that described above for Mvh and Fdh was used to identify homologs of Met in metagenomes. MetFVD are co-localized but are not co-localized with HdrABC. Since genes encoding for HdrABC are often not co-localized with those coding for MetFVD, we assumed that

if MetFVD were detected that HdrABC were also likely to be encoded in that same genome in our metagenomic screens.

Statistical Analysis. A binary table was created to represent the presence or absence of specified Bf enzymes in each genome and metagenome. The binary table was used as input to generate an abundance plot and was also subjected to co-occurrence analysis using the co-occur package in R (Griffith et al., 2016).

Phylogenetic analysis. All 16S ribosomal RNA (rRNA) genes from genomes that encoded at least one homolog of a Bf enzyme were compiled via BLASTn. The extracted sequences were subjected to multiple sequence alignment using the SILVA rRNA gene database (Quast et al., 2013) as an alignment reference and the mothur program (version. 1.39.5) (Schloss et al., 2009). The aligned sequences were filtered to remove all gaps and incomplete 16S rRNA gene sequences as previously described (Lindsay et al., 2017). The filtered 16S rRNA gene sequence alignment block was used to generate a phylogenetic tree with RAxML (version 7.3.0) (Stamatakis, 2014) specifying the LG substitution matrix and the GTRGAMMA option to cluster the sequences into unique groups. FigTree (version 1.4.2) (Bogaardt, 2014) was used to visualize the tree.

To further define the potential taxonomic origin of the twelve Bf enzyme complexes among extant organisms with available genome sequences, and to identify environment types that harbor organisms with the most deeply rooted homologs of each of the twelve Bf complexes, we also subjected our curated database of oxidoreductase catalytic subunit homologs (i.e., HydA, HytA, MvhA, FdhA, FdhF2, NfnSL, FixAB,

Bcd, CarC, HylA, Ldh, and MetF) from complete genomes and metagenomes to phylogenetic analysis. We did not conduct phylogenetic analysis of the catalytic subunit of Hdr homologs (i.e., HdrA and Hdr2A) identified in metagenomes because of difficulty in identifying whether the subunits of Hdr2 (i.e., HdrABC) are from the same genome or not or if they function alone or in association with Met, Mvh, or Fdh, as described above. Likewise, while phylogenetic analyses were conducted on HdrA, we did not assign homologs as Met, Fdh, Mvh, or Hdr2 since it is not possible to link these functions based on genome context. Briefly, paralogs (described below) of each of the aforementioned catalytic subunits were identified and the curated putative Bf homologs were aligned using Clustal Omega (Sievers et al., 2011). The aligned sequences were then subjected to phylogenetic reconstruction using RAxML specifying the LG substitution matrix and the PROTGAMMA option to empirically cluster the sequences into unique groups. Itol was used to visualize the trees (Letunic and Bork, 2016).

Paralogs of the catalytic subunits for each Bf enzyme complex have already been identified in previous studies for many of the enzymes and these served as outgroups in phylogenetic reconstructions. These include the eukaryotic Nar-like protein that lacks the conserved L1, L2, and L3 motifs present in HydA/HytA (Meyer, 2007; Wang et al., 2010), non-Bf [NiFe]-hydrogenase large subunits from group 3d for MvhA (Boyd et al., 2014), the PyrK subunit of dihydroorotate dehydrogenase and the beta subunit of glutamate synthase, for Bf NfnSL, respectively (Demmer et al., 2015), the non-Bf group 5 EtfBA for FixAB (Garcia Costas et al., 2017), thioredoxin reductase for HdrA (Hedderich et al., 1994), and non-Bf MetF for Bf MetF (Bertsch et al., 2015). Paralogs of

the remaining catalytic subunits (FdhA, FdhF2, Bcd, CarC, and Ldh) were empirically determined by subjecting the representative catalytic sequences to BLASTp against the non-redundant protein database of NCBI. FdhA and FdhF2 are closely related paralogs and thus were used as outgroups for each other in phylogenetic reconstructions. Likewise, Bcd and CarC are closely related paralogs and thus were used as outgroups for each other in phylogenetic reconstructions. The most closely related sequence to Ldh was identified as alkyl dihyroxyacetone phosphate synthase, and this was used as an outgroup for Ldh in phylogenetic reconstructions.

The numbers of Nfn and Met homologs in metagenomes were much larger (i.e., >1,500 homologs) than those associated with other classes of Bf enzymes. Therefore, to reduce the computational time necessary to analyze the evolutionary history of homologs of these two enzymes, we first clustered them into homolog 'bins' using CD-HIT (Li and Godzik, 2006; Fu et al., 2012) at the 60% sequence identity level. As such, each unique bin contained closely related homologs. Representative sequences from each bin of NfnSL and MetF were used in the phylogenetic analyses and we then cross-referenced the sequences in the phylogeny with those within the bins to identify patterns in the distribution of homologs on the final tree and the environment types that hosted those homologs.

Pairwise sequence identity. To determine the extent to which individual subunits comprising each of the Bf complexes have co-evolved, we first aligned homologs of individual subunits using Clustal Omega (Sievers et al., 2011). Pairwise distances using the p-distance model (Kumar et al., 2016) were calculated in MEGA (Tamura et al.,

2007), and a dissimilarity matrix was generated using these distances for each specified protein subunit. Mantel tests were performed to calculate matrix correlations and their statistical significance with the R package ape (Paradis et al., 2004; Team R, 2014).

Lastly, we determined the variance in amino acid identities of homologs of the catalytic subunits of Bf complexes identified among the complete genome datasets as a proxy for the functional diversity of those homologs. Pairwise sequence comparisons of the homologs of each Bf enzyme catalytic subunit identified in complete genomes were used to calculate pairwise e-values using phmmer. E-values identified for each pairwise comparisons were used as a proxy for amino acid identity differences. E-values were normalized by multiplying by -10^4 and are presented as this transformed value. High e-values indicate that homologs exhibit a lower diversity and thus are more similar phylogenetically. Low e-values indicate that homologs are more diverse and are less similar phylogenetically. We then subjected homologs of the catalytic subunits of Bf complexes identified among the metagenome datasets to pairwise sequence comparisons against the dataset comprising the homologs identified in genome sequences using phmmer. This analysis was performed to assess the extent that homologs identified in the complete genome dataset adequately captured the natural diversity detected in metagenomes for each Bf enzyme. Homologs identified in metagenomes that exhibit a lower average or range of e-values than those identified in complete genomes are underrepresented in complete genome databases (i.e., there exists unsampled diversity).

RESULTS AND DISCUSSION

Distribution of electron Bf enzyme homologs in genome sequences. Of the total 4,588 archaeal, bacterial, and eukaryal genomes available, 681 (15%) coded for at least one homolog of a Bf enzyme (**Supp. Table 1**). Of these 681 genomes, 169 were from Archaea while 512 were from Bacteria while genomes from Eukarya did not code for homologs of any of the twelve FBEB enzymes. Among Archaea, homologs of Bf enzymes were detected in genomes from three out of four phyla considered in our database: Euryarchaeota (n=126 genomes), Crenarchaeota (n=42 genomes) and the candidate division Korarchaeota (n=1 genome). The majority (83%) of Bf enzyme homologs were identified in genomes from the bacterial phyla Firmicutes (n=184 genomes), Proteobacteria (n=147 genomes), Bacteroidetes (n=44 genomes), Thermotogae (n=25 genomes), and Spirochaetes (n=22 genomes) (**Supp. Table 1**).

We also classified organisms whose genomes coded for Bf enzyme homologs as a function of their ability to integrate oxygen (O_2) into their energy metabolism based on information acquired from the DOE-IMG database and previous physiological characterizations. Of the 681 identified organisms whose genomes encoded at least one Bf enzyme homolog, 16% (n=110 genomes) were aerobes, 9% (n=59 genomes) were facultative anaerobes, and 74% (n=503 genomes) were strict anaerobes. The prevalence of Bf enzyme homologs among obligate anaerobes is consistent with our hypothesis suggesting the importance of FBEB for organisms living in energy limited environments.

[FeFe]-hydrogenase (Hyd). Among the total 961 Hyd (putatively Bf and non-Bf) homologs detected in complete genome sequences, 24% were multimeric (i.e., HydABC or HydABCD) and hence can potentially bifurcate electrons (**Fig. 3A**). In total, 229

homologs of Bf Hyds were identified in complete genome sequences, of which 95% were from the genomes of anaerobes, 4% were from the genomes of facultative anaerobes, and 1% were from the genomes of putative aerobes that belong to the phylum Spirochaetes [e.g., *Turneriella parva* DSM 21527 (AFM13386) and *Salinispira pacifica* (AHC13718)] (**Fig. 3B);** the ability of *T. parva* and *S. pacifica* to use O_2 has not been robustly tested. The genomes that coded for homologs of Bf Hyds belonged exclusively to the bacterial domain, which is consistent with previous studies (Meyer, 2007; Calusinska et al., 2010; Mulder et al., 2010; Peters et al., 2015; Poudel et al., 2016).

Hyd is one of three evolutionarily distinct hydrogenases that catalyze the reversible reduction of protons to hydrogen (H_2) (Vignais et al., 2001; Peters et al., 2015). Bf Hyds couple the simultaneous and reversible reduction of NAD^+ and Fd^+ to the oxidation of H_2 [**Fig. 2A**; (Schut and Adams, 2009; Schuchmann and Muller, 2012)]. Bf Hyd thus plays a key role in balancing the ratio of oxidized to reduced NAD(H) and Fd pools in the cell. Although Bf Hyd homologs were primarily identified among the genomes of Firmicutes (63% of the total genomes) and Thermotogae (13% of the total genomes) (**Supp. Table 2**), they were also identified, albeit sporadically, among a wide diversity of organisms (**Supp. Fig. 1).** These organisms are supported by a variety of metabolisms that include metal reduction (e.g., *Alkaliphilus metalliredigens* QYMF), arsenate reduction (e.g., *Alkaliphilus oremlandii* OhILAs), nitrate reduction (e.g., *Clostridium perfringens*), sulfur reduction (e.g., *Pelobacter carbinolicus*), and sulfate reduction (e.g., *Desulfotomaculum acetoxidans* DSM 771), although it is not necessarily clear if these homologs are active under these physiological conditions. Moreover,

homologs of the Bf Hyd were identified in the genomes of organisms that inhabit a range of environments based on where they were isolated from, including those that are of moderate (e.g., *Clostridium beijerinckii*) to high temperature (e.g., *T. maritima*).

[NiFe]-hydrogenase (Mvh). Only 1% of the total 3,808 homologs of [NiFe]-hydrogenase that were identified (putatively Bf and non-Bf) in complete genomes were predicted to bifurcate electrons based on those genomes also coding for MvhADG and HdrABC and the presence of unique sequence motifs in MvhA (**Fig. 3A**). Mvh homologs were identified in the genomes of obligate anaerobes (**Fig. 3B**) and were largely confined to members of the archaeal domain (74% of the total Mvh encoding genomes) with the remaining 26% of the genomes that encode Mvh being from members of the bacterial domain (**Supp. Table 2**).

Bf Mvh couples the simultaneous and reversible reduction of Fd^+ and heterodisulfide from CoM-CoB to the oxidation of H_2 [**Fig. 2B**; (Thauer et al., 2008; Thauer et al., 2010; Kaster et al., 2011)]. Mvh was primarily identified in the genomes of methanogens that belong to the phylum Euryarchaeota (**Supp. Table 2**), although homologs were also detected in the genomes of sulfate-reducing bacteria (e.g., *Thermodesulfatator indicus* DSM 15286), halophilic bacteria (e.g., *Desulfohalobium retbaense* DSM 5692), ammonifying bacteria (e.g., *Ammonifex degensii* KC4), and sulfate-reducing archaea (e.g. *Archaeoglobus profundus* DSM 563). This suggests that the reversible reduction of Fd^+ and CoM-CoB with H_2 may be involved in a diversity of metabolisms (**Supp. Fig. 1**).

Heterodisulfide-linked formate dehydrogenase (Fdh). Formate dehydrogenase can form a Bf complex with *i*) a Hdr complex (termed Fdh; **Table 1**), *ii*) a [FeFe]-hydrogenase-like complex (termed Hyl), and *iii*) a NADP(H)-dependent [FeFe]-hydrogenase complex (termed Hyt). Among the total 2,516 homologs of formate dehydrogenase that were detected in available complete genomes, 2% were predicted to form a complex that would allow for bifurcation of electrons (either as Fdh, Hyl, or Hyt); the remaining formate dehydrogenase homologs are unlikely to bifurcate electrons and likely function as canonical formate dehydrogenases which reversibly oxidize formate to CO_2 [**Fig. 3A**; (Ferry, 1990)]. Among the putative Bf Fdh, 73% were associated with Hdr (termed Fdh), all of which were from obligate anaerobes (**Fig. 3B**) that were primarily within the archaeal phylum Euryarchaeota (**Supp. Table 2**).

Fdh couples the simultaneous and reversible reduction of Fd^+ and heterodisulfide from CoM-CoB to the oxidation of formate [**Fig. 2C**; (Costa et al., 2013)]. Interestingly, genes encoding homologs of Fdh were primarily detected in the genomes of methanogens and in the genomes of only two sulfate-reducing Proteobacteria, *Desulfobacterium autotrophicum* HRM2 and *Desulfobacula toluolica* Tol2 (**Supp. Fig. 1**). The role of putative Bf Fdh enzymes in the metabolism of these sulfate reducing bacterial taxa is not known.

NADP(H)-dependent formate dehydrogenase (Hyt). In total, the genomes of five acetogenic and obligately anaerobic bacteria that belong to the genus *Clostridium* within the phylum Firmicutes encoded Hyt (**Figs. 3A**, **3B** and **Supp. Table 2**). Hyt couples the simultaneous and reversible reduction of Fd^+ and $NADP^+$ to the oxidation of formate or

H_2 [**Fig. 2D**; (Wang et al., 2013a)]. Formate is suggested to bind to FdhA and it is thought that either formate or H_2 can serve as an electron donor to the Hyt complex (Wang et al., 2013a). However, it is not clear from previous biochemical characterizations where Fd, NADH, and H_2 bind in the complex. Several of the subunits that comprise Hyt (i.e., HytABC) are homologous to HydABC of the Bf [FeFe]-hydrogenase complex (Hyd). However, Hyt appears to be specific for NADP(H) whereas Hyd utilizes NAD(H) (Wang et al., 2013a). Nonetheless, based on homology, we suggest that H_2 binds to HytA in a manner like H_2 binds to HydA (Peters et al., 1998), Fd binds to HytC in a manner like Fd binds to HydC (Schut and Adams, 2009), and NADP(H) binds to HytB in a manner like NAD(H) binds to HydB (Schut and Adams, 2009) (**Fig. 2D**). The roles of HytE1, HytE2, and HytD in the functioning of Hyt are not presently known, however, they harbor conserved cysteine residues that may serve as ligands for FeS clusters. This may suggest that they are involved in electron transfer within the enzyme complex.

NAD(H)-dependent reduced ferredoxin:NADP oxidoreductase (Nfn). There are several classes of Nfn that have been termed NfnI, NfnII (also termed Xfn), and NfnIII, which are all proposed to bifurcate electrons (Nguyen et al., 2017). However, the high potential electron acceptor in NfnII is not defined but is known to not be NAD(H) like in NfnI. Based on this precedent in the literature, all Nfn identified in genome sequences herein are proposed to bifurcate electrons (**Fig. 3A**). Of the 397 archaeal and bacterial genomes that encoded Nfn, 72 were from Archaea and 325 were from Bacteria. The majority (82%) of the archaeal and bacterial genomes that coded for Nfn belonged to the bacterial phyla Firmicutes (n=138 genomes), Proteobacteria (n=53 genomes),

Bacteroidetes (n=44 genomes), Thermotogae (n=24 genomes), and the archaeal phylum Euryarchaeota (n=66 genomes) (**Supp. Table 2**). All the genomes that encoded Nfn were from anaerobic organisms with the exception of *Candidatus* Koribacter versatilis Ellin345 (ABF41796), which has been characterized as an aerobe (**Fig. 3B** and **Supp. Table 2**) (Sait et al., 2002).

Nfn is one of the four unique Bf enzymes whose crystal structure has been solved (Demmer et al., 2015; Lubner et al., 2017). Biophysical data collected on Nfn from *Pyrococcus furiosus* shows that flavin is indeed the site of bifurcation (Lubner et al., 2017). Nfn catalyzes the simultaneous and reversible reduction of NAD^+ and Fd^+ with the oxidation of NADPH [**Fig. 2E**; (Demmer et al., 2015; Lubner et al., 2017)] and hence plays a critical role in balancing the ratio of oxidized and reduced pyridine [(i.e., NAD(H) and NADP(H)] and Fd pools. Our taxonomic distribution data strongly suggests that Nfn is present in a diversity of microorganisms that operate a variety of metabolisms in diverse environmental settings. For example, Nfn is encoded in the genomes of strict anaerobes and facultative anaerobes (and possibly an aerobe, as described above) and in mesophilic and thermophilic taxa (**Supp. Fig. 1**). Moreover, Nfn was identified in the genomes of organisms that catalyze methanogenesis (e.g. *Methanosarcina barkeri*), acetogenesis (e.g. *M. thermoacetica*), elemental sulfur reduction (e.g. *Desulfurococcus mucosus*, *Thermosulfidibacter takaii*), dinitrogen reduction (e.g., *Rhodopseudomonas palustris*), heterotrophy (e.g., *P. furiosus*), and phototrophy (e.g., *Chlorobium tepidum*, *R. palustris*), among others (**Supp. Table 2**).

Electron transfer flavoprotein involved in nitrogen fixation (Fix). Among the 930 Etf homologs identified in sequenced genomes, 19% were demarcated as Fix and are predicted to bifurcate electrons (**Fig. 3A**). Unlike other Bf enzymes that are largely confined to anaerobic or facultatively anaerobic taxa, the majority (53% of total) of the homologs of Fix were identified in the genomes of aerobes (**Fig. 3B**). Of the 178 genomes that coded for homologs of Fix, 25% were from Archaea within the phylum Crenarchaeota (n=36) (**Supp. Table 2**). The remaining 75% of the Fix encoding genomes were from Bacteria within the phyla Proteobacteria (n=85 genomes) and Firmicutes (n=33) (**Supp. Table 2**).

A recent structural bioinformatics and phylogenetics study categorized Etfs into five unique groups, of which only group 2 contained putative Bf Etfs (Garcia Costas et al., 2017). Formally, Etf enzyme complexes that are involved in supplying Fd^- for use in nitrogen fixation are termed Fix (Ledbetter et al., 2017) and these belong to group 2D2 (Garcia Costas et al., 2017). Fix catalyzes the simultaneous and reversible reduction of Fd^+ and quinone to the oxidation of NADH [**Fig. 2F**; (Earl et al., 1987; Edgren and Nordlund, 2004; Ledbetter et al., 2017)]. Additional Fix homologs were identified among group 2 Etf that have the subunit architecture and sequence motifs that suggest an ability to bifurcate. These were identified in the genomes of elemental sulfur oxidizing aerobic archaea (e.g. *Sulfolobus islandicus*), iron reducing anaerobic archaea (e.g. *Pyrobaculum islandicum*), anaerobic acetogenic bacteria (e.g. *M. thermoacetica*), anaerobic sulfate-reducing bacteria (e.g. *Caldimicrobium thiodismutans*), and anaerobic fermentative thermophilic bacteria (e.g. *T. maritima*) (**Supp. Fig. 1**). The role of these enzymes in the

physiology of these cells is not known, nor has it been biochemically shown that these homologs bifurcate electrons. However, for simplicity we have designated all Fix-like Etf (that include Etf identified in diazotrophs and non-diazotrophs) as Fix in this study.

Electron transfer flavoprotein/butyryl-CoA dehydrogenase (Bf-Bcd). Butyryl-CoA dehydrogenase (Bcd) can form a complex with EtfAB and bifurcate electrons (Herrmann et al., 2008). In total, 309 homologs of Bcd were identified among genome sequences and 31% of these are predicted to form a complex with EtfAB (termed Bf-Bcd; **Table 1**) and hence can potentially bifurcate electrons (**Fig. 3A**). Ninety-eight percent of Bf-Bcd encoding genomes were from anaerobic taxa (**Fig. 3B**). Bf-Bcd homologs were confined to Bacteria where the enzyme was detected in the phyla Firmicutes (n=56 genomes), Fusobacteria (n=15 genomes), and Bacteroidetes (n=12 genomes) (**Supp. Table 2**).

Bf-Bcd was the first enzyme reported to be able to bifurcate electrons (Herrmann et al., 2008; Li et al., 2008). Composed of only three subunits (i.e. EtfAB and Bcd), the enzyme complex was shown to couple the simultaneous and reduction of Fd^{+} and crotonoyl-CoA to the oxidation of NADH during clostridial fermentation [**Fig. 2G**; (Chowdhury et al., 2014)]. Thermodynamically the reaction should be reversible; however, the enzyme has not been shown to be reversible yet. In addition to clostridia, pathogenic bacteria from the phyla Bacteroidetes (e.g. *Porphyromonas gingivalis*), Firmicutes (e.g. *Clostridium botulinum*), and Fusobacteria (e.g. *Fusobacterium nucleatum*) encoded Bf-Bcd. In addition, homologs of Bf-Bcd were detected in the genomes of metal reducers (e.g., *A. metalliredigens* QYMF), acetogens (e.g., *C.*

beijerinckii), and thermophilic heterotrophs (e.g., *Fervidobacterium nodosum*) (**Supp. Fig. 1**).

Electron transfer flavoprotein/caffeyl-CoA reductase (Car). Caffeyl-CoA reductase was identified by itself (i.e., CarC) or in complex with Etfs (i.e., CarCDE) (Bertsch et al., 2013). In total, 41 CarC homologs were identified and 56% of these are predicted to form a complex with CarDE (termed Car; **Table 1**), and thus are predicted to bifurcate electrons. (**Fig. 3A**). The Car complexes were identified in the genomes of anaerobic bacteria (**Fig. 3B**), primarily within the phylum Firmicutes (**Supp. Table 2**).

The Car complex was first described in *A. woodii* where it was shown to couple the simultaneous and reversible reduction of Fd^{+} and caffeyl-CoA with the oxidation of NADH [**Fig. 2H**; (Bertsch et al., 2013)]. Although it was first described in an acetogenic bacterium, data presented here indicates that Car may also function in the metabolism of an arsenite oxidizer (e.g., *A. oremlandii* OhILAs), a fermentative halophile (i.e., *Halanaerobium praevalens*), and in several fermentative thermophiles within the Thermotogales (e.g., *Thermosipho africanus*), among others (**Supp. Fig. 1**), based on the presence of homologs encoded in these organisms' genomes.

NAD(H)-dependent formate dehydrogenase (Hyl). As previously mentioned, formate dehydrogenases can function alone or form a complex with Hdr, Hyt, or a trimeric hydrogenase-like complex comprising HylABC [termed Hyl; (Wang et al., 2013b)]. Of the 2,516 formate dehydrogenase homologs identified, 19% are predicted to

form a complex with HylABC (**Fig. 3A**) and these homologs were detected only in the genomes of anaerobes (**Fig. 3B**) within the phylum Firmicutes (**Supp. Table 2**).

Hyl couples the simultaneous and reversible reduction of Fd^+ and NAD^+ with the oxidation of formate [**Fig. 2I**; (Wang et al., 2013b)]. Based on biochemical characterization of this complex (Wang et al., 2013b), it is not yet clear where substrates bind. However, the subunits of Hyl (i.e., HylABC) are homologous with those of the Bf Hyd (HydABC) with the exception that HylA lacks three conserved motifs (L1, L2, and L3) that are involved in ligating the H cluster in HydA (Meyer, 2007). Thus, it is possible that Fd binds to HylC in a manner like Fd binds to HydC and NAD(H) binds to HytB in a manner like NAD(H) binds to HydB (Schut and Adams, 2009; Schuchmann and Muller, 2012). Given the presence of conserved cysteine motifs, it is possible that HylA plays a role in electron transfer. Hyl was detected in the genomes of fermentative bacteria (e.g., *Clostridium acetobutylicum*), metal reducers (e.g., *A. metalliredigens* QYMF), arsenite oxidizers (e.g., *A. oremlandii* OhILAs), acetogens (e.g., *M. thermoacetica*), sulfate-reducers (e.g., *D. acetoxidans* DSM 771), and thermophilic heterotrophs (e.g., *Thermacetogenium phaeum* DSM 12270) (**Supp. Fig. 1**).

Electron transfer flavoprotein/lactate dehydrogenase (Bf-Ldh). Lactate dehydrogenase (Ldh) can form a complex with EtfAB (termed Bf-Ldh; **Table 1**) and bifurcate electrons (Weghoff et al., 2015). Among the total 6,158 homologs of Ldh that were identified, 1% are predicted to form a complex with EtfAB and thus potentially bifurcate electrons (**Fig. 3A**). The Bf-Ldh homologs identified belonged to the bacterial domain and were from either obligate anaerobes or facultative anaerobes (**Fig. 3B**).

Bf-Ldh couples the reduction of Fd^+ and pyruvate with oxidation of NADH [**Fig. 2J**; (Weghoff et al., 2015)]. Most Bf-Ldh homologs were detected among genomes from the phylum Firmicutes (n=55) and these included metal reducers (e.g., *A. metalliredigens* QYMF), acetogens (e.g., *A. woodii*), sulfate-reducers (e.g., *Desulfosporosinus acidiphilus* SJ4), and anaerobic phototrophs (e.g., *Heliobacterium modesticaldum* Ice1) (**Supp. Fig. 1**).

$F_{420}H_2$-dependent heterodisulfide reductase (Hdr2). A total of 770 homologs of heterodisulfide reductase (HdrABC) were identified, of which 46% were predicted to form a complex with either FdhAB, MvhAGD, or MetFV (described below) or to function alone [termed Hdr2; **Table 1**)]. In total, 63% (n=226 homologs) of the HdrABC homologs identified were identified in genomes that encoded copies of HdrABC attributable to Fdh, Mvh, or Met complexes. The extra copies of HdrABC in these genomes were thus designated as Hdr2 (**Fig. 3A**). Hdr2 homologs were primarily identified in genomes from the phyla Euryarchaeota (n=66 genomes), Proteobacteria (n=44 genomes) and Firmicutes (n=25 genomes) (**Supp. Table 2**). Furthermore, these homologs were identified in the genomes of anaerobes (i.e., 80% of the total genomes), aerobes (i.e., 7% of the total genomes) and facultative anaerobes (10% of the total genomes) (**Fig. 3B**). Hdr2 couples the reduction of Fd^+ and CoM-CoB with the oxidation of coenzyme $F_{420}H_2$ [**Fig. 2K**; (Yan et al., 2017)]. Hdr2 homologs were detected among methanogens (e.g., *M. acetivorans*), as previously noted (Yan et al., 2017), acetogens (e.g., *M. thermoacetica*), sulfate-reducers (e.g., *D. acidiphilus* SJ4), and anaerobic phototrophs (e.g., *H. modesticaldum* Ice1) suggesting that it is widely distributed among

organisms with diverse metabolic backgrounds (**Supp. Fig. 1**). While HdrABC complexes are widely found in strict anaerobes (Thauer et al., 2008; Buan and Metcalf, 2010; Yan et al., 2017), we were surprised to detect homologs of Hdr2 in the genomes of aerobic and microaerophilic organisms. These organisms primarily belonged to the phylum Aquificae that include *Aquifex aeolicus*, *Hydrogenobaculum* sp., and *Thermocinis albus* (**Supp. Table 2**). However, a recent study showed that the HdrABC (i.e., Hdr2 since it could not be accounted for by Fdh, Mvh, or Met) in *Aquifex aeolicus* was expressed in cultures grown with thiosulfate or elemental sulfur leading to the speculation that Hdr could be involved in sulfur oxidation through an undefined mechanism (Boughanemi et al., 2016).

Methylene-H_4F reductase/ heterodisulfide reductase (Met). A total of 1,017 methylene-H_4F reductase homologs were identified and 3% (n=25 homologs) of these were from genomes that also encoded MetFV, MvhD, and HdrABC. Thus, these homologs can putatively bifurcate (termed Met; **Table 1**) (**Fig. 3A**). Homologs of Met were primarily confined to the phyla Firmicutes (n = 12 genomes) and Proteobacteria (n=8 genomes) and all were from obligate anaerobes (**Fig. 3B**).

The Bf Met complex couples the reduction of methylene-H_4F to the oxidation of NADH and a yet to be identified electron acceptor that is thought to be Fd^+ [**Fig. 2L**; (Mock et al., 2014)]. Met homologs were also detected in a methanogen (i.e., *Methanomassiliicoccus intestinalis*), acetogens (e.g., *M. thermoacetica*), sulfate-reducers (e.g., *D. acidiphilus* SJ4), and nitrate reducers (e.g., *A. degensii*) (**Supp. Fig. 1**), suggesting a widespread distribution in anoxic habitats.

Putative interactions among Bf enzymes. Homologs of multiple (up to seven) different Bf enzymes were detected in a single genome and this was particularly true for organisms belonging to the phyla Firmicutes, Thermotogae, and Euryarchaeota (**Supp. Fig. 2; Supp. Table 2**). Among Firmicutes genomes, varying combinations of homologs of Hyd, Bf-Ldh, Car, Bf-Bcd and Nfn were often detected whereas among Thermotogae genomes it was common to detect varying combinations of the homologs of Hyd, Nfn, and Fix. Likewise, euryarchaeote genomes, in particular those from methanogens, often encoded a combination of Nfn, Mvh, Hdr2, and Fdh.

To quantify the extent that homologs of the twelve Bf complexes co-distribute in complete genome sequences, we conducted a co-occurrence statistical analysis. Fdh, Mvh, and Hdr2, which are primarily found in methanogens (**Fig. 2C**, **Fig. 2B** and **Fig 2K**, respectively), positively co-occurred (**Fig. 4**). This is consistent with biochemical data that indicates both [NiFe]-hydrogenase (MvhAG) and formate dehydrogenase (FdhAB) compete to bind MvhD that is associated with HdrABC (Costa et al., 2013). Maintaining homologs of both MvhAG and FdhAB likely allows the host to generate reduced Fd in environments with a dynamic supply of H_2 or formate (Costa et al., 2010; Costa et al., 2013). Hdr2 was also positively associated with the distribution of Mvh and/or Fdh in genomes (**Fig. 4)**, consistent with previous reports (Buan and Metcalf, 2010; Yan et al., 2017).

The distributions of Bf-Ldh, Car, Bf-Bcd, Hyd, and Nfn, all of which were primarily detected in bacterial genomes, were also positively correlated (**Fig. 4**). Importantly, since these five enzymes share common substrates [i.e., NAD(H) and Fd],

several authors have proposed that the reductant generated from one Bf enzyme can be utilized by another Bf enzyme (Bertsch et al., 2013; Buckel and Thauer, 2013; Weghoff et al., 2015), thereby allowing for tight control on the balance of oxidized/reduced substrates/products in a cell. For example, it has been proposed that Car partially relies on Hyd to generate NADH for use in reducing caffeyl-CoA to hydrocaffeyl-CoA, and in the process Fd^- is generated (Bertsch et al., 2013). The membrane bound Rnf complex, which catalyzes the reversible NAD^+-dependent oxidation of Fd^- by coupling to the electrochemical gradient (Biegel et al., 2011; Sarkar et al., 2012; Hess et al., 2013) could use the Fd^- to generate more NADH which can then feed into the bifurcation reaction carried out by Car (Bertsch et al., 2013). By coordinating these Bf enzymes and Rnf, oxidants (NAD^+, Fd^+) are generated and caffeyl-CoA is kept from building up, thereby allowing for metabolism to continue (Bertsch et al., 2013).

We further examined the co-distribution of Bf-Ldh, Car, Bf-Bcd, Hyd, and Nfn in complete genome sequences to identify evidence for potential interactions between these Bf complexes. Genomes that encode Bf-Bcd often also encode Nfn (68% of the Bf-Bcd encoding genomes) and Hyd (46% of the Bf-Bcd encoding genomes) (**Supp. Table 1**) suggesting Fd^- generated during the reduction of crotonyl-CoA may be used by either Nfn or Hyd to generate either NADPH or H_2, respectively. The bifurcation reaction for Nfn and confurcation reaction of Hyd requires NADH and this has been suggested to be provided through glycolytic reactions (Buckel and Thauer, 2013). Other examples of co-distributed Bf enzyme homologs are presented in **Supp. Table. 2**. Together, these observations suggest that coupling multiple redox reactions through Bf reactions allows

microorganisms a continual provision of NAD^+ or Fd^+ to allow efficient balancing of energy metabolism in oxidant limited anoxic environments inhabited by anaerobes.

Evolution of Bf enzymes. ***Overview.*** Twenty eight of the 41 phyla represented in our complete genome sequence dataset coded for at least one homolog of a Bf enzyme (**Supp. Table 2**). We examined the distribution of enzyme homologs in bacterial and archaeal genomes to, at first order, provide evidence of whether these enzymes were likely a property of the LUCA (i.e., homologs present in both archaeal and bacterial domains) or if they more likely evolved after the divergence of Archaea and Bacteria from the LUCA (i.e., homologs present in only a single domain). Similar analyses have been conducted on [FeFe]-hydrogenase (Mulder et al., 2010), molybdenum-dependent nitrogenase (Boyd et al., 2011a), and mercuric reductase (Barkay et al., 2010) to determine the likelihood that these enzyme complexes were a property of the LUCA. Secondly, we reconstructed the phylogeny of 16S rRNA genes (proxy for taxonomic evolution) recovered from genomes that code for homologs of at least one of the twelve Bf enzymes for use in examining the phylogenetic distribution of Bf homologs (**Fig. 5**).

Six (i.e., Nfn, Fix, Mvh, Fdh, Hdr2, and Met) of the twelve Bf enzymes were identified in genomes from both bacterial and archaeal domains indicating that they may have been a property of the LUCA. To more precisely assess this possibility, we examined *i*) the extent that subunits comprising a complex co-evolved and *ii*) the phylogeny of the oxidoreductase subunit for each of the Bf enzymes relative to paralogs as described below; the phylogeny of Hdr2 was not assessed (see materials and methods). Moreover, the oxidoreductase phylogenies were used to assess if subunits that allow for

bifurcation (flavoproteins) were recruited or lost (or both through multiple events) during the evolution of the oxidoreductases.

Hyd, Hyl, and Hyt. Hyd, Hyl and Hyt were only detected in bacterial genomes, suggesting that these three enzymes evolved after the divergence of Archaea and Bacteria from the LUCA. This conclusion agrees with a previous analysis of the evolution of Hyd (Mulder et al., 2010), but is the first to suggest this for Hyl and Hyt. The reconstructed phylogeny reveals that monomeric, non-Bf HydA homologs predate Bf HydA, HylA, and HytA homologs with high bootstrap support (**Supp. Fig. 3**). This suggests that additional subunits (i.e., Hyd/Hyl/HytBC and D in the case of tetrameric Hyd) were recruited to function in complex with Hyd/Hyl/HytA, enabling FBEB.

A mantel test conducted on the pairwise distances of all the subunits of the Hyd complex (i.e., HydABC) reveals positive correlations (**Supp. Fig. 4**) suggesting that once these subunits were recruited they were largely maintained and were under selection to co-evolve. The earliest evolving lineage of HydA homologs were from thermophiles that belonged to the phyla Thermotogae [e.g., *T. maritima* (NP_229226) and *T. africanus* (ACJ75249)] and Dictyoglomi [e.g., *Dictyoglomus thermophilum* (ACI20052)]. These lineages were well supported, suggesting that Hyd likely evolved and diversified in an ancestor of Thermotogae or Dictyoglomi (**Table 2**). Importantly, the taxa comprising the earliest evolving clade are thermophilic anaerobes, suggesting that Bf Hyd may have evolved in an anoxic, high temperature environment. Numerous clades of Bf HydA homologs formed interspersed lineages among monomeric non-Bf HydA lineages. The positive, strong relationships observed in the pairwise distances of HydABC (**Supp. Fig.**

4) suggests this observation to be the result of multiple loss events of genes encoding for HydBC rather than independent recruitment events.

Homologs of HylA exhibited a very narrow taxonomic distribution and were only identified in the bacterial phylum Firmicutes (**Fig. 5**). HylA homologs formed a single monophyletic clade suggesting that Hyl evolved once during its evolution (**Supp. Fig. 3**) through recruitment of HylBC and FdhF2, enabling FBEB. HylA homologs were nested among lineages comprising homologs of both Bf and non-Bf HydA, making it difficult to discern whether the ancestor of Hyl was already capable of FBEB (i.e., BC subunits were in place) or whether these subunits were recruited. Nonetheless, the pairwise distances of all the subunits of Hyl (i.e., HylABC) showed strong, positive correlations suggested they have been under selection to co-evolve (**Supp. Fig. 5**). The weak correlation between pairwise distances of FdhF2 and HylABC suggests that FdhF2 may have been recruited independently of HylABC and multiple times during the evolution of Hyl, potentially representing the trigger to diversify from Bf Hyd or non-Bf HydA.

The earliest evolving Hyl is from an anaerobic thermophile [*Thermincola potens* (ADG81609)] belonging to the phylum Firmicutes (100% bootstrap support). The next branching lineage also comprised homologs from thermophilic Firmicutes. This suggests an origin for Hyl among an anaerobic ancestor of the Firmicutes in an anoxic high temperature environment (**Table 2**).

Like Hyl, Hyt also exhibited a limited taxonomic distribution (**Fig. 5**) and homologs were only detected in the genomes of autotrophic and acetogenic *Clostridium* strains (**Table 2**), suggesting an origin in an anoxic environment. Phylogenetically, HytA

homologs were confined to a single lineage nested among non-Bf, monomeric HydA homologs. This indicates that HytBCDE1E2 and FdhA were likely recruited to an ancestor of HytA thereby allowing for FBEB (**Supp. Fig. 3**). Mantel tests conducted on matrices describing the pairwise distances of the subunits that comprise Hyt (i.e., HytABCDE1E1 and FdhA) show positive correlations suggesting that all these subunits are under selection and have co-evolved (**Supp. Fig. 6**). Interestingly, the HytA lineage comprises additional non-Bf HydA homologs, and these branching differences are well supported (i.e., >96% bootstrap support). Thus, like Hyd, additional subunits that were recruited to HytA and that allow for FBEB were apparently easily lost.

Mvh. Previous phylogenetic reconstructions of the large subunit of Mvh (i.e., MvhA) suggests that it is derived from a non-Bf [NiFe]-hydrogenases through recruitment of the Hdr complex (i.e., HdrABC) and other subunits (i.e., MvhGD) (Schut et al., 2013; Boyd et al., 2014), an assessment that is supported by our phylogenetic reconstruction (**Supp. Fig. 7**). Mantel tests conducted on matrices describing the pairwise distances of the subunits that comprise Mvh (i.e., MvhAGD and HdrABC) show strong positive correlations (**Supp. Fig. 8**), indicating that once these subunits were recruited they were maintained and co-evolved. Phylogenetic reconstruction of MvhA homologs revealed a single monophyletic clade that comprised both bacterial and archaeal homologs. The earliest evolving MvhA in this clade contained archaeal homologs from anaerobic methanogens (i.e., *Methanobacterium* and *Methanocella*) within the order Methanomicrobiales, with bacterial sequences nested among methanogen lineages. A second more recently evolved clade of MvhA was identified that comprised homologs

from anaerobic members of the Chloroflexi and Proteobacteria, suggesting acquisition of MvhA via horizontal gene transfer from a methanogen (**Supp. Fig. 7**). This observation, coupled with the limited distribution of Mvh in archaeal and bacterial genomes at a taxonomic level (**Fig. 5**), suggests that Mvh was not a property of the LUCA. Rather, Mvh likely emerged in an anaerobe after the divergence of Archaea and Bacteria from the LUCA through recruitment of several additional subunits, including HdrABC, to an existing [NiFe]-hydrogenase isoform.

Fdh. Bf Fdh homologs exhibit a narrow taxonomic distribution but were identified in both archaeal (Euryarchaeota) and bacterial (Proteobacteria) genomes. Phylogenetic reconstruction of the oxidoreductase subunit of Fdh (i.e., FdhA) suggests that Bf Fdh postdates non-Bf formate dehydrogenases (**Supp. Fig. 9**). This suggests that the subunits (i.e., FdhB and HdrABC) that form Bf Fdh were recruited to FdhA. The earliest branching putatively Bf FdhA lineages comprise homologs from anaerobic Bacteria [e.g., the δ-proteobacterium *D. autotrophicum* (ACN14546)] with high bootstrap support, suggesting that Bf Fdh likely emerged in a bacterium in an anoxic environment. Bf archaeal FdhA lineages are nested among and are paraphyletic to bacterial FdhA lineages. This suggests that FdhB/HdrABC may have been recruited to an ancestor of proteobacterial FdhA. Moreover, these observations suggest that once FdhB/HdrABC were recruited to function with FdhA as a Bf Fdh complex, the Fdh complex was then laterally transferred to Archaea. This interpretation is supported by strong positive correlations between the pairwise distances of the subunits of Fdh complex (i.e. HdrABC and FdhAB) (**Supp. Fig. 10**), indicating that once these subunits were recruited, they

were largely maintained and co-evolved. We interpret the patchy taxonomic distribution of Bf Fdh (only identified among Euryarchaeota and Proteobacteria; **Fig. 5**), the lack of monophyly among archaeal and bacterial Bf FdhA homologs, and the observation that euryarchaeote homologs are nested among bacterial FdhA homologs to indicate that Bf Fdh was unlikely to be a property of the LUCA.

Nfn. In the case of Nfn, all known homologs have been proposed to be capable of Bf electrons based on conserved motifs (Nguyen et al., 2017). Due to this observation, we assume that the ancestor of Nfn also bifurcated electrons. Moreover, paralogs of Nfn with similar activities but which lack one of the subunits (NfnS or NfnL) were not detected in genomes which precludes a detailed assessment of the nature of the first Nfn complexes. Nonetheless, a mantel test conducted on the pairwise distances of NfnS and NfnL show a strong positive correlation (**Supp. Fig. 11**), suggesting that NfnS and NfnL have been under selection and co-evolved.

Despite evidence presented here indicating that NfnSL have likely co-evolved, our phylogenetic reconstruction revealed polyphyly with regards to taxonomic domains, with several interspersed bacterial and archaeal lineages. This suggests several successful lateral gene transfers of *nfnSL* have taken place between Bacteria and Archaea and these events may have occurred after the divergence of Bacteria and Archaea from the LUCA. The earliest evolving well supported Nfn lineage comprises a homolog from *Mobiluncus curtisii* (ADI67453) belonging to the phylum Actinobacteria (**Table 2** and **Supp. Fig. 12**). The next branching lineage comprised a homolog from a proteobacterium. This suggests a bacterial origin for Nfn with early diversification among an ancestor of

Actinobacteria and possibly Proteobacteria. Consistent with our interpretation that Nfn originated after the divergence of Archaea and Bacteria from the LUCA, several studies have indicated that both Actinobacteria and Proteobacteria are recently evolved phyla (Forterre, 2015; Hug et al., 2016).

Fix, Car, and Bf-Bcd. Previous bioinformatics analyses conducted on Etf suggests that non-Bf Etf, Bf Car, and Bf Bcd predate the emergence of Bf Fix, an assessment that was based on high bootstrap support (Garcia Costas et al., 2017) and that is consistent with the analyses conducted here. This suggests that FixCX were recruited to function with EtfBA (i.e., FixAB). Once these subunits were recruited, they co-evolved as indicated by a mantel tests conducted on the pairwise distances of FixABCX that reveal strong positive correlations (**Supp. Fig. 13**). Phylogenetic reconstruction of FixAB reveals multiple monophyletic clades of archaeal and bacterial homologs that are paraphyletic, suggesting several lateral transfers of this gene between these domains (Garcia Costas et al., 2017). Early evolving Fix lineages were generally not well supported and comprised members of the Crenarchaeota as well as the Proteobacteria (**Table 2**). Paraphyly among bacterial and archaeal Fix lineages, combined with data indicating that Fix is primarily encoded in the genomes of aerobes and facultative anaerobes (**Fig. 3B**), suggests that Fix was not a property of the LUCA. Rather, these features suggest a more recent origin for this Bf complex and that this likely occurred after the advent of oxygenic photosynthesis and the increase of atmospheric O_2 ~ 2.4 Ga [as summarized in (Lyons et al., 2014)]. This interpretation is consistent with a recent

report suggesting that Fix emerged relatively recently to supply Fd^- for nitrogenase in aerobic and anoxygenic phototrophic taxa (Poudel et al., 2018).

Bf-Bcd and Car are closely related, but functionally distinct paralogs that were only detected in bacterial genomes (**Fig. 5**), suggesting that they emerged after the divergence of Archaea and Bacteria from the LUCA. Phylogenetically, the catalytic subunits of Bf-Bcd and Car (i.e., Bcd and CarC, respectively (**Table 1**)] each form coherent lineages that are interspersed with putatively non-Bf lineages (**Supp. Fig. 14**). This suggests that recruitment of electron transfer flavoprotein protein subunits (i.e., EtfAB and CarDE, respectively) that allow FBEB in these lineages each took place in singular events, but that loss of these subunits was common during the evolution of each of these functional complexes. This interpretation is supported by results from mantel regressions of matrices showing the pairwise phylogenetic distances of Bf-Bcd (i.e. EtfAB and Bcd) and Car (i.e., CarCDE) subunits, all of which show strong positive correlations (**Supp. Fig. 15** and **Supp. Fig. 16**, respectively). This indicates that subunits that comprise Bf-Bcd and Car have been under selection and have co-evolved, with several instances of loss of genes coding for EtfAB and CarDE, respectively, during evolution.

The earliest evolving putatively Bcd (the catalytic subunit of Bf-Bcd) lineage is from *Eubacterium limosum* (ALU16222) within the phylum Firmicutes (100% bootstrap support) which is followed by a several lineages comprising additional homologs from Firmicutes. These lineages are well supported, suggesting that Bf-Bcd likely evolved first in an anaerobic member of the Firmicutes. Likewise, the earliest evolving putatively Bf

CarC (the catalytic subunit of Car) homologs were identified in the genome of *Natranaerobius thermophilus* (ACB85453) within the phylum Firmicutes (100% bootstrap support) (**Supp. Fig. 14**), followed by several additional well supported lineages comprising homologs from anaerobic Firmicutes. Early evolving Bch and CarC homologs that putatively associate with EtfAB and CarDE, respectively, and are therefore predicted to bifurcate, form sister lineages. This suggests that these enzyme complexes may have evolved by duplication and subsequent diversification of the catalytic subunit after EtfAB/CarDE homologs had been recruited. Collectively, these results suggest that both Bf-Bcd and Car likely evolved in an anaerobic ancestor of the Firmicutes, possibly through duplication of genes, and that this event(s) postdated the divergence of Archaea and Bacteria from the LUCA.

Bf-Ldh. Homologs of Ldh (the catalytic subunit of Bf-Ldh) were only detected among bacterial taxa (**Fig. 5**) suggesting that Bf-Ldh also evolved after the divergence of Archaea and Bacteria from the LUCA (**Supp. Fig. 17**). Moreover, Ldh that are predicted to associate with EtfAB based on gene context and thus are predicted to bifurcate are nested among putatively non-Bf Ldh (not associated with EtfAB) suggesting that EtfAB were recruited to Ldh allowing for FBEB. Three distinct lineages of Bf-Ldh were identified, and these were interspersed among non-Bf homologs. Mantel regressions conducted on matrices describing the pairwise phylogenetic distances of subunits that comprise Bf-Ldh (i.e., EtfAB and Ldh) show positive correlations (**Supp. Fig. 18**). This suggests that the multiple lineages of Bf-Ldh likely have a common origin and that the non-Bf lineages that intersperse lineages of putatively Bf-Ldh result from gene loss.

The earliest evolving Bf-Ldh lineage comprised a homolog from the genome of the anaerobic, moderate thermophile *Thermovirga lienii* (AER67312) within the phylum Synergistetes (100% bootstrap support). The next branching lineage comprised homologs from genomes affiliated with the Thermotogae. Collectively, these observations suggest that Bf-Ldh evolved in an anaerobic ancestor of the Synergistetes or Thermotogae after the divergence of Archaea and Bacteria from the LUCA (**Supp. Fig. 17)**

Hdr2. HdrABC homologs were identified in both archaeal and bacterial genomes that were unaccounted for by Mvh, Fdh, and Met, and therefore were denoted as Hdr2. However, it is not possible to definitively determine which HdrABC homologs associate with Mvh, Fdh, or Met, unless HdrABC homologs are identified in genomes that do not encode Mvh, Fdh, or Met, which was uncommon. This in turn, makes it impossible to evaluate the evolutionary history of HdrA as it relates to bifurcation. However, regardless of whether HydrABC homologs can be definitively shown to form a complex with Bf complexes (i.e., Mvh, Fdh, Met and Hdr2), we can evaluate the evolutionary history of HdrA to determine whether HdrABC were likely to be a property of the LUCA.

Early evolving homologs of HdrA were from bacterial genomes that belonged to the phyla Proteobacteria (e.g. *Desulfobacca acetoxidans*) and Firmicutes (e.g. *M. thermoacetica*) (**Supp. Fig. 19**). Archaeal HdrA homologs were nested among bacterial HdrA homologs suggesting that Archaea acquired HdrA through horizontal gene transfer from Bacteria. These observations suggest that HdrA is not a property of the LUCA.

Met. A narrow taxonomic distribution of Bf Met homologs was detected among both archaeal (Euryarchaeota) and bacterial (Firmicutes) taxa. Phylogenetic reconstruction of homologs of MetF suggests that Bf Met postdates non-Bf methylene-H_4F reductases (**Supp. Fig. 20**). This suggests that the subunits that comprise the Met complex (MetV, MvhD, and HdrABC) were recruited to MetF to enable FBEB. Two distinct lineages of putatively Bf MetF were identified, and these were interspersed by non-Bf homologs. Mantel regressions conducted on matrices describing the pairwise phylogenetic distances of subunits that comprise Met (i.e., MetFV, MvhD and HdrABC) show positive correlations (**Supp. Fig. 21**). This suggests that once the subunits were recruited they were under selection and have co-evolved.

The earliest evolving lineage of Met comprised of only one homolog from the anaerobe *D. acetoxidans* (AEB08698) that belongs to the phylum Proteobacteria, with high bootstrap support. This suggests that Met likely evolved in an anaerobic ancestor of Proteobacteria. Only one archaeal Met homolog was identified, and this was encoded in the genome of the methanogen *M. intestinalis*. Phylogenetic reconstruction of MetF from this putative Bf complex reveals that it is nested among bacterial MetF homologs suggesting that it was likely acquired via horizontal gene transfer from Bacteria. Collectively, our findings suggest that Met was not a property of the LUCA but rather emerged in a bacterial anaerobe and was laterally transferred to Archaea.

Environmental distribution and phylogenetic ecology of electron Bf enzyme homologs. Results from our screening and analysis of complete genome sequences reveal that homologs of Bf enzymes are enriched in anaerobes and strongly suggest that each

enzyme system has an independent origin among anaerobic taxa, with the potential exception being Fix. We therefore hypothesized that microbial communities inhabiting anoxic environments would *i*) be enriched in homologs of Bf enzymes relative to oxic environments and *ii*) host the earliest evolving homologs of enzymes involved in FBEB. Specifically, we hypothesized that subsurface environments, which are characteristically limited in redox gradients and feature a low flux of oxidants and other nutrients (Edwards et al., 2012; Colman et al., 2017), would select for organisms reliant on anaerobic metabolism and Bf enzymes and may have promoted the origin of Bf enzyme complexes.

To test these hypotheses, we compiled 3,136 available metagenomic sequences and classified them as being from 'surface' or 'subsurface' environments based on a previously described classification scheme [(**Supp. Table 3**); (Colman et al., 2017)]. Briefly, metagenomes were classified as surface or subsurface based on available metadata associated with the metagenome, where subsurface environments were defined as the following environments: groundwater, deep subsurface, hydrothermal vents/springs, subsurface sediments, or marine sediments > 1 meter below sea floor. Surface environments included all others that did not fit the above classification, including saline environments, surface waters, and soils. We then screened these metagenomes for homologs of the eleven characterized Bf enzymes and subjected the catalytic oxidoreductase subunits to informatics analysis and phylogenetic reconstruction (metagenomes were not screened for Hdr2; see material and methods).

Screening of the 3,136 metagenomes revealed that 893 metagenomes coded for a homolog of at least one Bf enzyme (**Supp. Table 3** and **Supp. File 1-10**). We identified

homologs of nearly all eleven Bf enzymes in at least one metagenome except for homologs of the heptameric NADP(H)-dependent formate dehydrogenase (Hyt) complex. The lack of Hyt detection may be due to the stringent requirements that we imposed for identification of a Hyt complex (all six genes must be present in near synteny to be considered a Hyt homolog). However, while the number of subunits for a Bf enzyme (2 to 7 subunits) and the abundance of those enzyme homologs in metagenome sequences was inversely correlated, the relationship was not significant ($R^2 = 0.18$) (data not shown). This suggests that the lack of detection of homologs with more subunits was not due simply to the lower probability of assembling contigs with the minimal number of specified genes in near synteny.

In general, the abundance of Bf homologs identified in metagenomes (**Supp. Fig. 22A)** reflected the distribution identified in complete genomes (**Supp. Table 1**), with a few important exceptions. Mvh, and Met were enriched in the metagenomic dataset when compared to the genomic dataset. These enzyme complexes are all either involved in alkane-related metabolisms such as methanogenesis/methanotrophy, sulfate reduction, acetogenesis, or other potentially diverse anaerobic metabolisms (Buan and Metcalf, 2010; Kaster et al., 2011; Buckel and Thauer, 2013; Mock et al., 2014; Yan et al., 2017; Buckel and Thauer, 2018a; Buckel and Thauer, 2018b). Conversely, Bf-Bcd and Fix were enriched in the complete genome dataset when compared to the metagenomic dataset. Bf-Bcd is involved in fermentation (Djordjevic et al., 1995; Li et al., 2008; Chowdhury et al., 2014) and is predominantly identified among members of the Firmicutes and Bacteroidetes (**Supp. Table 1)** whereas Fix is involved in nitrogen fixation (Earl et al.,

1987; Edgren and Nordlund, 2004; Ledbetter et al., 2017) within the Proteobacteria and Firmicutes, although homologs predicted to bifurcate were identified in other non-diazotrophic taxa (**Supp. Table 1)**.

Differences in enrichment of Bf enzyme homologs in genomic versus metagenomic datasets might be explained by differences in the coverage of certain metabolic guilds in whole-genome databases (generally derived from cultivars) when compared to metagenomic datasets. For example, several taxonomic groups of methanogens or anaerobic alkane degraders have only recently been discovered or characterized via environmental genomics surveys (Paul et al., 2012; Evans et al., 2015; Laso-Perez et al., 2016; Vanwonterghem et al., 2016; Sorokin et al., 2017), suggesting that they exhibit much greater diversity than is currently represented in genomic databases from cultivars. This could account for the relative enrichment of Mvh in metagenomic datasets. In contrast, Bf-Bcd and Fix may be overrepresented in the whole-genome dataset due to overrepresentation of cultivated fermenters that tend to encode Bf-Bcd (Firmicutes and Bacteroidetes phyla) and diazotrophic Proteobacteria that tend to encode Fix in whole-genome datasets (Rinke et al., 2013).

Of the 893 metagenomes that encode Bf enzyme homologs, 275 were classified as being derived from subsurface-like environments, and 618 were classified as being derived from surface-associated environments (**Supp. Table 3**). Metagenomes with Bf enzyme homologs were further classified into eight sub-categories (**Fig. 6**), with most being derived from surface soils (30% of the total metagenomes) and surface waters (25% of the total metagenomes) that included lakes, freshwater streams, and oceans

(**Supp. Table 3**). When metagenomes from the subcategories of subsurface environments or surface environments were considered together, those from subsurface locales harbor a significantly greater abundance of Bf enzyme homologs per mega base pair (Mbp) of assembled sequence than those from surface environments (Welch's two sample *t* test, $P < 2.2 \times 10^{-16}$) (**Supp. Fig. 22B**). This is consistent with our hypothesis that the capacity to bifurcate electrons should be enriched in subsurface-like environments that generally exhibit minimal redox gradients, highly reducing conditions, and low nutrient fluxes (Edwards et al., 2012; Kieft, 2016; Colman et al., 2017). In particular, homologs of the ten Bf enzymes (Hyt was not detected and Hdr2 was not considered), when considered together, were significantly enriched in deep subsurface rocks and fracture fluids as well as hydrothermal springs/vents (**Fig. 6**) but were surprisingly not enriched in subsurface sediment microbiomes relative to those from surface environments.

We further scrutinized the distribution of Bf enzyme homologs among more specific environmental types (**Fig. 7**). Most of the Bf complexes were identified in at least seven of the eight metagenome subcategories, with the exception of Hyt (not detected), Hdr2 (not included) and Car, the latter of which was only identified in one deep subsurface and one soil metagenome. Homologs of Bf enzyme complexes were generally overrepresented in metagenomes associated with deep subsurface habitats, such as subsurface rock fracture fluids and were overrepresented in environments that share characteristics with subsurface environments, such as hydrothermal springs/vents (**Fig. 7**). This was particularly true for Nfn, Hyd, Fix, and Bf-Bcd. Like deep subsurface environments, hydrothermal springs/vents often exhibit highly reducing conditions and

minimal energetic gradients (Jannasch and Mottl, 1985; Amend and Shock, 2001; Martin et al., 2008; Shock et al., 2010; Martin, 2012; Amenabar et al., 2015). Consequently, the capacity to bifurcate electrons in these environments is likely ecologically advantageous, allowing anaerobic or facultatively anaerobic microorganisms to leverage minimal energetic gradients to support efficient metabolisms.

To assess whether subsurface-like environments promoted the origin of Bf enzymes complexes, we investigated the phylogenetic distribution of Bf homologs. We reconstructed phylogenies of homologs of catalytic subunits of each of the ten complexes identified in the metagenomic data, mapped the environments types to the phylogenies, and then determined the environment types that are associated with the earliest branching homologs for each complex. These analyses indicated that most (seven of the ten enzyme complexes that were detected/considered) of the Bf complexes featured deep-branching lineages comprising homologs from either deep subsurface or hydrothermal springs/vents metagenomes (**Fig. 7, Table 2, Supp. Fig. 23-30**), although other environmental origins could not be summarily excluded for several of these enzymes. Deep subsurface or hydrothermal spring/vent environments host early evolving lineages of the catalytic subunits of Nfn, Hyd, Fix, Mvh, Bf-Ldh, Hyl, and Car. The three enzymes (includes Met, Fdh, and Bf-Bcd) where phylogenetic data potentially suggest an origin in a surface like environment are nonetheless found almost exclusively in the genomes of anaerobes (**Fig. 3B**). Many surface environments considered to be broadly oxidized (e.g., soils, freshwaters, saline microbial mats) harbor anoxic microcompartments (Stein et al., 2001; Shen et al., 2016) that may support anaerobic taxa that encode Bf enzyme homologs,

potentially helping to explain this observation. In addition, high rates of microbial respiration have been observed in many surface environments such as in marine sediments of the ocean margins where oxygen (O_2) penetrates only millimeters to centimeters (D'Hondt et al., 2015). Thus, it is not unusual to find anaerobic microorganisms that may encode Bf enzyme homologs in environments that were classified as surface environments.

The taxonomic identities of the earliest branching homologs of Bf enzymes from metagenomes broadly agree with our analyses of complete genomes (**Table 2**). Specifically, both analyses indicate that several of the Bf enzymes are likely to have originated in thermophiles, sulfate reducers, and members of the Firmicutes. Thermophiles were well represented in the earliest branching homologs of Bf hydrogenases (Hyd and Mvh), Nfn, Fix, and Hyl. Origins for Bf enzymes in thermophiles is consistent with the generalized lack of O_2 in high temperature environments due to the inverse relationship between O_2 solubility and temperature (Amend and Shock, 2001) and to the limited availability of high potential oxidants in these environments (Shock et al., 2010). These characteristics would select for anaerobic taxa with metabolisms such as methanogenesis, acetogenesis, fermentation, and sulfur/sulfate reduction and for adaptations that maximize the efficiency of these metabolisms, such as FBEB (Buckel and Thauer, 2018b; Buckel and Thauer, 2018a).

For several classes of Bf enzymes, the deepest-branching lineages of homologs were phylogenetically divergent from those that are currently represented in our complete genome database (**Table 2**). It is possible that the function of these Bf homologs may

differ from those that have been biochemically characterized. For example, the deepest-branching [NiFe]-hydrogenase Mvh homolog was identified in a deep sea hydrothermal vent community and is distantly related to Mvh from *Acetomicrobium thermoterreneum* (**Table 2**). To date, Bf Mvh complexes have only been characterized in Archaea (Kaster et al., 2011), where they function to reduce the disulfide bond in CoM-CoB during methanogenesis (Thauer et al., 2008; Buckel and Thauer, 2013). The bacterium harboring this Bf Mvh-like complex almost certainly is not a methanogen, since methanogenesis has not been observed outside of Archaea (Vanwonterghem et al., 2016; Sorokin et al., 2017). Importantly, in addition to this deeply rooted Mvh homolog from a putative hydrothermal vent bacterium, homologs of Mvh were identified in the genomes of other non-methanogenic taxa, including those that catalyze sulfate, sulfur, and nitrate reduction (**Supp. Fig. 1**). This may suggest a unique functional role for Mvh in these non-methanogenic taxa.

With few exceptions, the amino acid sequence identities of earliest branching homologs for the ten Bf enzymes identified/considered in our metagenomic sequence dataset to the nearest characterized cultivar in our genome dataset were low (**Table 2**). The lack of close representation in existing genomic databases or culture collections for many of these early evolving homologs from metagenomic datasets suggests the presence of a greater diversity of Bf enzyme complexes in natural systems than is currently available in genome databases from cultivars. To assess the diversity of Bf enzyme homologs in genome and metagenomic datasets and to determine the extent that the diversity present in metagenomic datasets is captured by genomic datasets, we calculated

the pairwise sequence identities of each pair of homologous sequences within each of the ten Bf enzymes that were detected/considered in our metagenomic sequence analysis. Specifically, we conducted pairwise sequence comparisons of 1) metagenome versus complete genome homologs and 2) complete genome versus complete genome homologs. We quantified an e-value (see Materials and Methods) as a proxy for protein identity distance, where higher values indicate that average pairwise identities among sequences are higher and that they are thus less diverse on average.

Calculation of e-values for homologs indicates that those identified in complete genomes have a relatively narrow distribution of amino acid identities when compared to the homologs identified in the metagenomes (**Fig. 8**). In particular, the significantly lower median e-values for Bcd, FdhA, HylA, Ldh, and MetF homologs in metagenomic sequences when compared to those from complete genome sequences indicates that a greater diversity of Bf enzymes exists in natural environments than currently exists in genomic databases from cultivars. Although the median e-values for the remaining Bf enzymes were not significantly different between the datasets, they were all comparatively lower in the metagenomes than in the complete genomes and many metagenomes comprised unique outlier homologs (e.g., HydA), suggesting the presence of novel protein sequences. Collectively, these results suggest that there is a far greater diversity of uncharacterized homologs of Bf enzymes in natural environments that may harbor unique functional properties that are distinct from those homologs that have been biochemically characterized.

CONCLUSION

Complete and publicly available archaeal, bacterial, and eukaryal genomes were screened for homologs of the twelve biochemically characterized enzyme complexes that have been biochemically shown to bifurcate electrons. Homologs of putative Bf enzymes were identified in organisms with diverse metabolisms, including those of aerobes and anaerobes, chemotrophs and phototrophs, and autotrophs and heterotrophs. Homologs were identified in both archaeal and bacterial taxa, however, they were heavily concentrated in the genomes of anaerobic taxa and were enriched among members of the Euryarchaeota, Firmicutes, Bacteroidetes, and Proteobacteria. The limited taxonomic distribution of homologs of the twelve Bf complexes, the observed paraphyly of archaeal and bacterial homologs of catalytic subunits of eleven of these twelve complexes (Hdr2 not evaluated), and inferred origins of homologs among different taxonomic domains suggests that the ability to bifurcate electrons evolved after the divergence of Archaea and Bacteria from the LUCA and that these enzyme complexes evolved independently multiple times. This suggests an ecological advantage to integrate FBEB into the energy metabolism of functionally distinct and phylogenetically diverse taxa, specifically those that are anaerobic. Importantly, interspersed lineages of Bf and non-Bf homologs in phylogenetic reconstructions and strong evidence for co-evolution of protein subunits indicates that loss of flavoproteins was common during the evolution of the oxidoreductase subunits.

Interestingly, our conclusion that homologs of the twelve currently known Bf enzymes were not likely a property of the LUCA indicates that extant Bf enzyme complexes are not primordial and contrasts with conclusions from previous reports on the

natural history of FBEB (Baymann et al., 2018) and the origins of FBEB (Martin, 2012; Martin et al., 2017; Sousa et al., 2018). Baymen et al. 2018 based their conclusions on apparent structural conservation within three groups of flavoproteins that form complexes with oxidoreductases that enable FBEB. In at least one case of one of these flavoprotein groups (i.e., the Etf group that comprises four of the 12 Bf enzyme families), phylogenetic approaches suggest that flavoproteins that associate with Bf enzymes are derived from non-Bf enzymes (Garcia Costas et al., 2017). This indicates that the Bf EtfB flavoproteins are recently evolved and is consistent with our conclusions based on taxonomic distribution of these complexes and the evolutionary history of the oxidoreductase subunits. While the evolutionary history of the other two structural groups of flavoproteins has not yet been fully resolved, the culmination of taxonomic distribution of the enzyme complexes and evolutionary analyses of the oxidoreductase subunits compiled here leads us to conclude that the Bf flavoproteins in the remaining two groups are also unlikely to be primordial. Regardless of whether the Bf flavoproteins themselves are primordial, the current configuration of these Bf flavoproteins and oxidoreductases are unlikely to be reflective of primordial complexes, based on the evolutionary analyses conducted herein.

Our analysis of the distribution and composition of Bf enzyme homologs in metagenomic sequences revealed enrichment of homologs in subsurface environments relative to surface environments. This is particularly true for deep subsurface rock fracture fluids and thermal spring/hydrothermal vent environments, where homologs of Bf enzymes were shown to be highly enriched. Moreover, phylogenetic reconstructions

of the catalytic subunits of Bf enzyme complexes identified in metagenomic sequences revealed that the earliest evolving lineages were often (seven out of ten instances; Hyt not identified in metagenomic datasets, Hdr2 not assessed) from deep subsurface rock fracture fluid and thermal spring/hydrothermal vent communities. This suggests that the characteristics of these environments that at one time promoted the origin of these Bf complexes may also help to explain the prevalence of Bf enzyme homologs in these environments today. The specific characteristics of these environments that may have promoted the origins of several of the Bf enzyme classes, and that promote the establishment of contemporary communities enriched in these protein complements, includes minimal redox gradients, highly reducing conditions, and low nutrient fluxes (Edwards et al., 2012; Kieft, 2016; Colman et al., 2017). Many of the homologs of Bf enzymes recovered from these environmental genomes, including those that branch early in phylogenetic reconstructions, exhibit low sequence identities with homologs available in complete genome databases. This suggests the existence of novel and diverse Bf homologs in natural environments that could be prioritized for characterization. Moreover, the apparent ease by which phylogenetically distinct flavoproteins can be recruited to function with oxidoreductases (twelve independent origins of Bf enzymes, as revealed here) suggests the possibility of numerous other yet to be discovered Bf systems in both genomic and metagenomic sequence datasets.

The observations made here suggest that FBEB is a widely distributed strategy in organisms inhabiting anoxic environments where the energy yield of redox reactions is generally low and where increased efficiency of energy capture would be highly

selectable (Peters et al., 2016; Buckel and Thauer, 2018a; Buckel and Thauer, 2018b; Peters et al., 2018). The prevalence of homologs of putatively Bf enzymes in the genomes of anaerobes and in subsurface environments that are presumably anoxic is consistent with the hypothesized role of FBEB in supporting the energy metabolism of anaerobic life on early Earth (Martin, 2012; Sousa et al., 2018), when O_2 concentrations were low (Lyons et al., 2014) and energetic gradients capable of supporting metabolism are thought to have been minimal (Schut et al., 2016). Phylogenetic data indicates that most Bf enzyme systems evolved from putatively non-Bf ancestors. Thus, as biology evolved new metabolic strategies allowing it to diversify into new ecological niches, many existing oxidoreductases were fine-tuned through recruitment of flavoproteins allowing FBEB. This in turn, would have allowed for coupling of new oxidation-reduction reactions involving a variety of substrates in cells that function to increase the metabolic efficiency and thus competitiveness of those cells. The near complete absence of Bf enzyme homologs in aerobic cells (exception being Fix) suggests that as biology diversified from anoxic to oxic environments, the selective advantage for Bf systems decreased, presumably due to toxicity of O_2 to the cells or the sensitivity of the enzymes themselves to O_2. In the case of Fix, it appears that aerobic diazotrophic cells specifically acquired and maintained this enzyme complex as an adaptation to generate Fd^- for use in supporting nitrogenase activity (Poudel et al., 2018), which evolved first in an anaerobe and later diversified into aerobes (Boyd et al., 2011a; Boyd et al., 2011b; Boyd and Peters, 2013; Boyd et al., 2015).

Significantly more work needs to be completed to identify the specific triggers that led to the emergence and extinction of Bf systems as biology diversified to occupy the breadth of niche space as it does today. Moreover, additional work needs to be conducted to identify the unique environmental characteristics that selected for the emergence of Bf systems. Such studies will shed new light on the evolutionary history of FBEB and provide new perspective on the physiological and ecological role of Bf enzymes in these cells and the natural environment. The insights and perspectives obtained from conductance of these evolutionary ecology studies, framed within a robust understanding of the physiology and biochemistry of these enzymes, will provide a plethora of exciting new questions to guide future thinking and research into the nature of microbial life on early Earth and the role of FBEB in the physiology and ecology of that life. Insights from such studies will ultimately be required to evaluate the merits of placing the process of FBEB as a property of the LUCA, as has been suggested elsewhere (Martin, 2012; Martin et al., 2017; Baymann et al., 2018; Sousa et al., 2018), and the merits of placing the emergence of FBEB after the divergence of Bacteria and Archaea from the LUCA, as suggested here. If it turns out that FBEB was not a property of the LUCA, as advocated here, new challenging questions will emerge such as how to reconcile the energy metabolism of cells whose contemporary energetics are dependent on FBEB. In the case of hydrogenotrophic methanogens, which are often advocated as operating energy metabolisms reminiscent of that which operated in the LUCA (Weiss et al., 2016), it would be necessary to envision a primitive energy metabolism within a specific environmental context that favored reaction energetics that were not dependent

on FBEB (Thauer et al., 2008). Similar scenarios would need to be envisioned for other fermentative and autotrophic cells whose contemporary energy metabolisms are also dependent on FBEB.

ACKNOWLEDGEMENTS

This work was supported as part of the Biological Electron Transfer and Catalysis Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, and Basic Energy Sciences under Award # DE-SC0012518.

AUTHORS CONTRIBUTIONS

S.P. and E.S.B. designed the study and wrote the paper. S.P. conducted informatics analyses. S.P., E.C.D., M.R.L., M.J.A., E.M.F., D.R.C. and E.S.B downloaded the metagenomes, provided feedback, and contributed to the writing of the manuscript.

COMPETING INTERESTS

The authors declare that they have no competing interests.

ADDITIONAL FILES

Supplemental Table 1

Supplemental Table 2

Supplemental Table 3

Supplemental Fig. 1

Supplemental Fig. 2

Supplemental Fig. 3

Supplemental Fig. 4

Supplemental Fig. 5

Supplemental Fig. 6

Supplemental Fig. 7

Supplemental Fig. 8

Supplemental Fig. 9

Supplemental Fig. 10

Supplemental Fig. 11

Supplemental Fig. 12

Supplemental Fig. 13

Supplemental Fig. 14

Supplemental Fig. 15

Supplemental Fig. 16

Supplemental Fig. 17

Supplemental Fig. 18

Supplemental Fig. 19

Supplemental Fig. 20

Supplemental Fig. 21

Supplemental Fig. 22

Supplemental Fig. 23

Supplemental Fig. 24

Supplemental Fig. 25

Supplemental Fig. 26

Supplemental Fig. 27

Supplemental Fig. 28

Supplemental Fig. 29

Supplemental Fig. 30

Supplementary_File_1_Hyd_meta

Supplementary_File_2_Mvh_meta

Supplementary_File_3_Fdh_meta

Supplementary_File_4_Nfn_meta

Supplementary_File_5_FixB_meta

Supplementary_File_6_Bcd_meta

Supplementary_File_7_Car_meta

Supplementary_File_8_Hyl_meta

Supplementary_File_9_Ldh_meta

Supplementary_File_10_Met_meta

Table 1. The twelve enzyme complexes that have been biochemically shown to bifurcate electrons to date. The minimum complement of subunits comprising each enzyme, a reference NCBI protein sequence ID, an e-value cutoff for demarcating subunit homologs from closely related paralogs, the number of open reading frames from the catalytic subunit (boldfaced) searched for homologs of companion subunits, and literature references are provided.

Enzyme	Subunits[a]	Protein ID	E-Value	Flanking Region	Reference
A) NAD(H)-dependent [FeFe]-hydrogenase (Hyd)	**HydA**	AAD36496	2e-20		(Schut and Adams, 2009)
	HydB	AAD36495	2e-13	+/- 5	(Schut and Adams, 2009)
	HydC	AAD36494	1e-07	+/- 5	(Schut and Adams, 2009)
B) [NiFe]-hydrogenase/heterodisulfide reductase (Mvh)	**MvhA**	CAF30379	1e-90		(Kaster *et al*, 2011; Thauer *et al*, 2008)
	MvhG	CAF30378	5e-07	+/- 3	(Kaster *et al*, 2011; Thauer *et al*, 2008)
	MvhD	CAF30377	4e-07	+/- 3	(Kaster *et al*, 2011; Thauer *et al*, 2008)
	HdrA	CAF30381	3e-08	All[b]	(Kaster *et al*, 2011; Thauer *et al*, 2008)
	HdrB	CAF30711	1e-03	All[b]	(Kaster *et al*, 2011; Thauer *et al*, 2008)
	HdrC	CAF30710	2e-06	All[b]	(Kaster *et al*, 2011; Thauer *et al*, 2008)
C) Formate dehydrogenase/ heterodisulfide reductase (Fdh)	**FdhA**	CAF30854	6e-12		(Costa *et al*, 2010; Costa *et al*, 2013)
	FdhB	CAF30853	8e-15	+/- 1	(Costa *et al*, 2010; Costa *et al*, 2013)
	HdrA	CAF30381	3e-08	All[b]	(Costa *et al*, 2010; Costa *et al*, 2013)

	HdrB	CAF30711	1e-03	All[b]	(Costa *et al*, 2010; Costa *et al*, 2013)
	HdrC	CAF30710	2e-06	All[b]	(Costa *et al*, 2010; Costa *et al*, 2013)
D) NADP(H)-dependent [FeFe]-hydrogenase (Hyt)	FdhA	AGT29705	6e-12	+/- 10	(Wang *et al*, 2013a)
	HytE2	AGT29714	1e-43	+/- 10	(Wang *et al*, 2013a)
	HytA	AGT29713	1e-04		(Wang *et al*, 2013a)
	HytE1	AGT29712	1e-43	+/- 10	(Wang *et al*, 2013a)
	HytD	AGT29711	3e-17	+/- 10	(Wang *et al*, 2013a)
	HytB	AGT29710	2e-13	+/- 10	(Wang *et al*, 2013a)
	HytC	AGT29709	1e-07	+/- 10	(Wang *et al*, 2013a)
E) NAD(H)-dependent reduced ferredoxin:NADP(H) oxidoreductase (Nfn)	NfnS	AAD36706	9e-22		(Demmer *et al*, 2015)
	NfnL	AAD36707	2e-96	+/- 1	(Demmer *et al*, 2015)
F) Electron-transfer flavoprotein involved in nitrogen fixation (Fix)	**FixA/EtfB**	WP_011665895	1e-43		(Edgren and Nordlund, 2004)
	FixB/EtfA	WP_011665894	6e-45	+/- 4	(Edgren and Nordlund, 2004)
	FixC	WP_011665893	4e-35	+/- 4	(Edgren and Nordlund, 2004)
	FixX	WP_011665892	3e-05	+/- 4	(Edgren and Nordlund, 2004)
G) Butyryl-CoA dehydrogenase / electron transfer flavoprotein (Bf-Bcd)	**Bcd**	EDK32509	3e-12		(Herrmann *et al*, 2008; Li *et al*, 2008)
	EtfA	EDK32511	6e-45	+/- 2	(Herrmann *et al*, 2008; Li *et al*, 2008)
	EtfB	EDK32510	1e-43	+/- 2	(Herrmann *et al*, 2008; Li *et al*, 2008)
H) Caffeyl-CoA reductase/ electron transfer flavoprotein (Car)	**CarC**	AFA48354	2e-12		(Bertsch *et al*, 2013)
	CarD/EtfB	AFA48355	1e-43	+/- 2	(Bertsch *et al*, 2013)
	CarE/EtfA	AFA48356	6e-45	+/- 2	(Bertsch *et al*, 2013)

I) NAD(H)-dependent formate dehydrogenase (Hyl)	FdhF2	AFS79904	6e-12	+/- 4	(Wang *et al*, 2013b)
	HylA	AFS79905	2e-39		(Wang *et al*, 2013b)
	HylB	AFS79906	3e-14	+/- 4	(Wang *et al*, 2013b)
	HylC	AFS79907	5e-38	+/- 4	(Wang *et al*, 2013b)
J) Lactate dehydrogenase/ electron transfer flavoprotein (Bf-Ldh)	**Ldh**	AFA47664	3e-29		(Weghoff *et al*, 2015)
	EtfA	AFA47663	6e-45	+/- 2	(Weghoff *et al*, 2015)
	EtfB	AFA47662	1e-43	+/- 2	(Weghoff *et al*, 2015)
K) $F_{420}H_2$-dependent heterodisulfide reductase (Hdr2)	**Hdr2A**	AAM06247	3e-08	All[c]	(Yan *et al*, 2017)
	Hdr2B	AAM07582	1e-03	+/- 2 HdrC[c]	(Yan *et al*, 2017)
	Hdr2C	AAM07581	2e-06	+/- 2 HdrB[c]	(Yan *et al*, 2017)
L) Methylene-tetrahydrofolate (H_4F) reductase/heterodisulfide reductase (Met)	**MetF**	YP_430048	9e-13		(Mock *et al*, 2014)
	MetV	YP_430049	0.1	+/- 3	(Mock *et al*, 2014)
	MvhD	YP_430050	4e-07	+/- 3	(Mock *et al*, 2014)
	HdrA	YP_430051	3e-08	All	(Mock *et al*, 2014)
	HdrB	YP_430052	1e-03	All	(Mock *et al*, 2014)
	HdrC	YP_430053	2e-06	All	(Mock *et al*, 2014)

[a]Subunits in bold are the primary oxidoreductase catalytic subunits that were used as bait sequences when identifying Bf enzyme homologs

[b]Co-localization of *hdrABC* with *mvh*, *fdh* or *met* is not necessary for Bf function in these enzyme complexes (Bertsch *et al*, 2015; Costa *et al*, 2013; Kaster *et al*, 2011). Thus, genomes containing these subunits in any location were considered as having a Bf homolog.

[c] Hdr2A is often not co-localized with Hdr2BC [i.e., as in *Methanosarcina acetovorans* where Hdr2ABC was first described (Yan *et al*, 2017)]. Thus, Hdr2A was allowed to be located anywhere in the genome, while Hdr2B and Hdr2C were required to be co-localized for inclusion as a Bf Hdr2 complex

Table 2. Proposed taxonomic origin of Bf enzyme complexes based on phylogenetic delineation of the earliest evolving extant lineage of the oxidoreductase subunits associated with each Bf enzyme complex in complete genome sequences and metagenomes (see **Supp. Figs 3, 7, 9, 12, 14, 17, 19, 20** and **Supp. Figs. 24 to 31**). The taxonomic rank and a representative with a homolog comprising the earliest evolving, well-supported lineages, are provided. Multiple taxa are shown when no single group was strongly supported. The percent sequence identity and percent sequence coverage of each early evolving homolog to the most closely related homolog from a complete genome sequence is provided. Names of abbreviated protein complexes are provided in **Table 1.**

Bifurcating System[a]	**Complete Genomes**		**Metagenomes**	
	Earliest Evolving Lineage	**Taxon Example**	**Environment Type**	**Closely related organism(s) (% sequence identity / % sequence coverage)**
HydA	Thermotoga/ Dictyoglomi	*Thermotoga maritima/ Dictyoglomus thermus*	Soil/ Surface Water/ Hydrothermal Vents/ Springs/Groundwater	*Thermodesulfatator indicus* (49 / 97)/ *Desulfotomaculum australicum* (53 / 96)
MvhA	Methanomicrobiales	*Methanobacterium* sp.	Hydrothermal Vents/Springs	*Acetomicrobium thermoterrenum* (57 / 94)
FdhA	δ-Proteobacteria	*Desulfobacula toluolica*	Surface Water	*Desulfobacterium autotrophicum* (50 / 94)
HytA	Firmicutes	*Clostridium scatologenes*	NA	NA
NfnSL	Actinobacteria/ Proteobacteria	*Mobiluncus curtisii/ Shewanella halifaxensis*	Surface Water/Hydrothermal Vents/Springs	*Carboxydothermus islandicus* (70 / 99)
FixAB	Crenarchaeota	*Aeropyrum camini*	Hydrothermal Vents/ Springs	*Desulfotomaculum thermocisternum* (52 / 99)
Bcd	Firmicutes	*Eubacterium limosum*	Soil	*Caldisalinibacter kiritimatiensis* (70 / 99)

CarC	Firmicutes	*Natranaerobius thermophilus*	Deep Subsurface/Soil	*Sporobacter termitidis* (60 / 99)
HylA	Firmicutes	*Thermincola potens*	Saline/Hydrothermal Vents/Springs	*Caloranaerobacter ferrireducens* (70 / 99)
Ldh	Synergistetes/ Thermotoga	*Thermovirga lienii*/ *Pseudothermotoga lettingae*	Deep Subsurface	*Anaerotruncus rubiinfantis* (100 / 97) / *Desulfotignum balticum* (58 / 99)
Hdr2A	Firmicutes	*Moorella thermoacetica*	NA	NA
MetF	δ-Proteobacteria	*Desulfobaca acetoxidans*	Surface Sediments	*Candidatus* Latescibacteria (95 / 100)

[a]The subunit or suite of subunits (i.e., concatenation of NfnSL and FixAB) of oxidoreductases subjected to phylogenetic analysis is shown for each Bf system.
Abbreviations: NA, not available

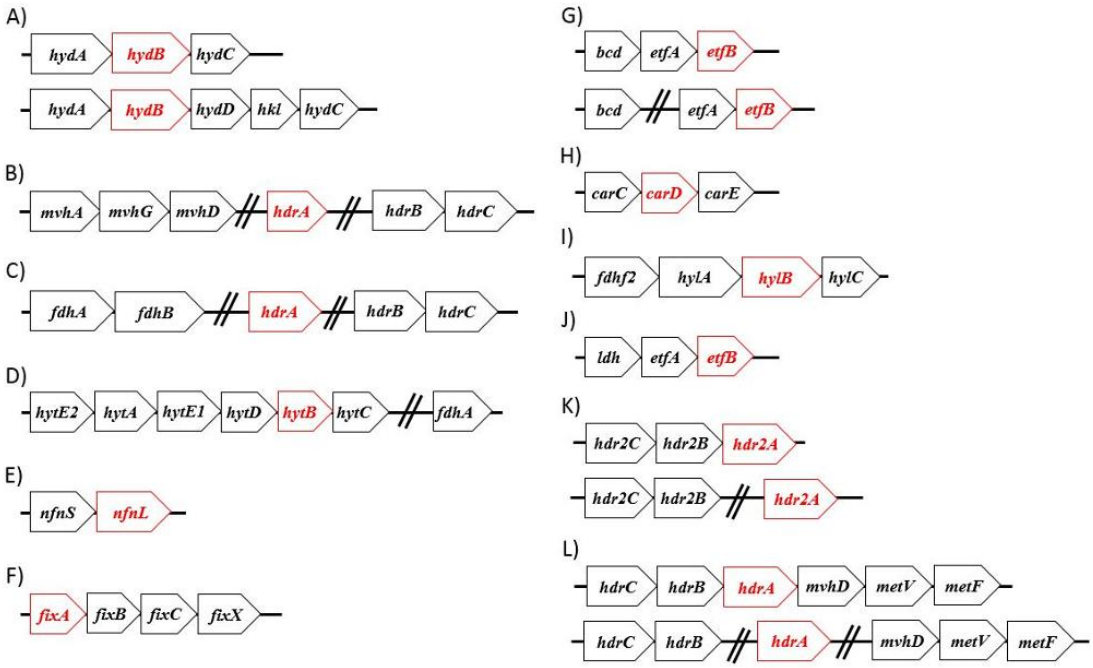


Figure 1. The most commonly identified operon structures for characterized Bf enzyme homologs: **A**) Hyd, **B**) Mvh, **C**) Fdh, **D**) Hyt, **E**) Nfn, **F**) Fix, **G**) Bf-Bcd, **H**) Car, **I**) Hyl, **J**) Bf-Ldh, **K**) Hdr2, and **L**) Met. Names of each abbreviated complex are presented in **Table 1**. Open reading frames that are highlighted in red represent genes encoding the proposed Bf flavoprotein subunit of the specified complex.

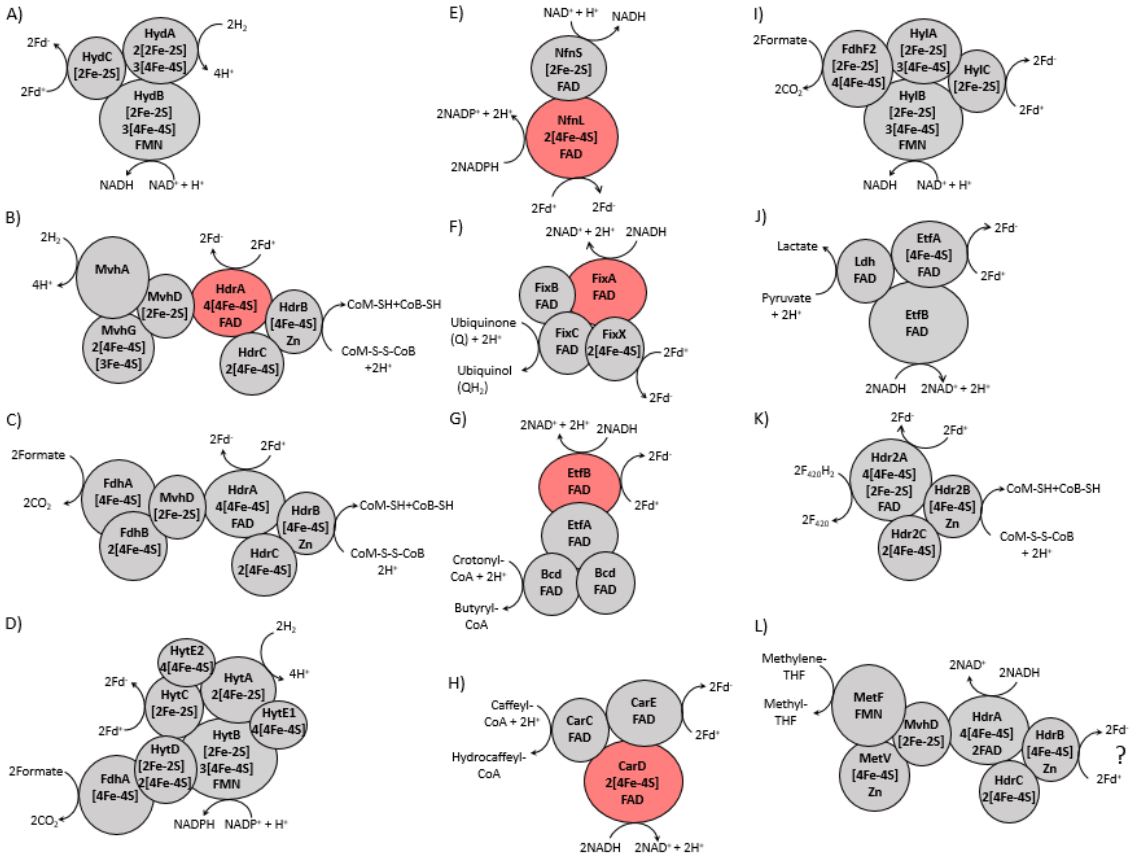


Figure 2. Proposed protein-protein interactions and the number and type of cofactors present in each protein subunit, as defined previously in references characterizing each Bf enzyme (see Table 1). A) Hyd, B) Mvh, C) Fdh, D) Hyt, E) Nfn, F) Fix, G) Bf-Bcd, H) Car, I) Hyl, J) Bf-Ldh, K) Hdr2, and L) Met. The '?' in L) indicates that the third substrate is not known for this enzyme but is predicted to be ferredoxin (Fd) (Mock et al., 2014). Names of each abbreviated complex are presented in Table 1. The reversibility of the enzyme complexes is not shown to better depict coupling of substrate oxidation/reduction. Subunits that are highlighted red are the flavoprotein subunits that have been biochemically shown to be the site of electron bifurcation.

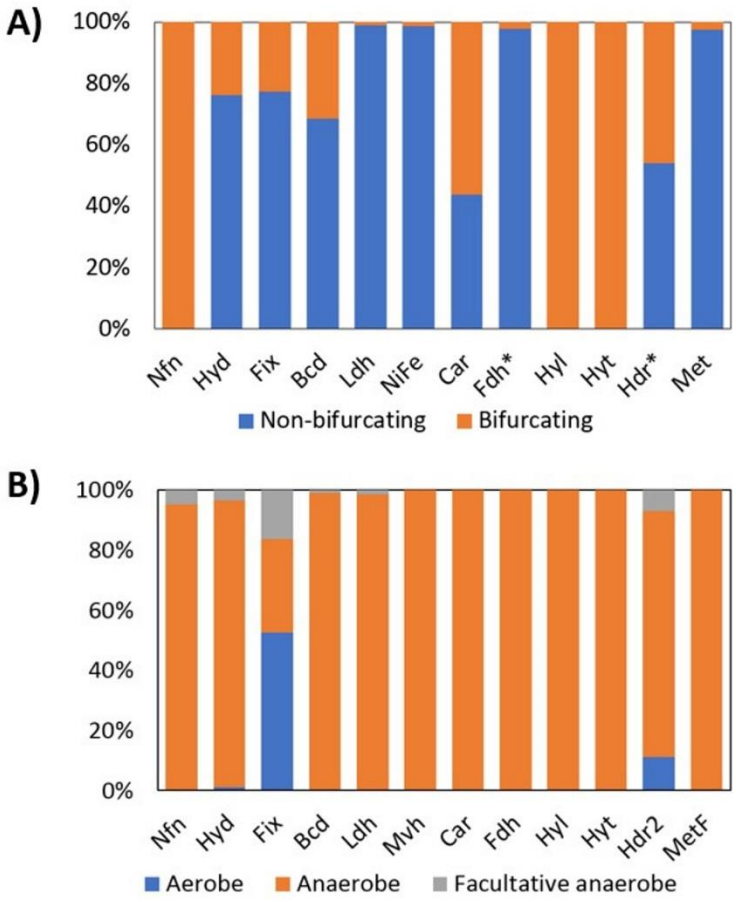


Fig. 3. Distribution of homologs of putative Bf and non-Bf enzymes in complete genome sequences. **A**) Histogram depicting the percent of homologs that are predicted to bifurcate electrons and those that are predicted to not bifurcate electrons in complete genomes. **B**) Percent of enzyme homologs predicted to bifurcate electrons that are derived from the genomes of aerobes, anaerobes, or facultative anaerobes. Fdh* and Hdr* includes all the Bf enzymes that form a complex with Fdh (i.e., Hyt, Hyl and Hdr associated Fdh) and Hdr (i.e., Fdh associated Hdr, Mvh, Hdr2 and Met), respectively. Names of each abbreviated enzyme complex are provided in **Table 1**. Note, Mvh are the only known Bf subclass of [NiFe]-hydrogenase (labeled [NiFe] in **Panel A**).

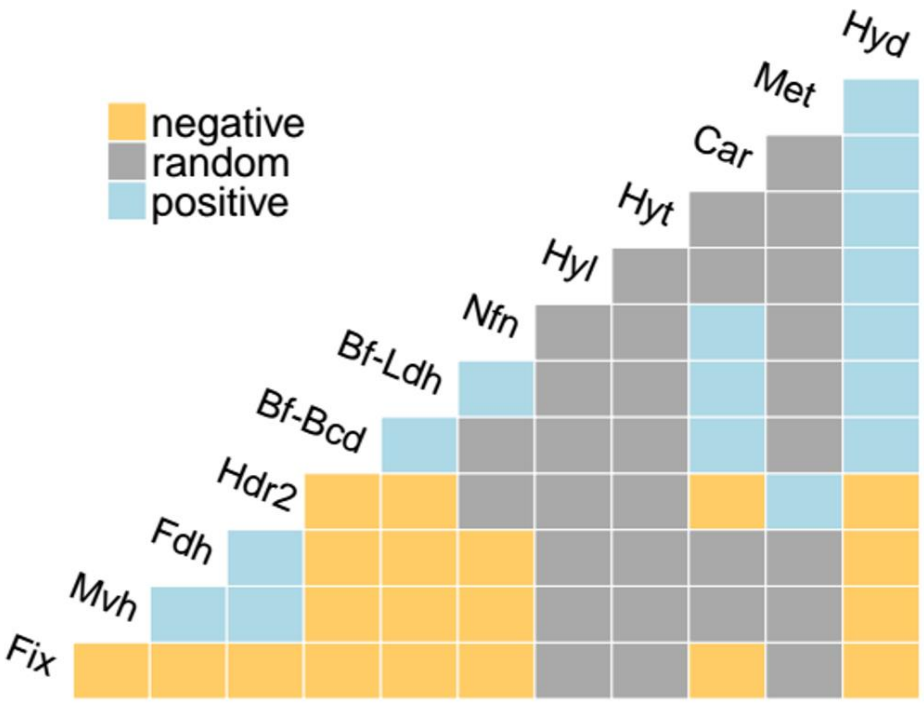


Fig. 4. Co-occurrence analysis of the distribution of homologs of Bf enzymes in complete genome sequences. Positive co-occurrences (exceeding the 95% statistical significance threshold) are represented in cyan, negative co-occurrences are represented in yellow, and random co-occurrences are represented in grey. Abbreviations for enzyme complexes are provided in **Table 1**.

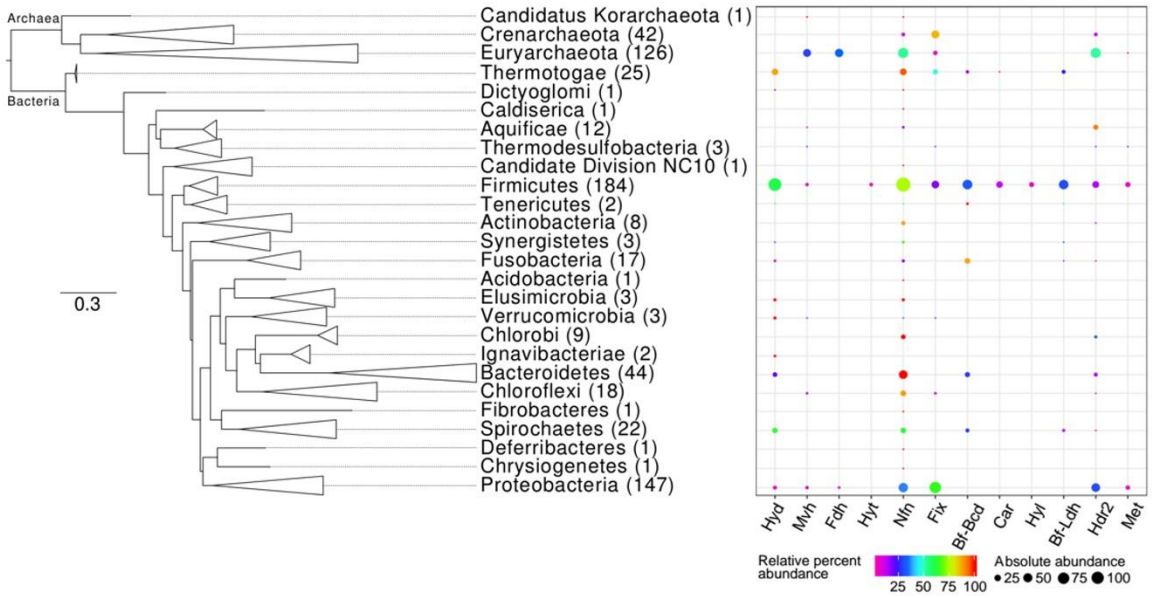


Figure 5. Phylogenetic distribution of homologs of enzymes that are predicted to bifurcate electrons in complete genome sequences. The Maximum-Likelihood phylogenetic tree was constructed using 16S rRNA gene sequences obtained from genomes that also encode at least one homolog of a Bf enzyme. Phylum-level lineages were collapsed to conserve space. The number in parentheses next to each phylum level designation indicates the total number of genomes within a specified lineage that code for homologs of at least one of the twelve Bf enzymes (**Supp. Table 2**). The percent of genomes within a given lineage that code for a homolog of the specified Bf enzyme (i.e., relative percent abundance) is indicated by the color of the bubble. The percent of the total number of homologs of the specified Bf enzyme that are encoded in genomes affiliated with the specified lineage (i.e., absolute percent abundance) is indicated by the size of the bubble. Names for each abbreviated enzyme complex are provided in **Table 1**.

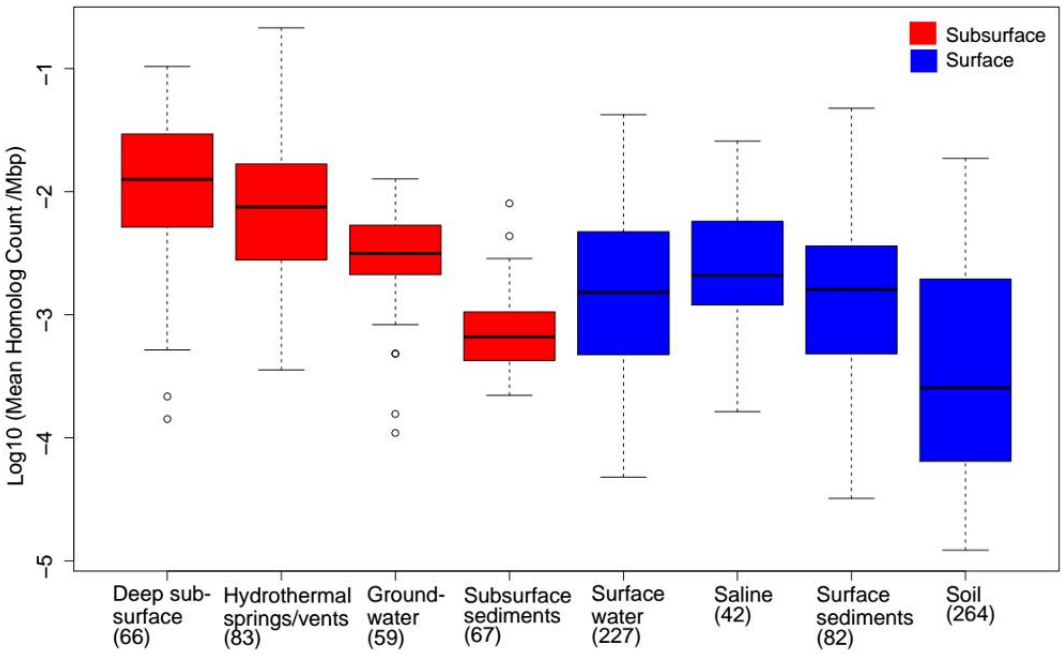


Fig. 6. Average abundance of homologs of the eleven (see **Table 1**) Bf enzyme complexes in metagenomic sequences (Hyt homologs were not identified in metagenomes and therefore are not included for this analysis). Here, the average abundance of homologs of all Bf enzyme complexes identified among non-redundant metagenomic contigs was normalized to the total number of megabase (Mb) pairs of sequence in those contigs to account for differences in sequence depth between datasets. Metagenomes were classified into different environmental groups based on metadata provided for each dataset (see **Supp. Table 3**). The number in parentheses next to the environmental groups, as specified on the x-axis, represents the total number of metagenomes classified into a given environmental group. Only environmental (e.g. non-engineered and non-host-associated) metagenomes were used in the analyses.

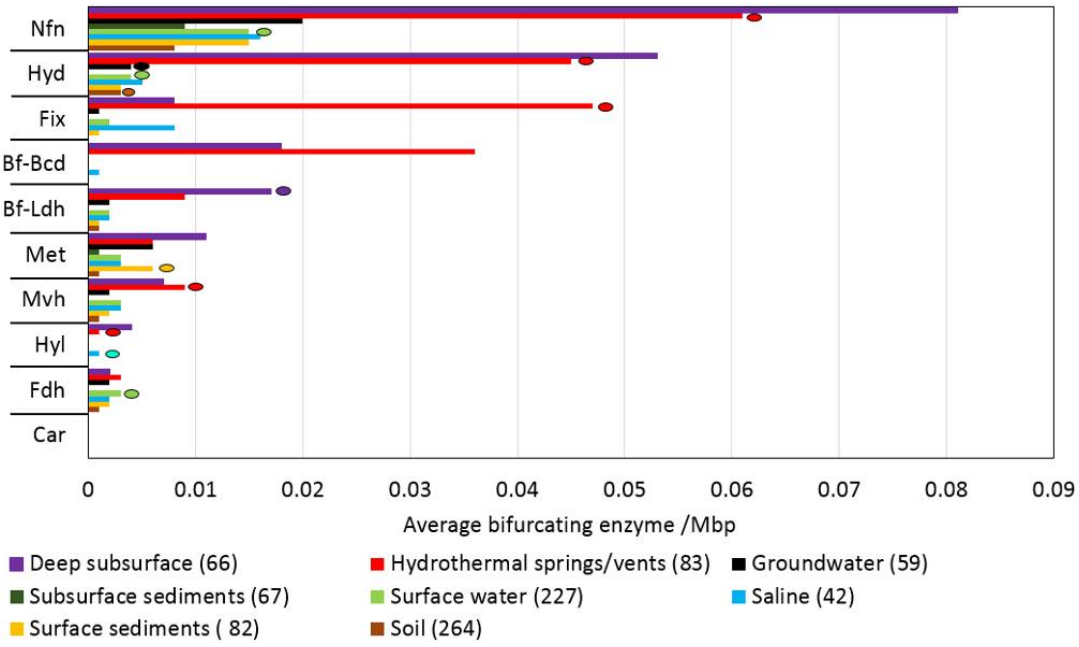


Fig. 7. Average abundance of homologs of the ten (see **Table 1**) Bf enzyme complexes within metagenomes (n=928 metagenomes). Hyt homologs were not identified in metagenomes and therefore are not depicted in the graph. Hdr homologs are difficult to demarcate from the ones that forms a complex with Mvh, Fdh or Met so were not included in metagenome analysis. The average abundances of homologs of all Bf enzyme complexes in non-redundant metagenomic contigs were normalized to the total number of megabase (Mb) pairs of sequence in those metagenomes to account for differences in sequence depth between metagenomes. The colored circle in the graph represents the environment (colored accordingly) that contained the early evolving homologs of specific Bf enzymes (also included in **Table 2**). Only a small number of Bf-Bcd and Car homologs were identified in metagenomes and thus, resolving the environment of their presumed origin was deemed not appropriate from their phylogenies. Metagenomes were classified using metadata provided for each dataset. The total number of metagenomes classified into each environmental group is given in parentheses. Names for each abbreviated protein complex are provided in **Table 1**.

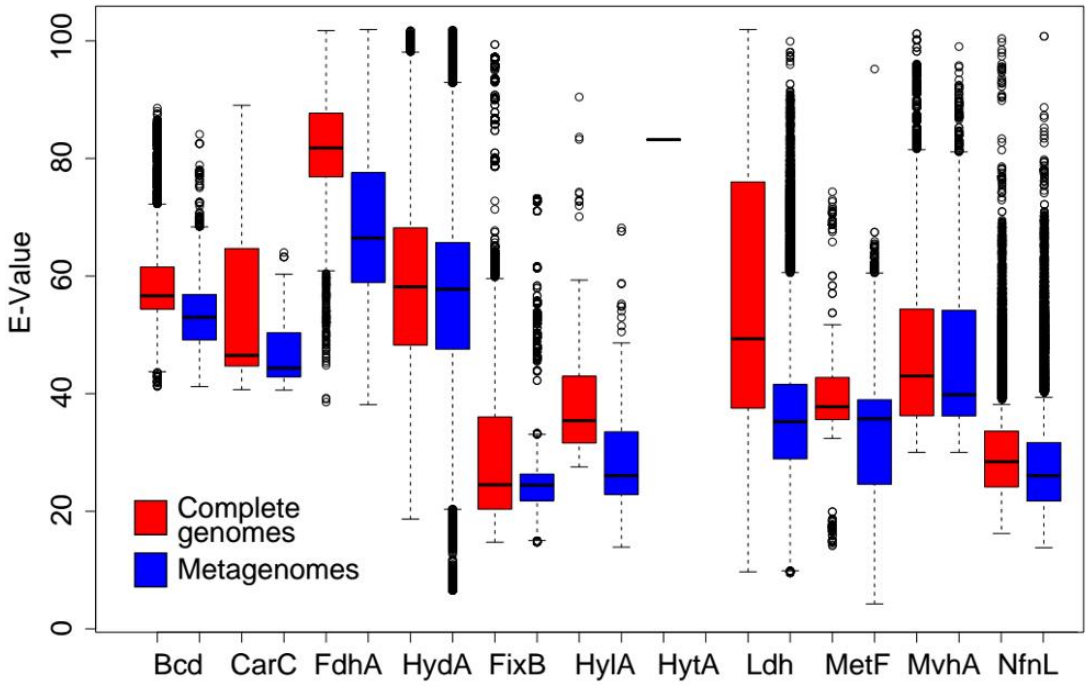
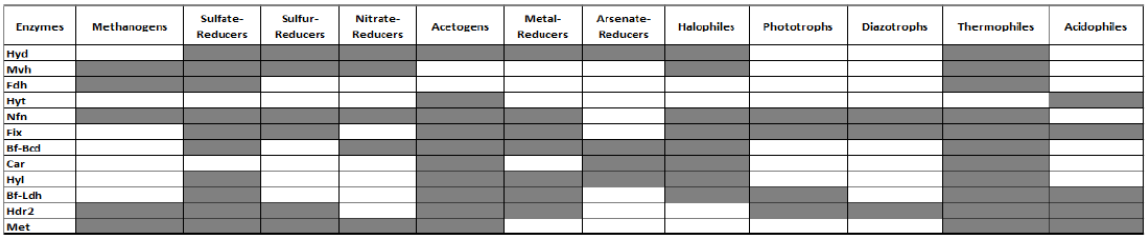
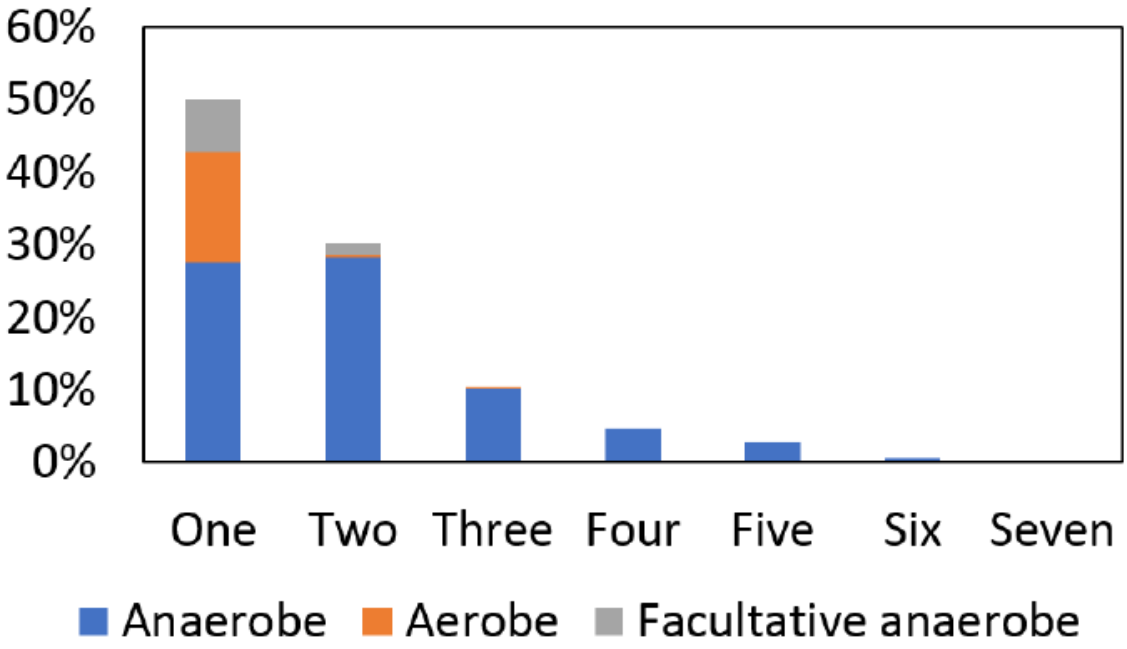


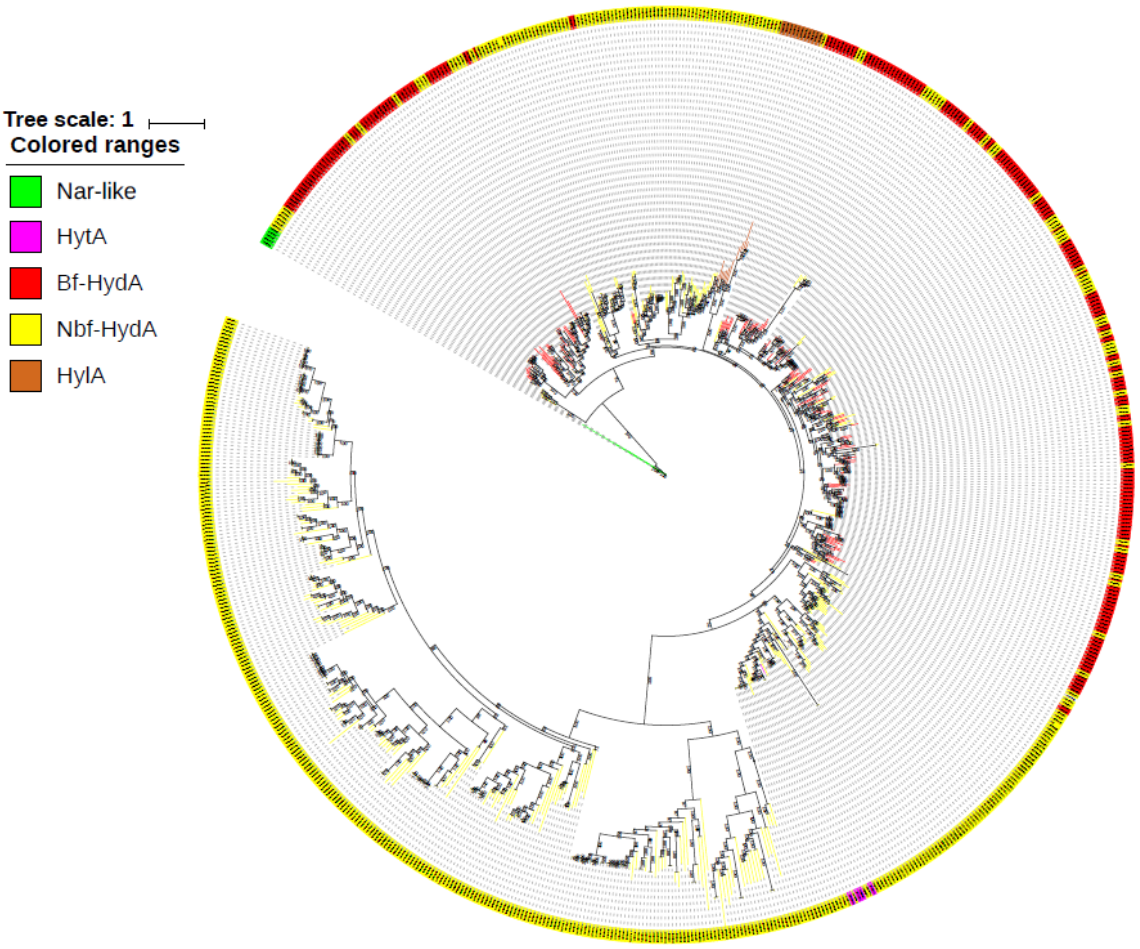
Figure 8. Distribution of e-values of homologs of the catalytic oxidoreductase subunit of the eleven (see **Table 1**; Hyt was not identified in our metagenomic sequence database) Bf enzyme complexes identified in metagenome (in red) and complete genome (in blue). Homologs identified in genome sequences were subjected to pairwise phmmer analysis to obtain e-values as an indication of their diversity. Similarly, homologs identified in metagenome sequences were subjected to phmmer against a database comprising homologs obtained from genomes to assess the extent of the diversity of sequences in taxa that are either yet to be cultivated or that lack genome sequences. To simplify data presentation, the e-values were then normalized by multiplying by -10000 which is depicted as 'E-Value' on the y-axis. Higher E-Values indicate that homologs exhibit a lower diversity and thus are more similar phylogenetically. Low E-Values indicate that homologs are more diverse and are less similar phylogenetically. The box represents the interquartile range with the whiskers show the full range of the data. The circles outside of those whiskers represent outliers and the horizontal black bold line within the box represents the median value. Homologs identified in metagenomes that exhibit a lower average or range of E-values than those identified in complete genomes are underrepresented in complete genome databases (i.e., there is unsampled diversity).



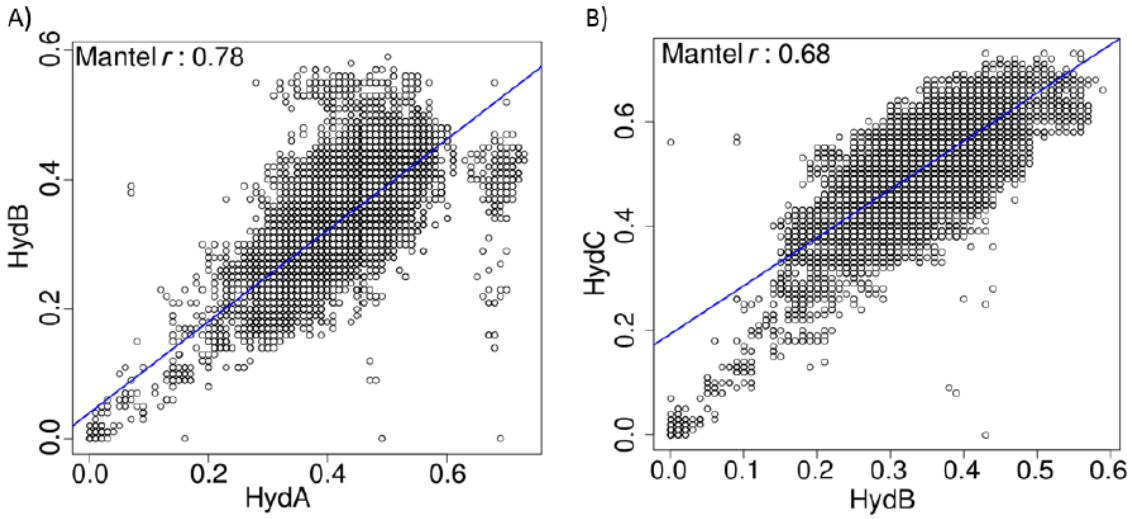
Supplementary Figure 1. Presence (grey) or absence (white) of homologs of one or more of the twelve Bf enzyme complexes encoded in the genomes of organisms with specified physiological attributes.



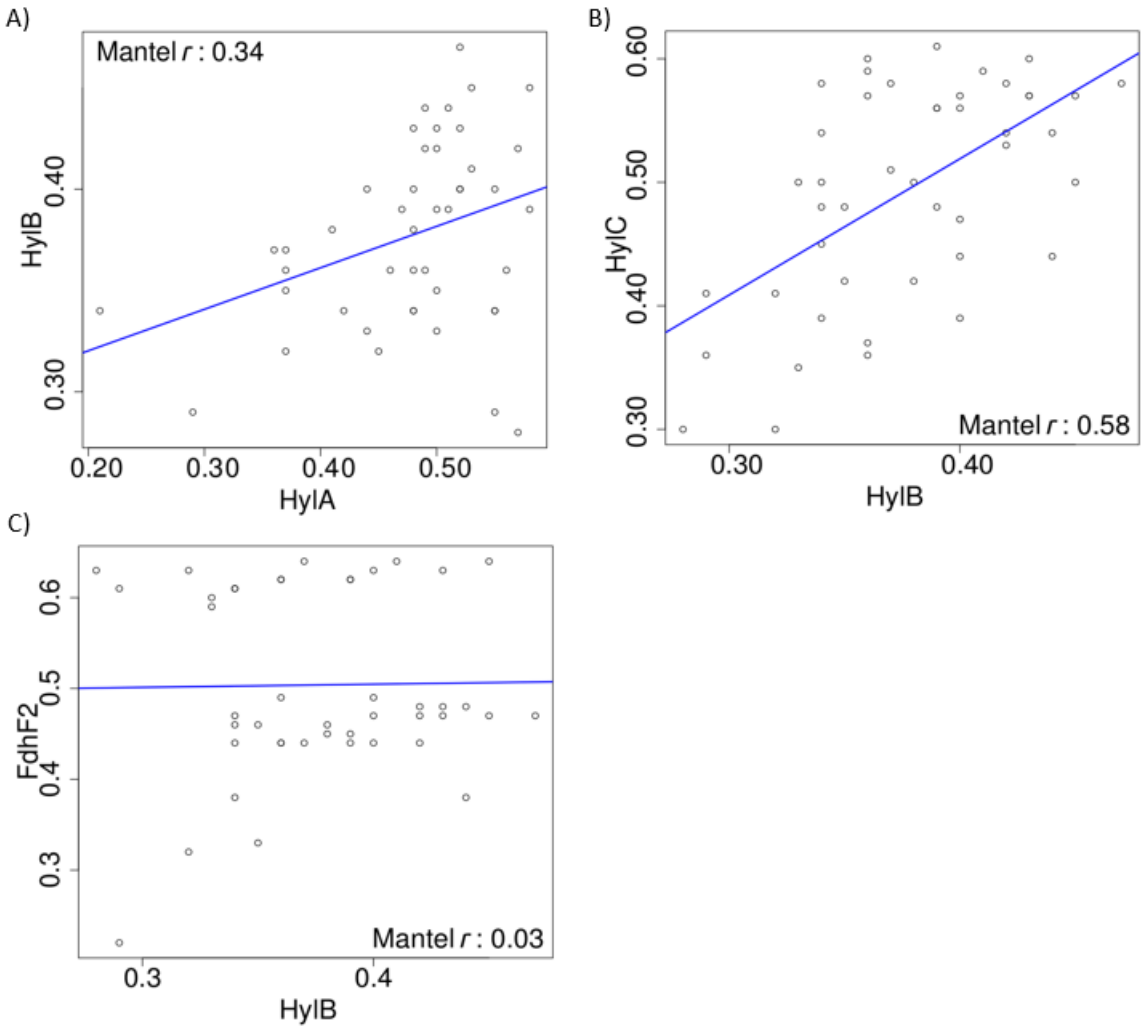
Supplementary Figure 2. Histogram depicting the percent of complete genome sequences that encode homologs of one to seven Bf enzyme complexes, out of 681 genomes found to contain at least one Bf enzyme complex. The bar chart also shows the distribution of these genomes as they relate to the ability of the host organism to integrate oxygen into their energy metabolism.



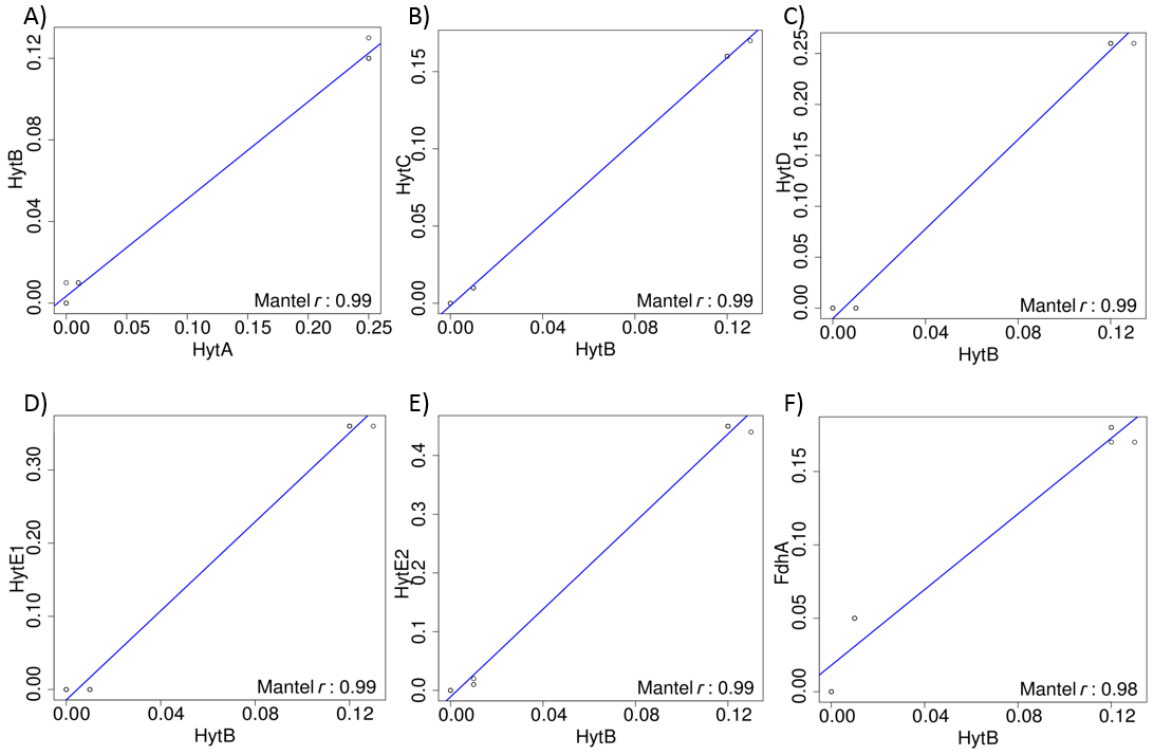
Supplementary Figure 3. Maximum-likelihood phylogenetic reconstruction of Bf [(Bf)-HydA, HytA, and HylA] and non-Bf HydA homologs [(Nbf)-HydA] and Nar-like proteins in complete genome sequences. The phylogeny was rooted with Nar-like proteins from *Homo sapiens* (NP_071938 and NP_036468), *Danio rerio* (A2RRV9), *Thalassiosira pseudonana* (XP_002289272) and *Ostreococcus lucimarinus* (XP_001416706). Sequence terminals are color coded, with Bf-HydA in red, HytA in purple, HylA in rust, Nbf-HydA in yellow, and Nar-like homologs in green. Names for each abbreviated protein complexes are provided in **Table 1**. Note, homologs of Bf-HydA, HytA, HylA were only identified among Bacteria, thereby negating the need to add a colored strip demarcating archaeal and bacterial homologs. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



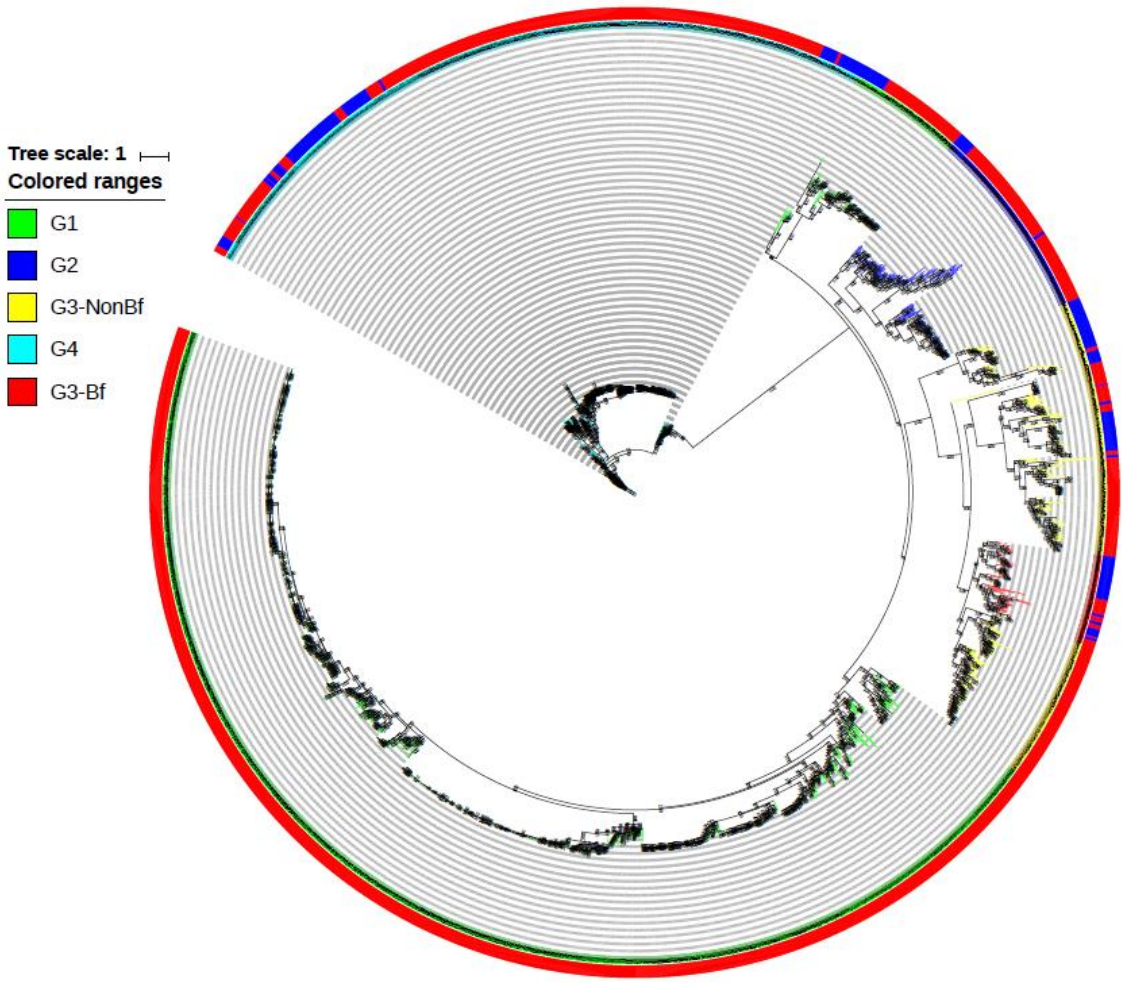
Supplementary Figure 4. Mantel regression of the pairwise distances of homologs of subunits of the Hyd complex plotted as a function of each other: **A**) HydA and HydB, **B**) HydB and HydC. Mantel *r* values are provided and the significance values (*p*-value) for both regressions were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1**.



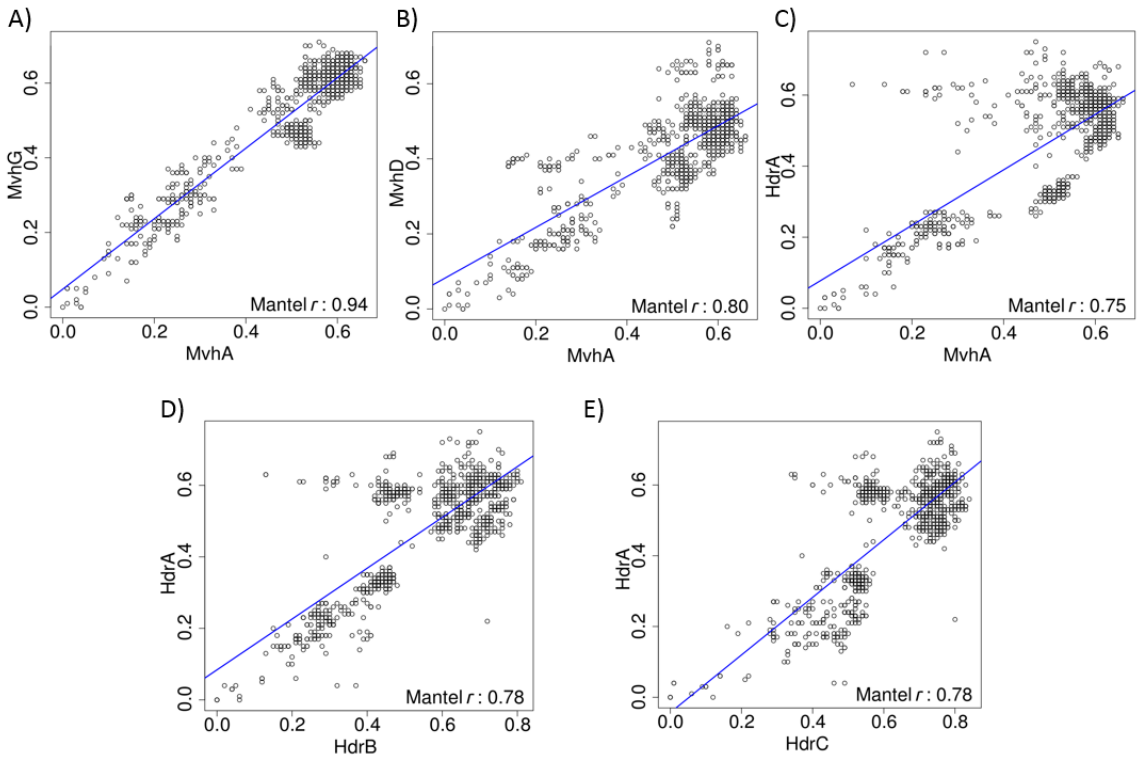
Supplementary Figure 5. Mantel regression of the pairwise distances of homologs of subunits of the Hyl complex plotted as a function of each other: **A**) HylA and HylB, **B**) HylB and HylC and **C**) HylB and FdhF2. Mantel *r* values are provided and the significance values (*p*-value) for all regressions were <0.02. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



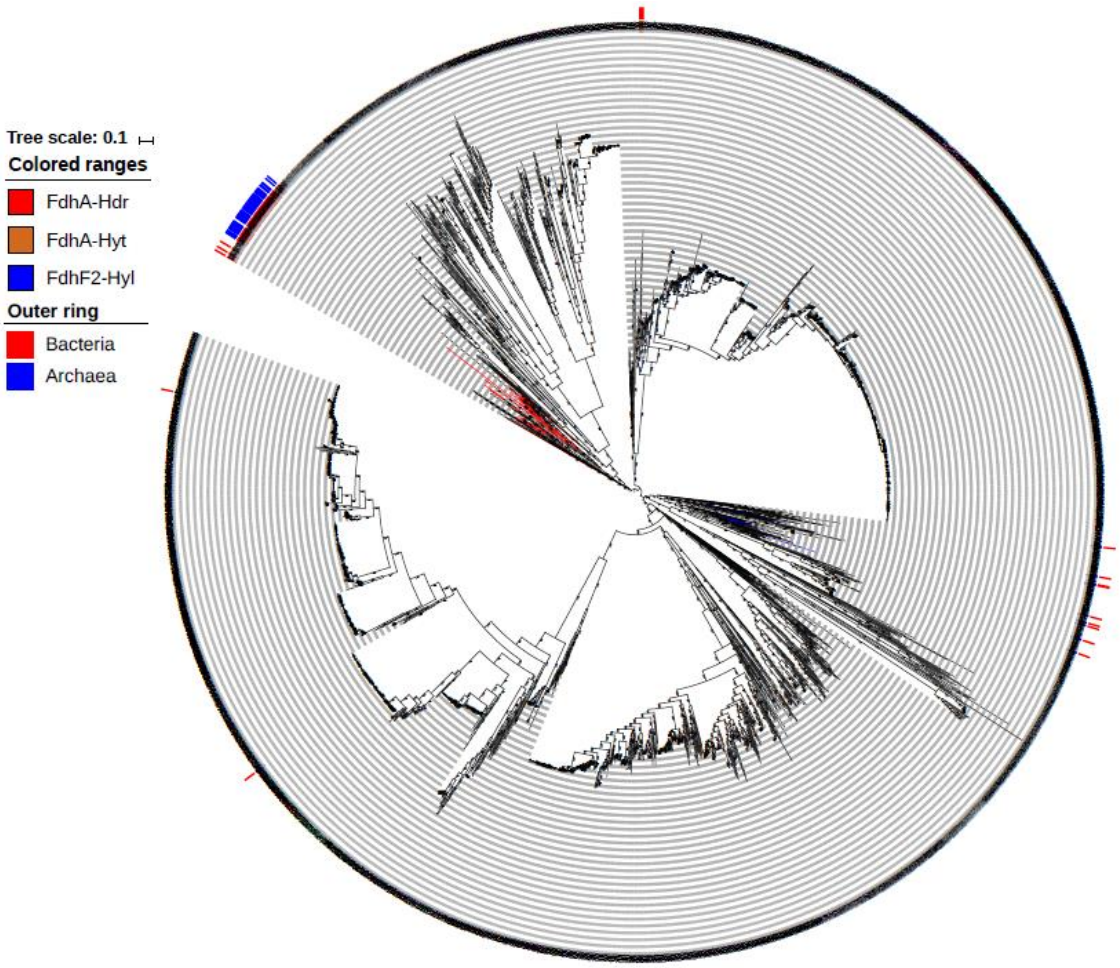
Supplementary Figure 6. Mantel regression of the pairwise distances of homologs of subunits of the Hyt complex plotted as a function of each other: **A**) HytA and HytB, **B**) HytB and HytC, **C**) HytB and HytD, **D**) HytE1 and HytB, **E**) HytE2 and HytB, and **F**) HytB and FdhA. Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.2. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



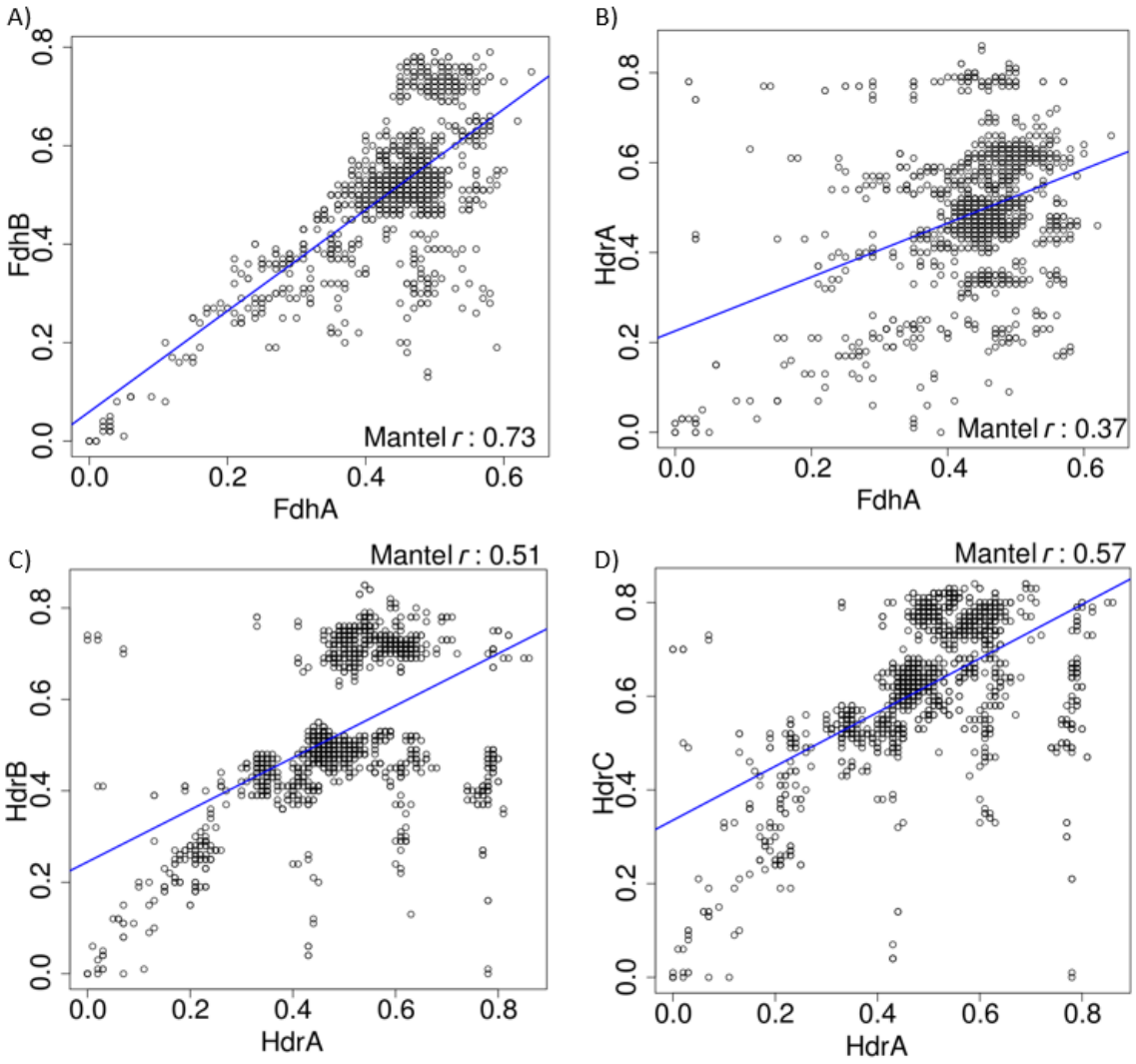
Supplementary Figure 7. Maximum-likelihood phylogenetic reconstruction of homologs of archaeal (blue in outer strip) and bacterial (red in outer strip) large subunits of Bf (i.e., MvhA that form group 3c) and non-Bf (groups 1, 2, 3a, 3b, 3d, and 4) [NiFe]-hydrogenase in complete genomes extracted from (Boyd et al., 2014). The phylogeny was by default rooted to the large subunit of group 4 by the RAxML which was used as the root to evaluate the evolution of MvhA homologs. The group designations for [NiFe]-hydrogenase are defined as in (Vignais et al., 2001; Boyd et al., 2014) and are represented by different colors at sequence terminals (inner ring and as projected on lineages). Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



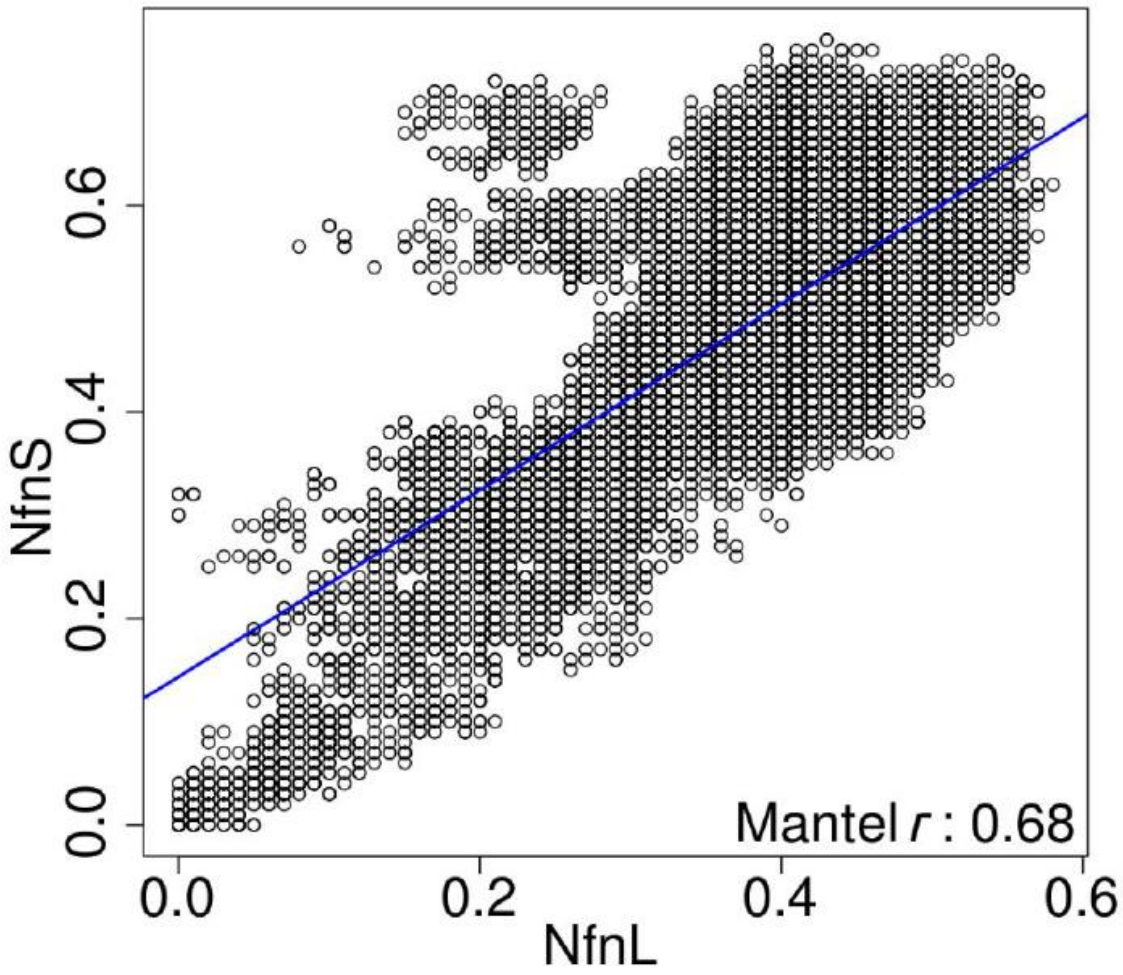
Supplementary Figure 8. Mantel regression of the pairwise distances of homologs of subunits of the Mvh complex plotted as a function of each other: **A**) MvhA and MvhG, **B**) MvhA and MvhD, **C**) MvhA and HdrA (only associated with Mvh complex), **D**) HdrA and HdrB (both associated with only Mvh complex), and **E**) HdrC and HdrA (both associated with only Mvh complex). Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



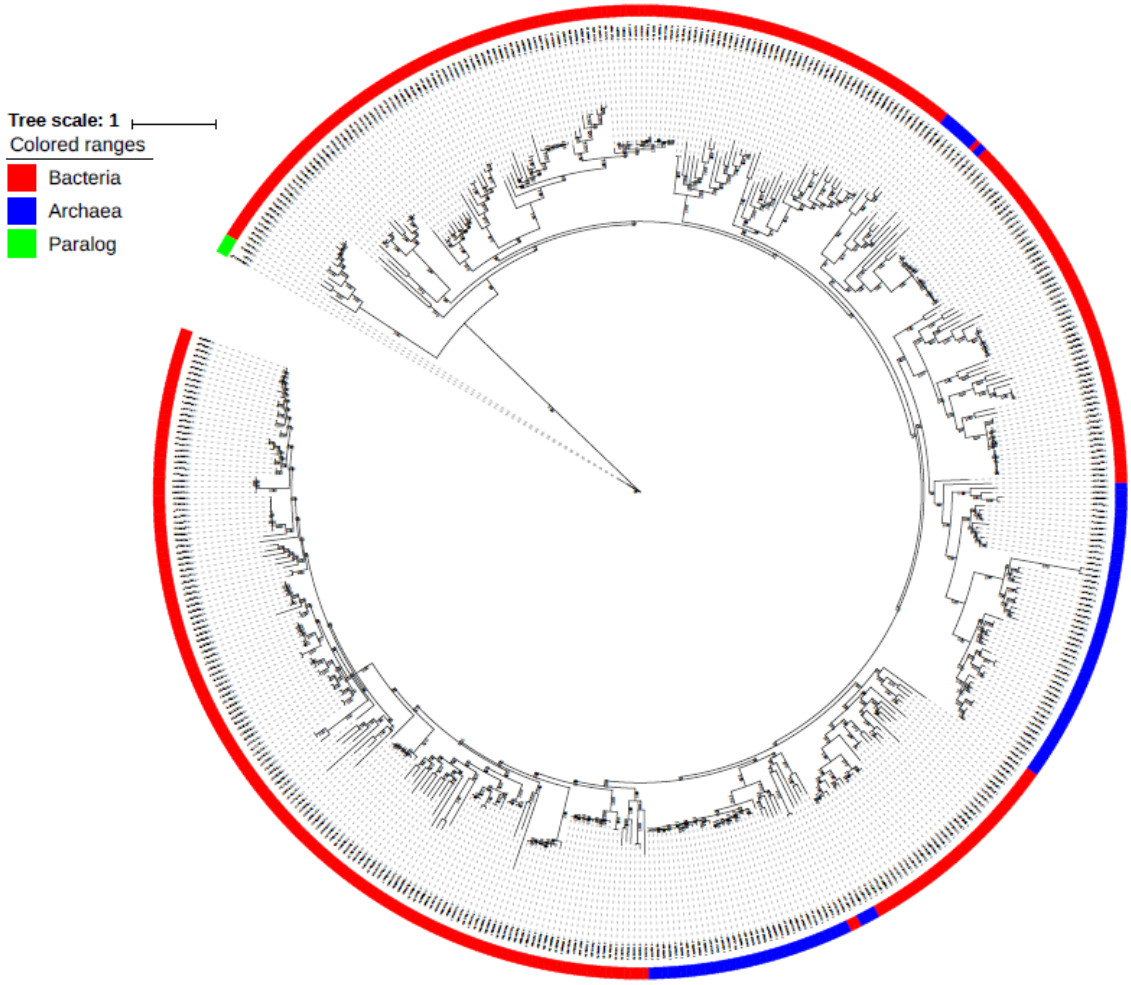
Supplementary Figure 9. Maximum-likelihood phylogenetic reconstruction of archaeal (blue in the outer strip) and bacterial (red in outer strip) Bf (FdhA-Hdr, FdhA-Hyt, and FdhF2-Hyl) and non-Bf FdhA homologs from complete genome sequences. Mid-point rooting was used since FdhA and FdhF2 are paralogous. Sequence terminals of homologs of bifurcating FdhA are colorcoded, with FdhA-Hdr in red, FdhA-Hyt in rust, and FdhF2-Hyl in blue. Sequence terminals that are not colored depict FdhA homologs that are not predicted to bifurcate. Names for each abbreviated protein complexes are provided in **Table 1**. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



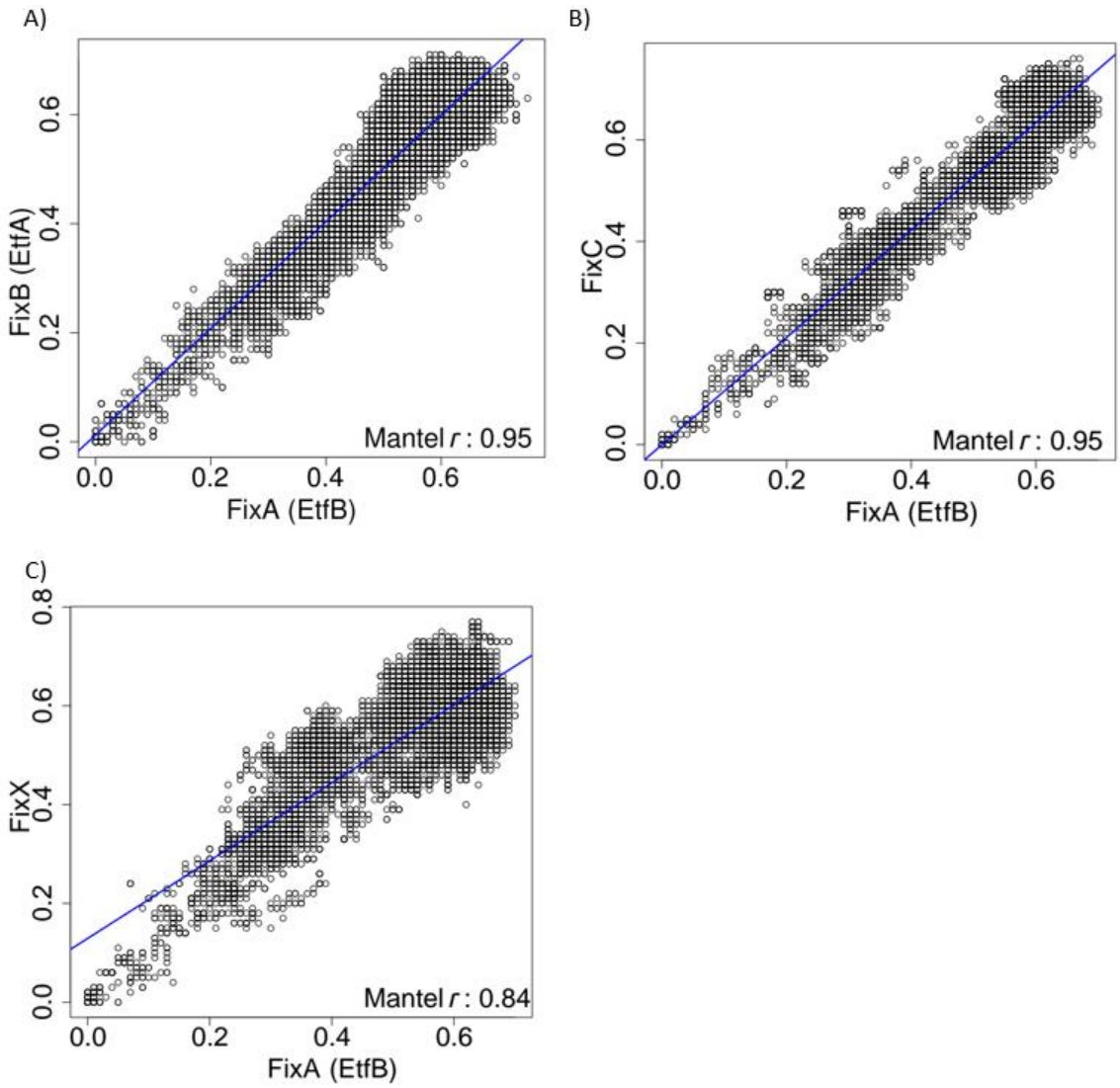
Supplementary Figure 10. Mantel regression of the pairwise distances of homologs of subunits of the Fdh complex plotted as a function of each other. **A**) FdhA and FdhB, **B**) FdhA and HdrA (only associated with Fdh complex), **C**) HdrB and HdrA (both only associated with Fdh complex), and **D**) HdrA and HdrC (both only associated with Fdh complex). Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



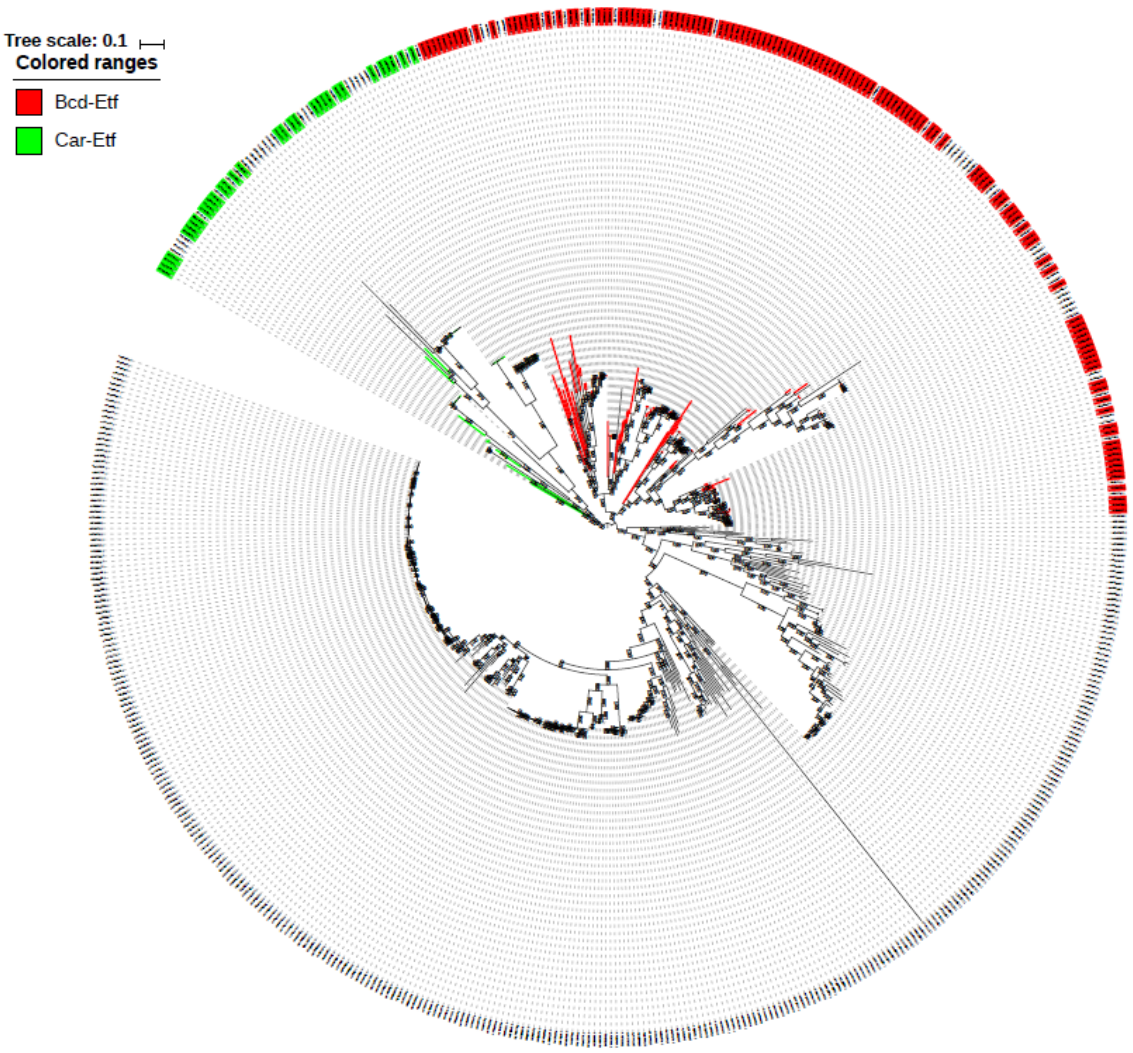
Supplementary Figure 11. Mantel regression of the pairwise distances of homologs of NfnS as a function of the pairwise distances of homologs of NfnL. The Mantel *r* value is provided and the significance value (*p*-value) for the regression shown was <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1**.



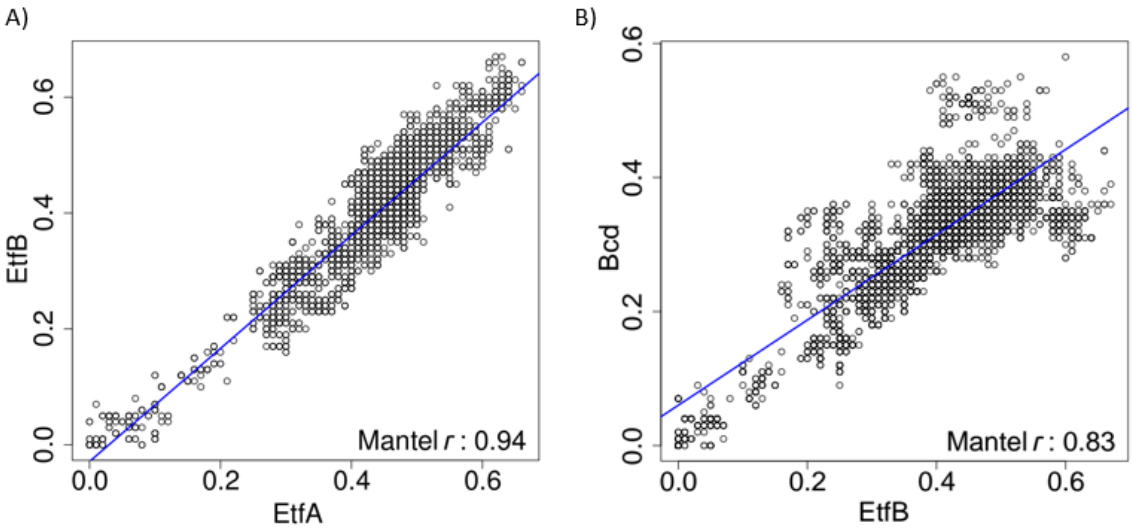
Supplementary Figure 12. Maximum-likelihood phylogenetic reconstruction of a concatenation of archaeal (blue in outer strip) and bacterial (red in outer strip) NfnSL in complete genome sequences rooted (green outer strip) with concatenated paralogous proteins which include dihydroorotate dehydrogenase from *Lactococcus lacticus* (WP_011835013) and glutamate synthase from *Azospirillum brasilense* (WP_035677957), dihydroorotate dehydrogenase from *Lactococcus garvieae* (BAK58851) and glutamate synthase from *Azospirillum oryzae* (WP_085087092), and dihydroorotate dehydrogenase from *Floricoccus tropicus* (WP_070791886) and glutamate synthase from *Azospirillum humicireducens* (WP_063635528). Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



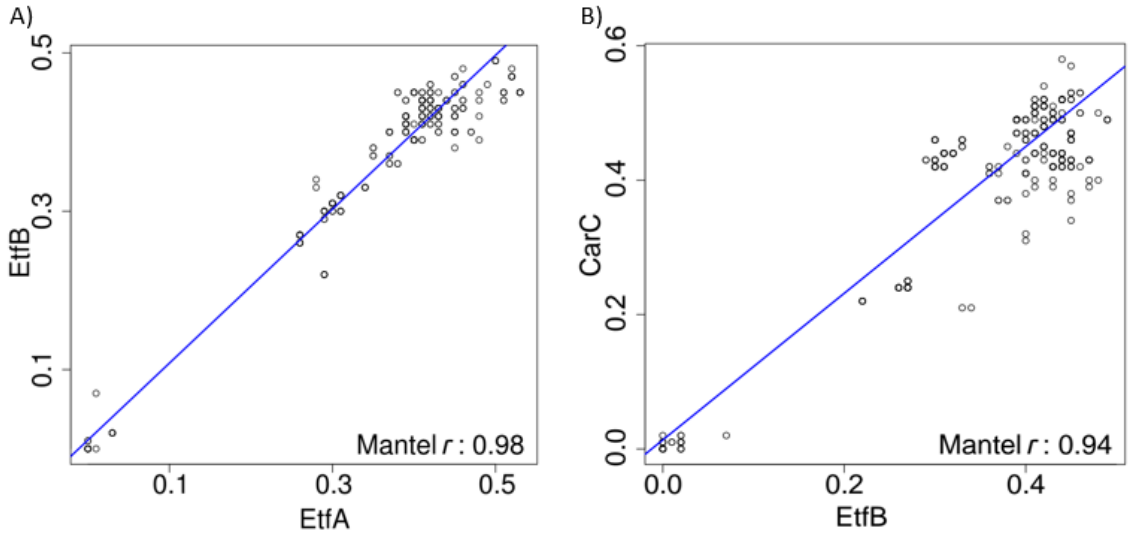
Supplementary Figure 13. Mantel regression of the pairwise distances of homologs of subunits of the Fix complex plotted as a function of each other. **A**) FixA (or only EtfB associated with Fix complex) and FixB (or only EtfA associated with Fix complex), **B**) FixC and FixA (or only EtfB associated with Fix complex), **C**) FixX and FixA (or only EtfB associated with Fix complex), Mantel *r* values are provided and the significance values (*p*value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



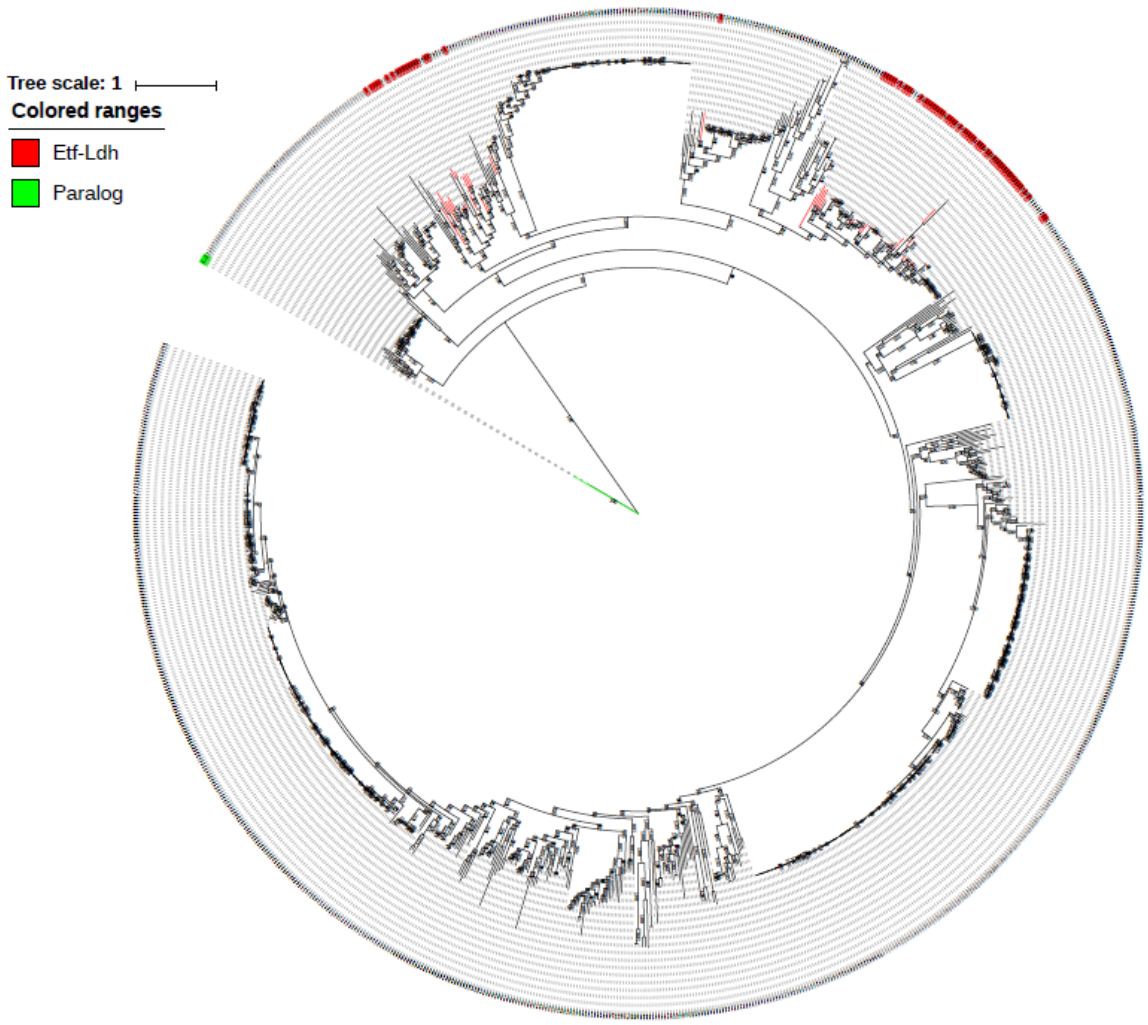
Supplementary Figure 14. Maximum-likelihood phylogenetic reconstruction of butyryl-CoA dehydrogenase (Bcd) (red outer strip) and CarC (green outer strip) subunits that are associated with Etf in complete genomes and proposed to bifurcate and those that are not associated with Etf and thus are predicted to not bifurcate (terminals are uncolored). Mid-point rooting was used since Bcd and CarC are paralogous. Names for each abbreviated protein complexes are provided in **Table 1**. Note, homologs of Bcd and CarC subunits associated with Etf were only detected among Bacteria, thereby negating the need to add a colored strip demarcating archaeal and bacterial homologs. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



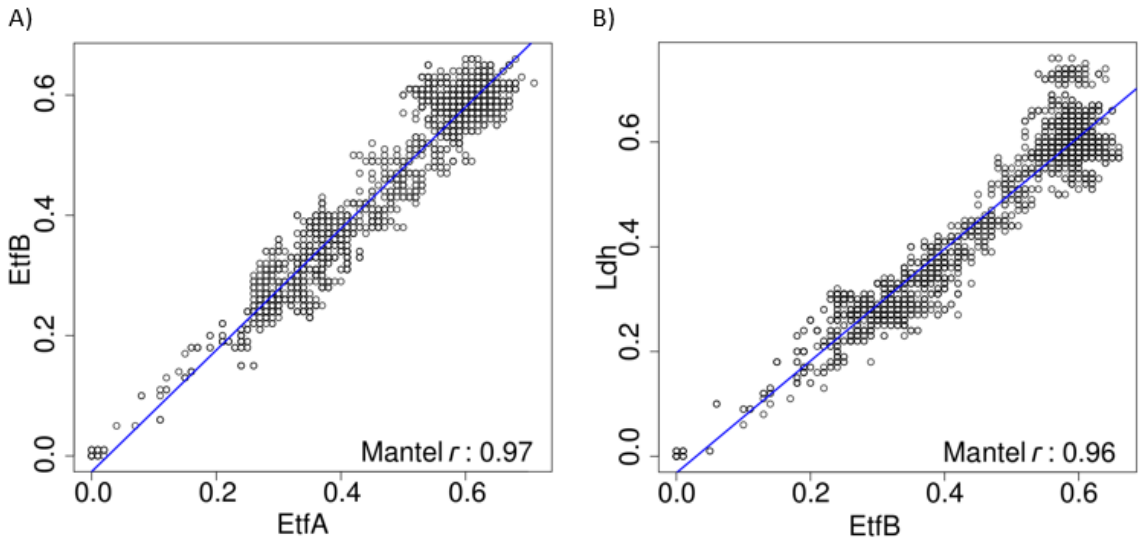
Supplementary Figure 15. Mantel regression of the pairwise distances of homologs of subunits of the Bf-Bcd complex plotted as a function of each other: **A**) EtfB and EtfA (both of which are associated with only Bf-Bcd complex), **B**) EtfB (associated with only Bf-Bcd complex) and Bcd. Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



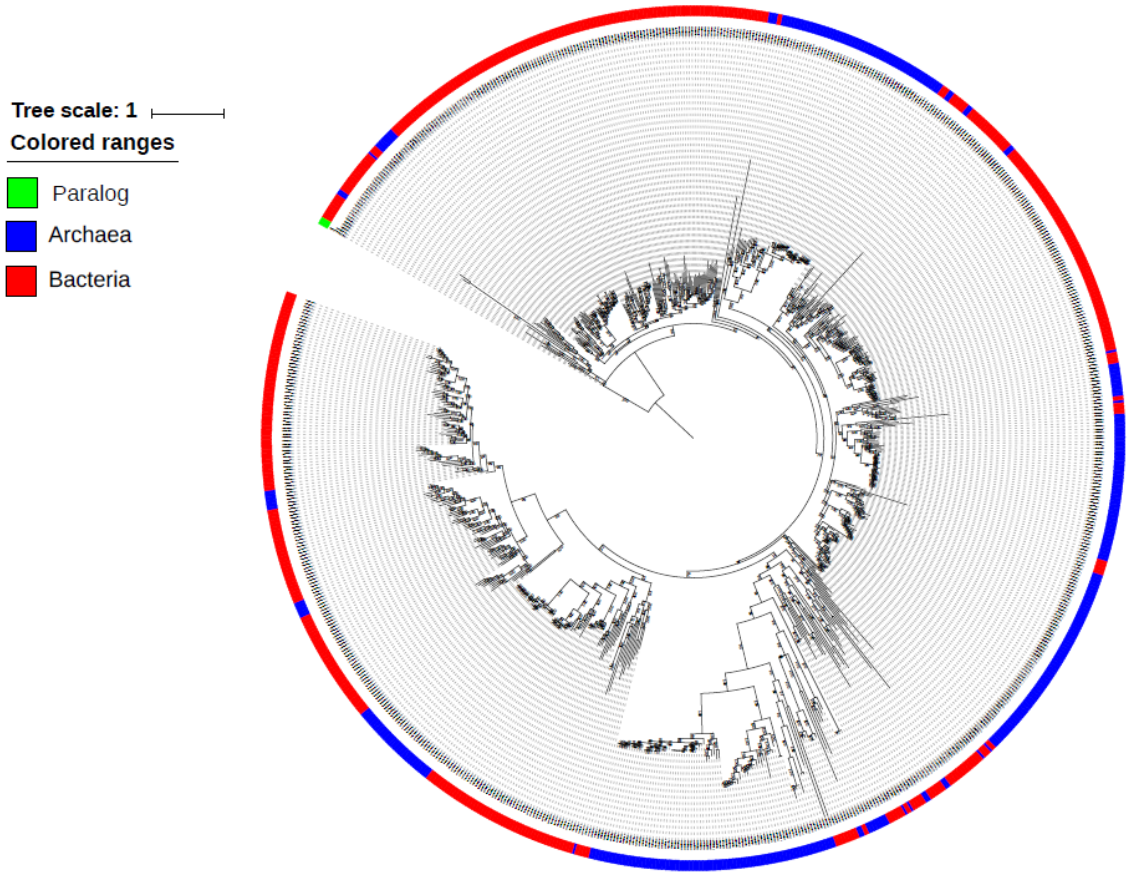
Supplementary Figure 16. Mantel regression of the pairwise distances of homologs of subunits of the Car complex plotted as a function of each other: **A**) EtfB and EtfA (both of which are associated with only Car complex), **B**) EtfB (associated with only Car complex) and CarC. Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



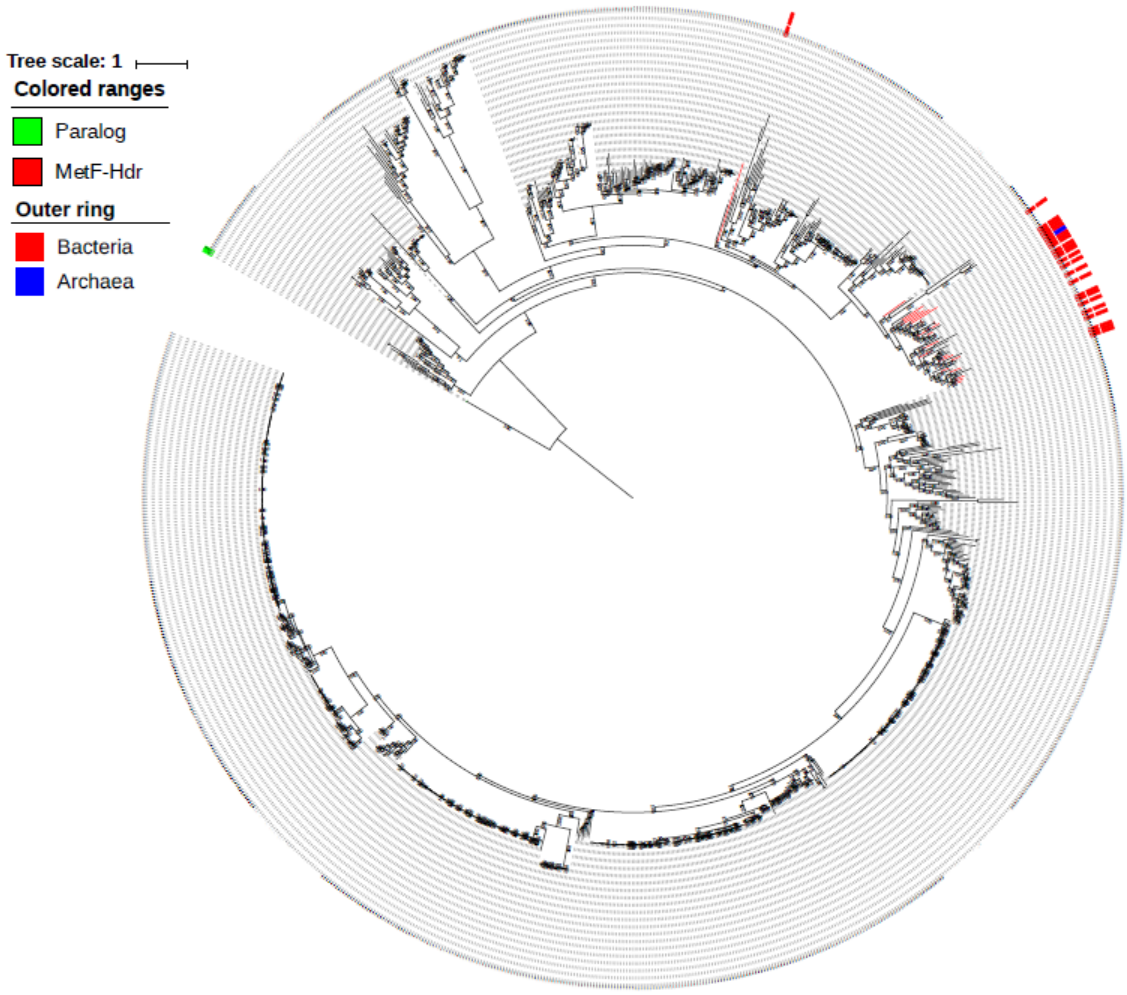
Supplementary Figure 17. Maximum-likelihood phylogenetic reconstruction of lactate dehydrogenase (Ldh) homologs that associate with Etf in complete genomes and are predicted to bifurcate (terminals colored in red, designated as Etf-Ldh) and those that do not associate with Etf and thus are not predicted to bifurcate (terminals are uncolored). The tree was rooted with the paralog (terminals colored in green) alkyl dihyroxyacetone phosphate synthase from *Dictyostelium discoideum* (XP_637836) and *Trypanosoma brucei* (XP_845272). Names for each abbreviated protein complexes are provided in **Table 1**. Note, homologs of Ldh associated with Etf were only detected among Bacteria, thereby negating the need to add a colored strip demarcating archaeal and bacterial homologs. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



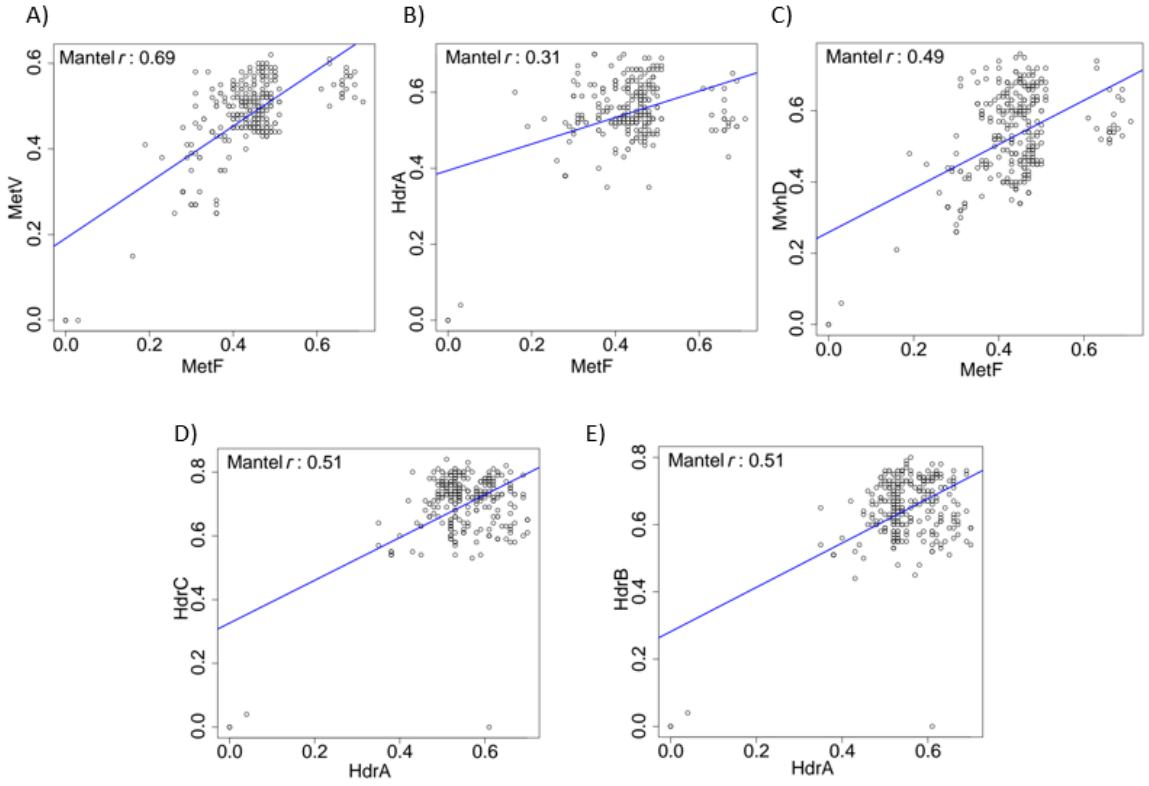
Supplementary Figure 18. Mantel regression of the pairwise distances of homologs of subunits of the Bf-Ldh complex plotted as a function of each other: **A**) EtfB and EtfA (both of which are associated with only Bf-L dh complex), **B**) EtfB (associated with only Bf-Ldh complex) and Ldh. Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



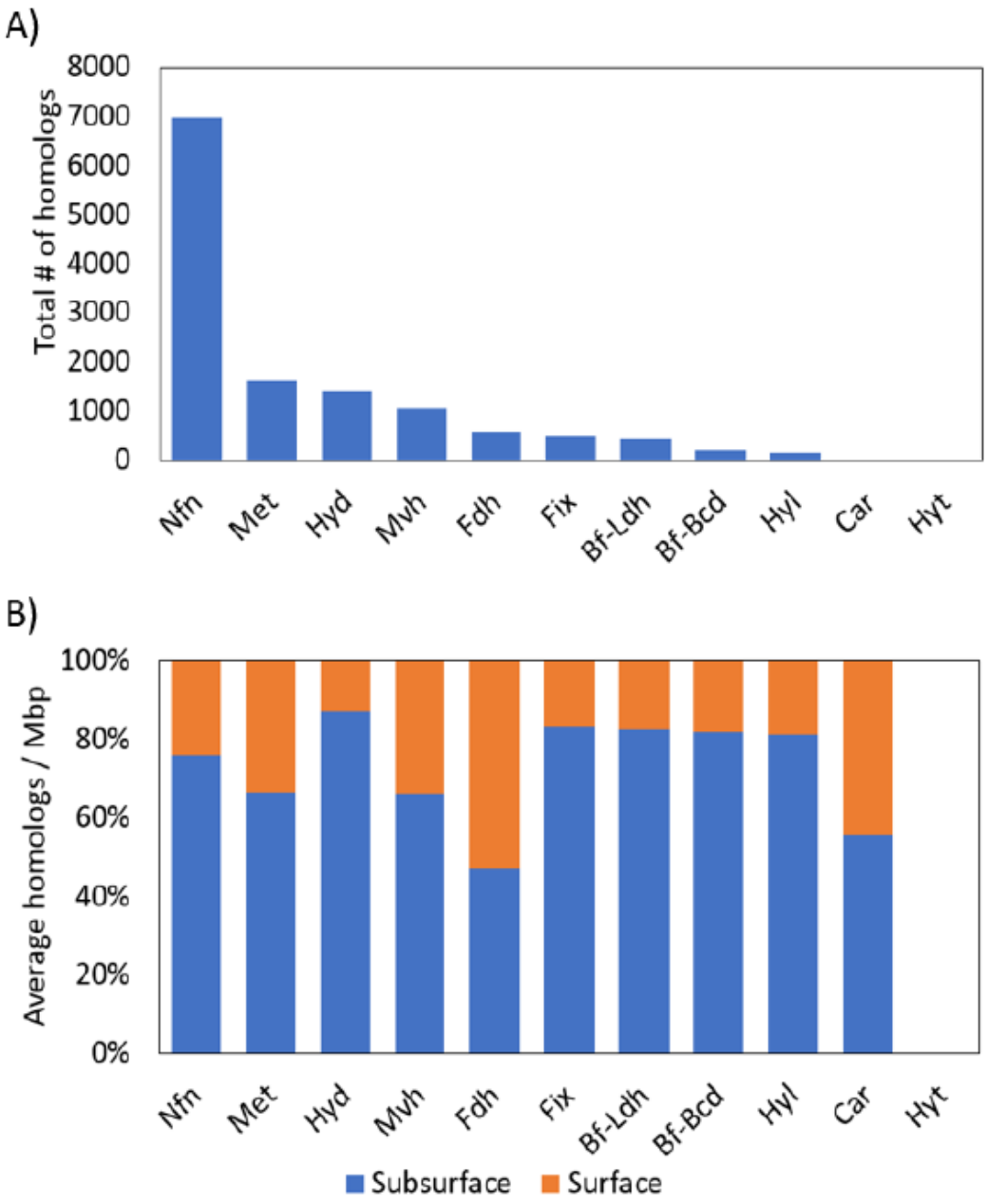
Supplementary Figure 19. Maximum-likelihood phylogenetic reconstruction of all Hdr homologs that associate with Fdh, Met, Mvh, and Hdr2 that functions in complete bacterial (red in outer strip) and archaeal (blue in outer strip) genomes. The tree was rooted with the paralog (terminals colored in green) thioredoxin reductase from *Escherichia coli* (P0A9P4) and *Shigella* sp. (WP_000537416). Names for each abbreviated protein complexes are provided in **Table 1**. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



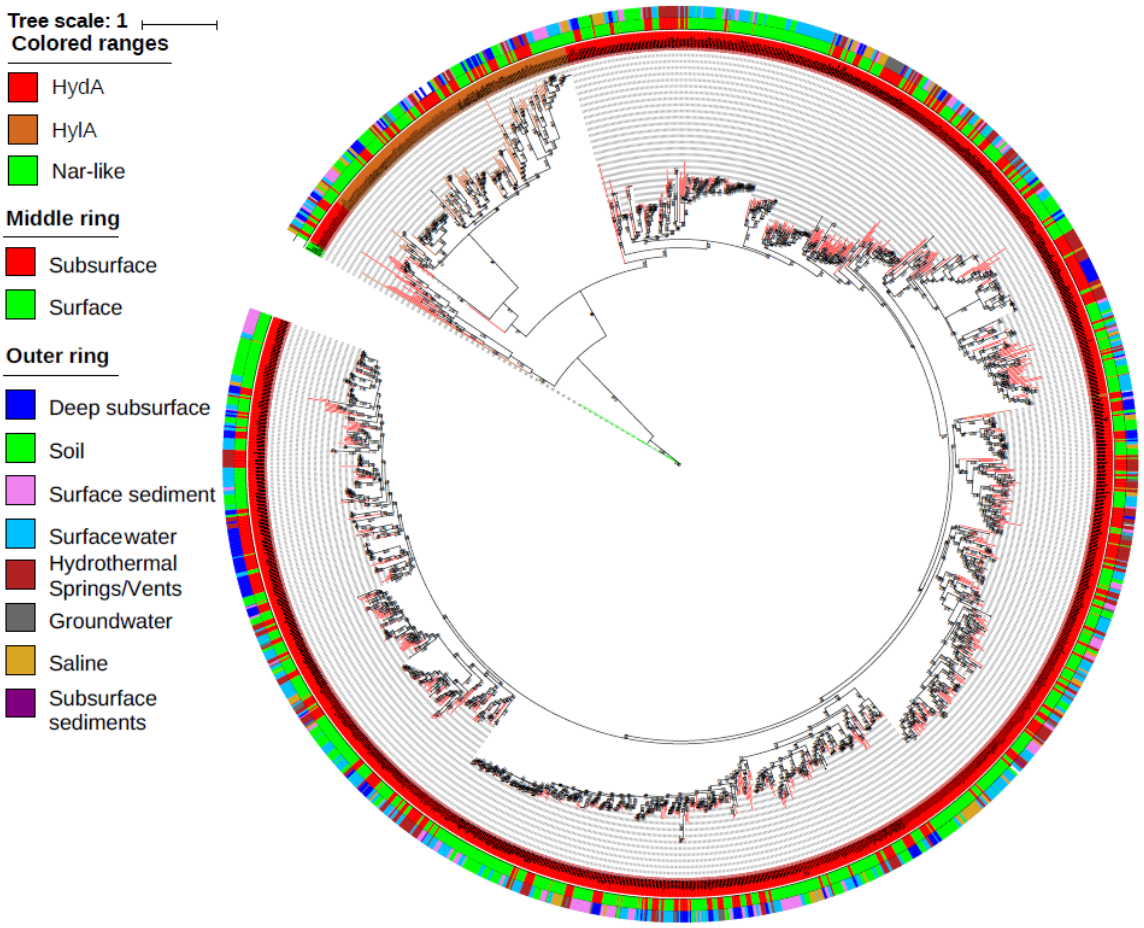
Supplementary Figure 20. Maximum-likelihood phylogenetic reconstruction of MetF homologs that associate with Hdr in complete bacterial (red in outer strip) and archaeal (blue in outer strip) genomes and are predicted to bifurcate (terminals colored in red). Those homologs that do not associate with Hdr and thus are not predicted to bifurcate are represented by uncolored terminals. The tree was rooted with the paralog (terminals colored in green) methylenetetrahydrofolate reductase from *Escherichia coli* (WP_089580839) and *Shigella flexneri* (OUZ65700). Names for each abbreviated protein complexes are provided in **Table 1**. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



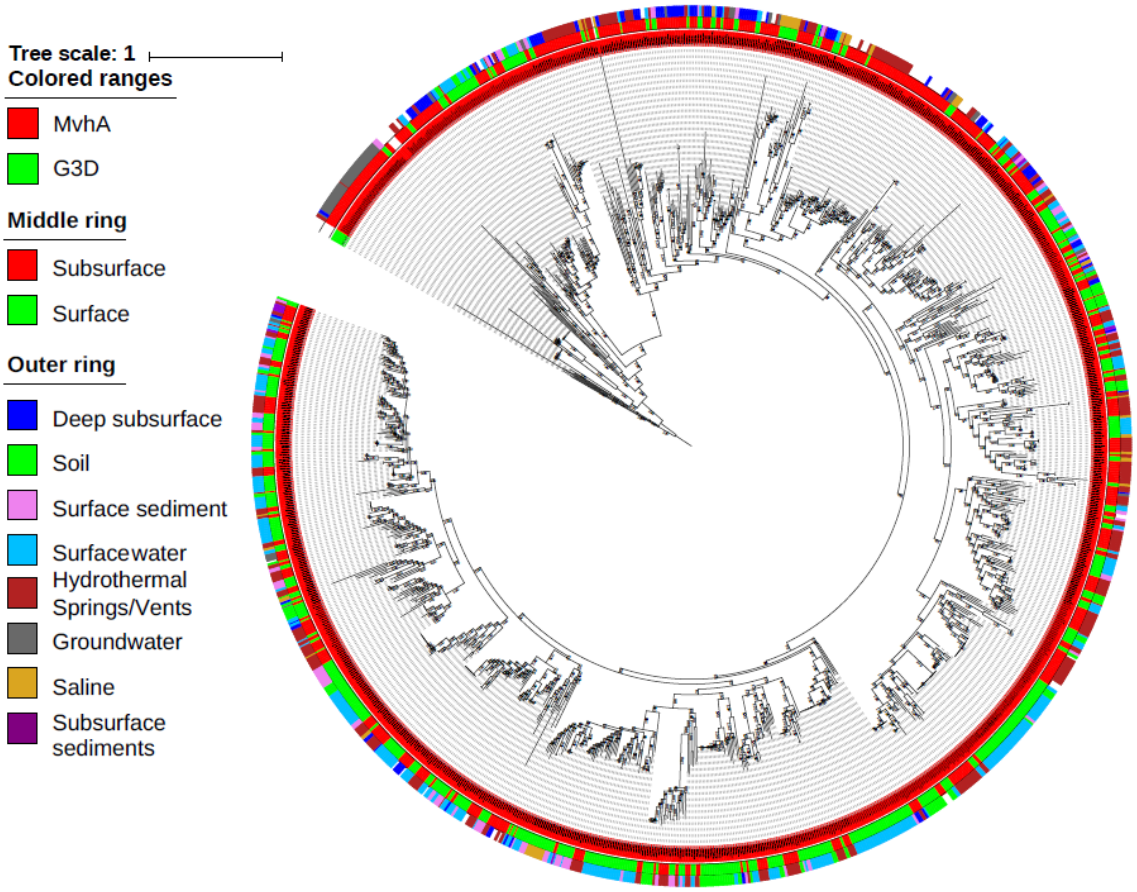
Supplementary Figure 21. Mantel regression of the pairwise distances of homologs of the subunits of Met plotted as a function of each other: **A**) MetF and MetV, **B**) MetF and HdrA, **C**) MetF and MvhD, **D**) HdrA and HdrC, **E**) HdrA and HdrB. Mantel *r* values are provided and the significance values (*p*-value) for all regressions shown were <0.001. Names for each abbreviated protein complex and their subunits are provided in **Table 1.**



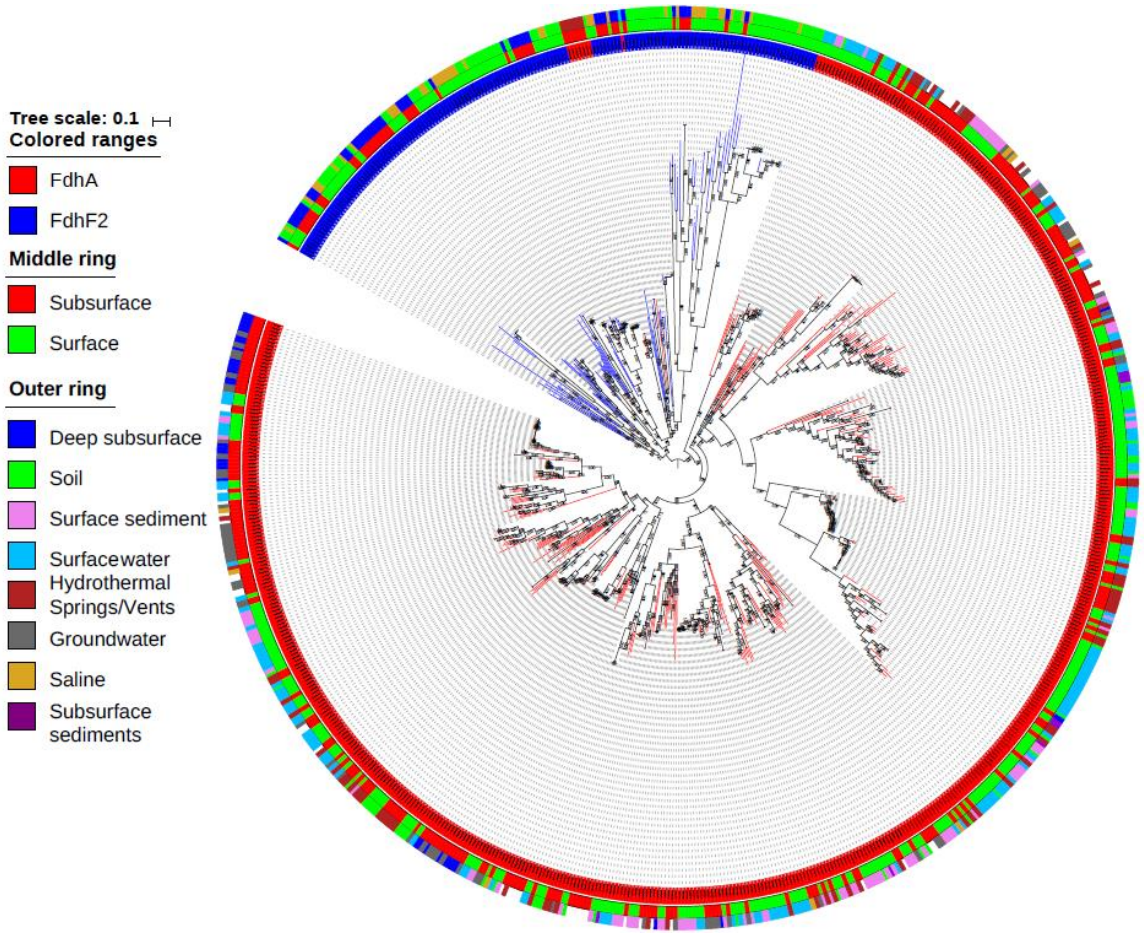
Supplementary Figure 22. Abundance and distribution of homologs of putative Bf enzymes among metagenomic sequences (n = 3,136 total metagenomes). **A**) Histogram depicting the total number of homologs for the specified type of Bf enzyme complex. **B**) The average abundances of homologs of all Bf enzyme complexes in non-redundant metagenomic contigs that were classified as 'surface' and 'subsurface' were normalized to the total number of megabase (Mb) pairs of sequence. Names for protein complexes are provided in **Table 1**.



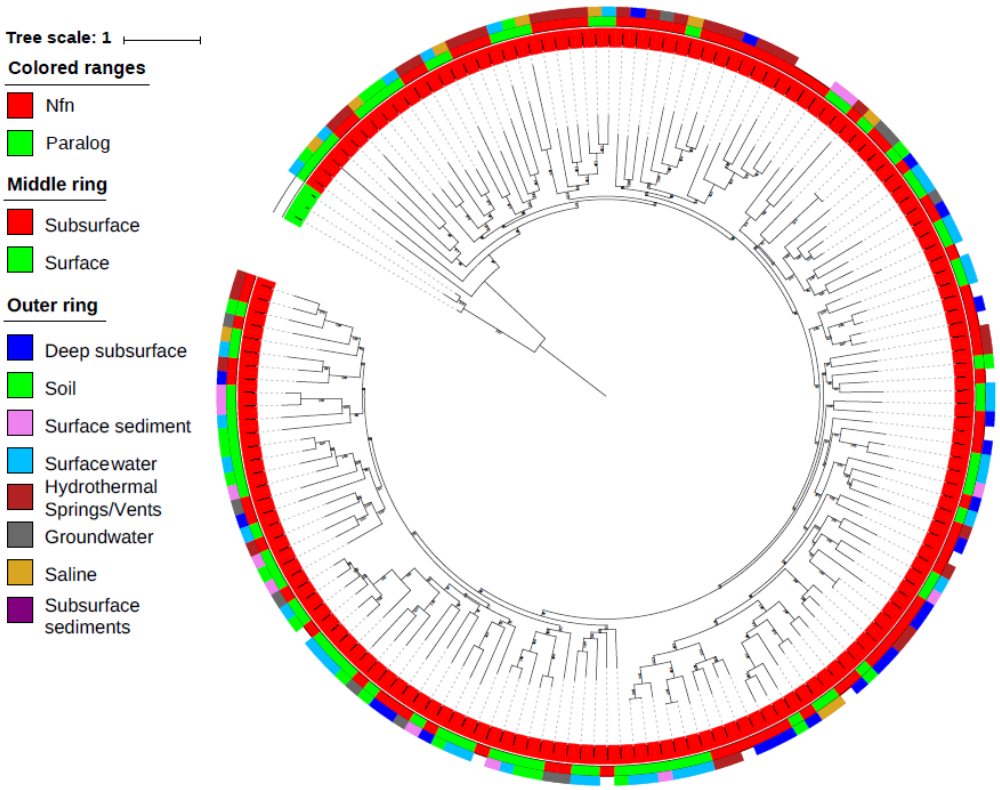
Supplementary Figure 23. Maximum-likelihood phylogenetic reconstruction of Bf homologs of Hyd (HydA and HylA; homologs of HytA were not identified among metagenomic sequences) in metagenomic sequences. The phylogeny was rooted with Nar-like proteins from *Homo sapiens* (NP_071938 and NP_036468), *Danio rerio* (A2RRV9), *Thalassiosira pseudonana* (XP_002289272) and *Ostreococcus lucimarinus* (XP_001416706). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to HydA (highlighted in red), HylA (highlighted in rust), or the paralog Nar-like proteins (highlighted in green). If a protein homolog was identified in a metagenome that lacked metadata allowing for coding of the outer or middle ring, it was left blank (white). Names of abbreviated protein complexes are provided in **Table 1.** Note, monomeric (non-Bf) HydA or HylA were not included in this analysis. Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



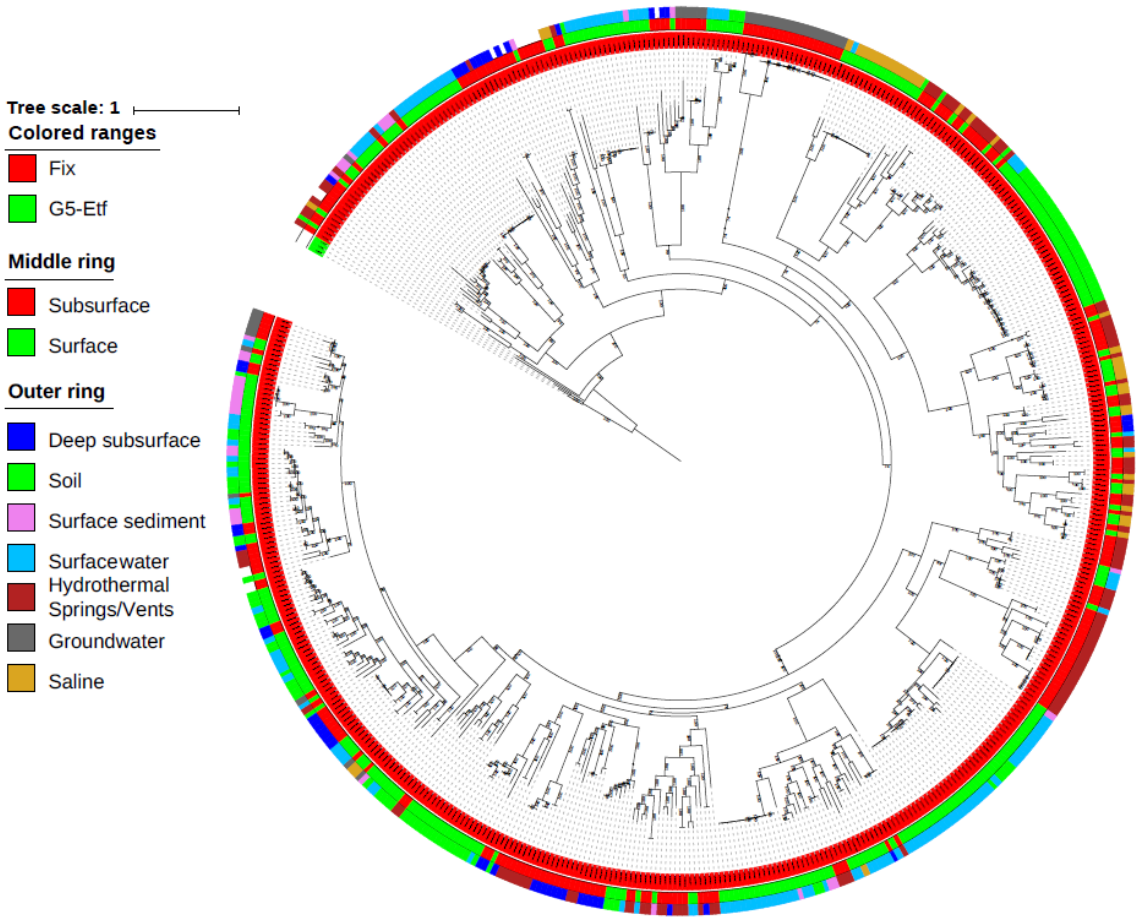
Supplementary Figure 24. Maximum-likelihood phylogenetic reconstruction of homologs of the large subunit of the Mvh complex (i.e., MvhA) identified in metagenomic sequences. The phylogeny was rooted with the large subunit of group 3d non-bifurcating [NiFe]-hydrogenase from *Cupriavidus necator* (WP_011154013), *Azoarcus olearius* (WP_011765148), *Paraburkholderia xenovorans* (WP_040123534), *Psychromonas ingrahamii* (WP_041766077), and *Rhodobacter capsulatus* (AAD38065). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to MvhA (highlighted in red) or the paralog (group 3d non-Bf [NiFe]-hydrogenase highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



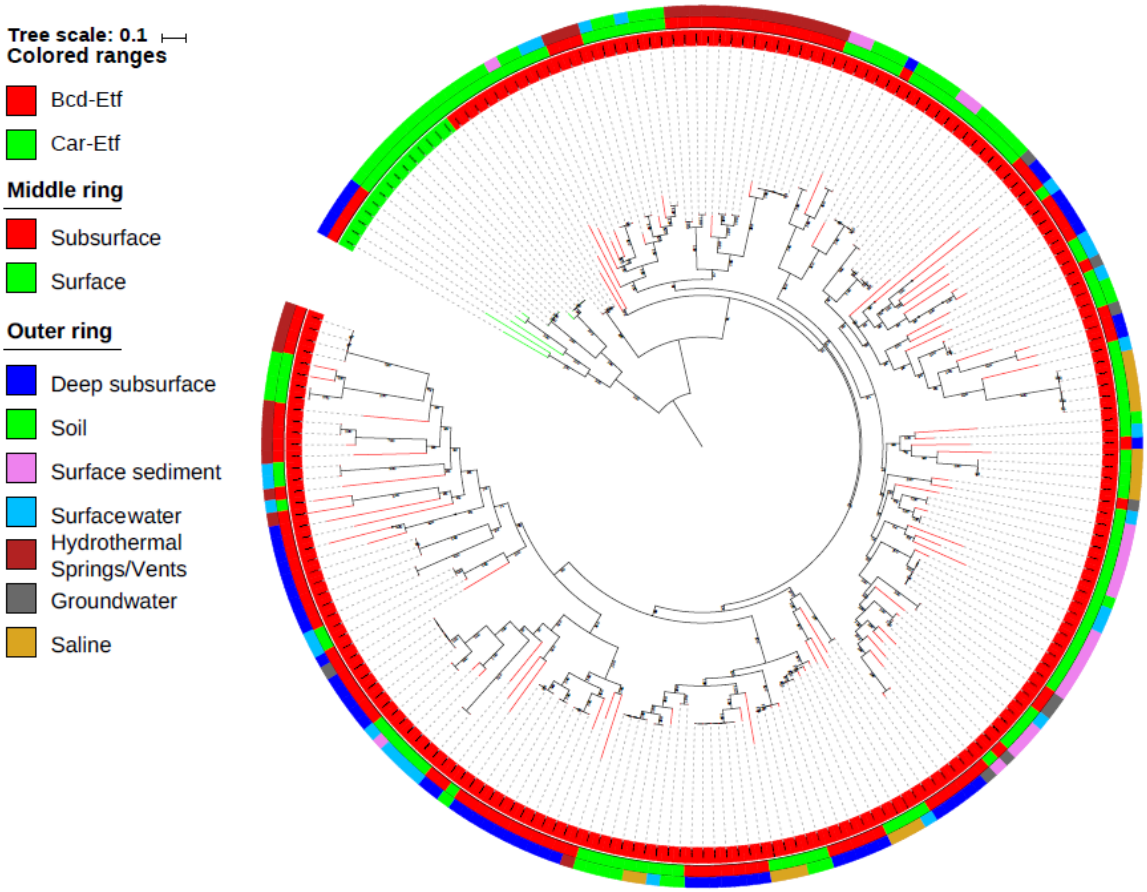
Supplementary Figure 25. Maximum-likelihood phylogenetic reconstruction of FdhA homologs that are predicted to form a Bf complex (i.e., Hyl and Hdr) in metagenomic sequences. FdhA homologs that are predicted to form a Bf complex with Hyt were not identified in metagenomic sequences and thus are not depicted. Mid-point rooting was used since FdhA and FdhF2 are paralogous. Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to FdhA (Hdr highlighted in red) or FdhF2 (Hyl highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



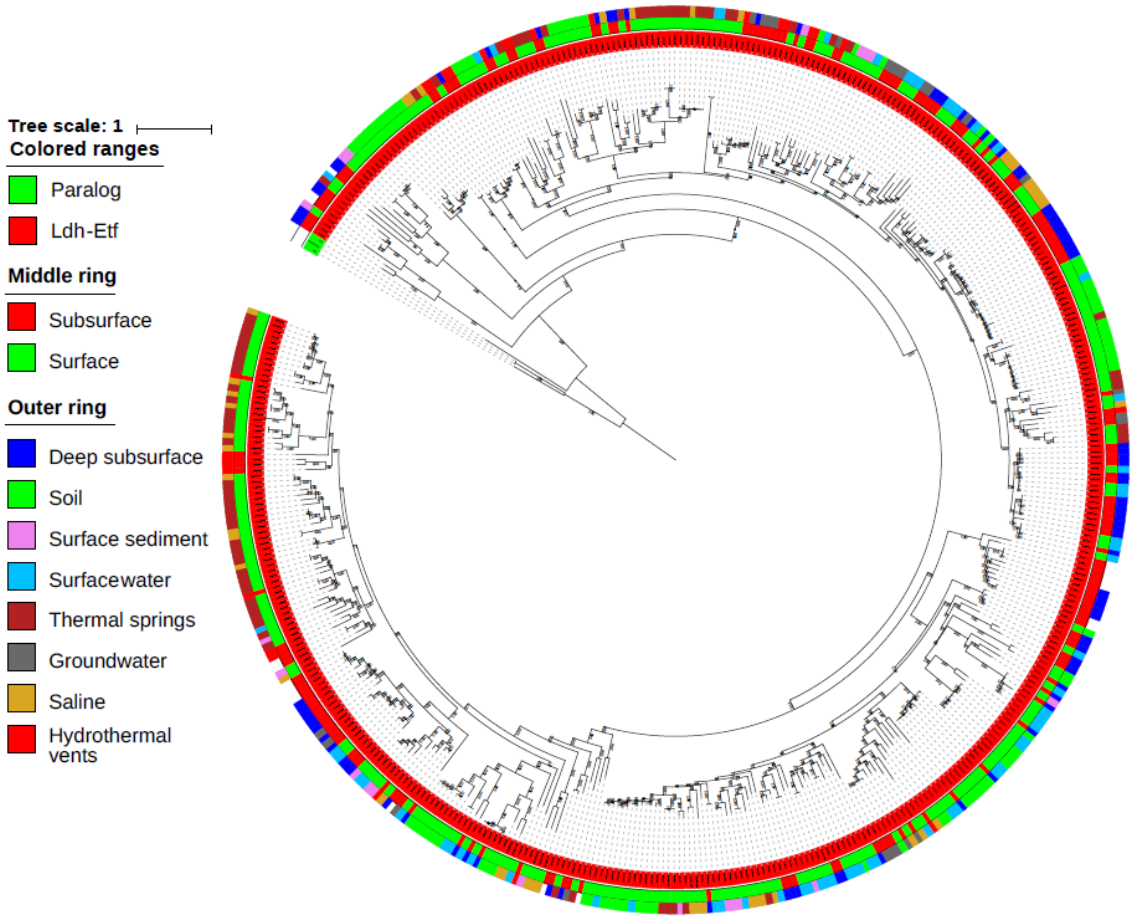
Supplementary Figure 26. Maximum-likelihood phylogenetic reconstruction of a concatenation of representative NfnSL homologs of each homolog 'bin' generated using a CD-HIT clustering approach applied to homologs identified in metagenomic sequences. All NfnSL homologs identified in metagenomes were first clustered into unique homolog 'bins' that contained closely related NfnSL homologs and the representative sequences of each 'bin' were extracted to reconstruct the phylogenetic tree of NfnSL. The phylogeny was rooted (green outer strip) with concatenated paralogous proteins which include dihydroorotate dehydrogenase from *Lactococcus lacticus* (WP_011835013) and glutamate synthase from *Azospirillum brasilense* (WP_035677957), dihydroorotate dehydrogenase from *Lactococcus garvieae* (BAK58851) and glutamate synthase from *Azospirillum oryzae* (WP_085087092), and dihydroorotate dehydrogenase from *Floricoccus tropicus* (WP_070791886) and glutamate synthase from *Azospirillum humicireducens* (WP_063635528). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to Nfn (highlighted in red) or the paralog (dihydroorotate dehydrogenase highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



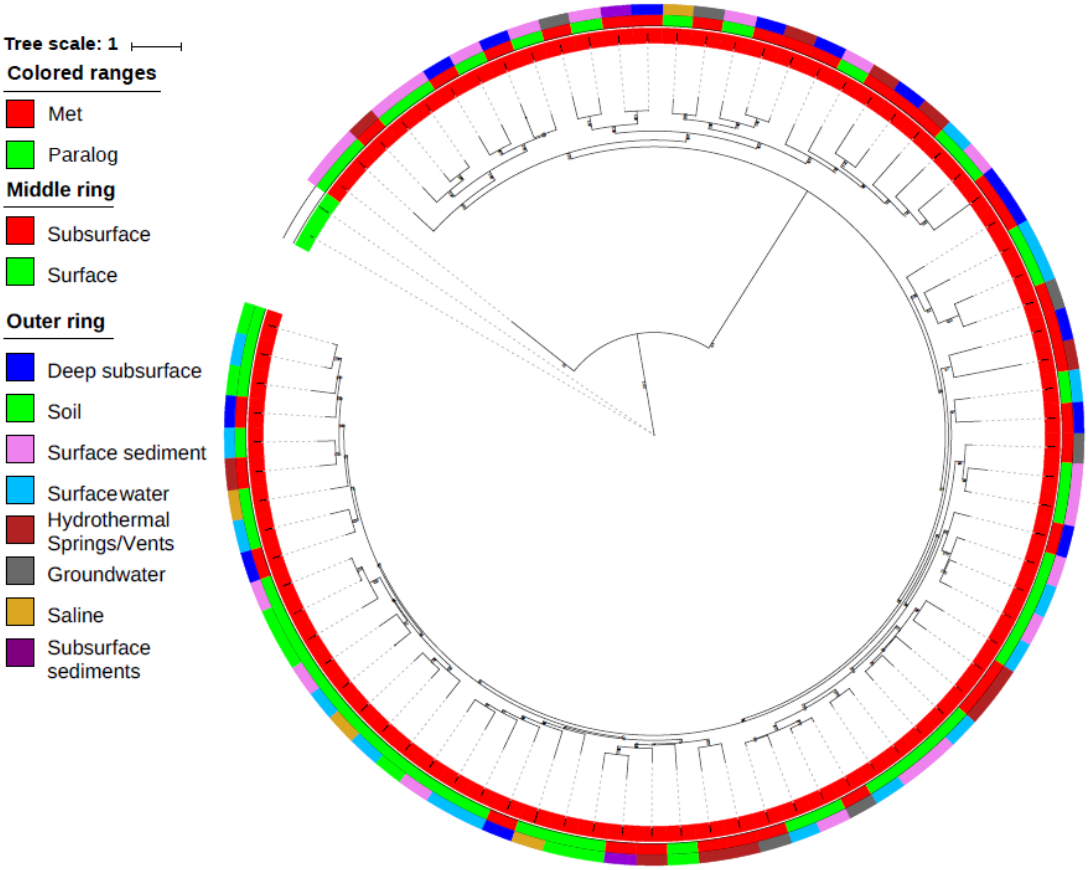
Supplementary Figure 27. Maximum-likelihood phylogenetic reconstruction of homologs of putative Bf EtfAB (i.e., Fix) identified in metagenomic sequences. The phylogeny was rooted with homologs of group 5 EtfAB from *Desulfitobacterium hafniense* (BAE84276-BAE84277), *Desulfotomaculum reducens* (ABO50304-ABO50305) and *Geobacillus kaustophilus* (BAD76971-BAD76972) that are not predicted to bifurcate (Garcia Costas et al., 2017). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to Fix (highlighted in red) or the paralog (non-Bf Etf highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



Supplementary Figure 28. Maximum-likelihood phylogenetic reconstruction of homologs of butyryl-CoA dehydrogenase (Bcd) and CarC identified in metagenomic sequences. Since, CarC and Bcd are paralogous enzymes, midpoint rooting was used. Sequence terminals are colored coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to Bcd (highlighted in red) or Car (highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



Supplementary Figure 29. Maximum-likelihood phylogenetic reconstruction of lactate dehydrogenase (Ldh) homologs that are predicted to form a complex with Bf Etf (i.e., Ldh) in metagenomic sequences. The phylogeny was rooted with the paralog alkyl dihyroxyacetone phosphate synthase from *Dictyostelium discoideum* (XP_637836) and *Trypanosoma brucei* (XP_845272). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to Ldh (highlighted in red) or the paralog (alkyl dihyroxyacetone phosphate synthase highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).



Supplementary Figure 30. Maximum-likelihood phylogenetic reconstruction of a concatenation of representative MetF homologs (that are predicted to form a complex with Bf Hdr) of each homolog 'bin' generated using a CD-HIT clustering approach applied to homologs identified in metagenomic sequences. All MetF homologs identified in metagenomes were first clustered into unique homolog 'bins' that contained closely related MetF homologs and the representative sequences of each unique 'bin' were extracted to reconstruct the phylogenetic tree of homologs of MetF. The tree was rooted with the paralog (terminals colored in green) methylenetetrahydrofolate reductase from *Escherichia coli* (WP_089580839) and *Shigella flexneri* (OUZ65700). Sequence terminals are color coded, with the outer ring indicating the environment type where a given homolog was identified, the middle ring indicating whether the environment type that the homolog was identified in was classified as a surface or subsurface environment, and the inner ring indicating homology to MetF-Hdr (highlighted in red) or the paralog (non-bifurcating methylenetetrahydrofolate reductase from highlighted in green). Protein homologs from metagenomes that lacked environmental classification were not colored in the middle and outer rings. Names of abbreviated protein complexes are provided in **Table 1.** Bootstrap values for each node are shown as a percentage (out of 1000 bootstrap replicates).

REFERENCES

Amenabar, M.J., Urschel, M.R., and Boyd, E.S. (2015). "Metabolic and taxonomic diversification in continental magmatic hydrothermal systems," in Microbial Evolution Under Extreme Conditions. Berlin: De Gruyter.), 57-96.

Amend, J.P., and Shock, E.L. (2001). Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and Bacteria. FEMS Microbiol. Rev. 25, 175-243.

Barkay, T., Kritee, K., Boyd, E., and Geesey, G. (2010). A thermophilic bacterial origin and subsequent constraints by redox, light and salinity on the evolution of the microbial mercuric reductase. Environ. Microbiol. 12, 2904-2917.

Baymann, F., Schoepp-Cothenet, B., Duval, S., Guiral, M., Brugna, M., Baffert, C., Russell, M.J., and Nitschke, W. 2018. On the natural history of flavin-based electron bifurcation. Frontiers in Microbiology.

Berry, L., Poudel, S., Tokmina-Lukaszewska, M., Colman, D.R., Nguyen, D.M., Schut, G.J., Adams, M.W., Peters, J.W., Boyd, E.S., and Bothner, B. (2017). H/D exchange mass spectrometry and statistical coupling analysis reveal a role for allostery in a ferredoxin-dependent bifurcating transhydrogenase catalytic cycle. Biochim. Biophys. Acta. 1862, 9-17.

Bertsch, J., Oppinger, C., Hess, V., Langer, J.D., and Muller, V. (2015). Heterotrimeric NADH-oxidizing methylenetetrahydrofolate reductase from the acetogenic bacterium *Acetobacterium woodii*. J. Bacteriol. 197, 1681-1689.

Bertsch, J., Parthasarathy, A., Buckel, W., and Muller, V. (2013). An electron bifurcating caffeyl-CoA reductase. J. Biol. Chem. 288, 11304-11311.

Biegel, E., Schmidt, S., Gonzalez, J.M., and Muller, V. (2011). Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell Mol. Life Sci. 68, 613-634.

Bogaardt, C. (2014). FigTree [Online]. Available: http://tree.bio.ed.ac.uk/software/figtree/ [Accessed].

Boughanemi, S., Lyonnet, J., Infossi, P., Bauzan, M., Kosta, A., Lignon, S., Giudici-Orticoni, M.T., and Guiral, M. (2016). Microbial oxidative sulfur metabolism: biochemical evidence of the membrane-bound heterodisulfide reductase-like complex of the bacterium Aquifex aeolicus. FEMS Microbiol. Lett. 363.

Boyd, E.S., Anbar, A.D., Miller, S., Hamilton, T.L., Lavin, M., and Peters, J.W. (2011a). A late methanogen origin for molybdenum-dependent nitrogenase. Geobiology 9, 221-232.

Boyd, E.S., Costas, A.M.G., Hamilton, T.L., Mus, F., and Peters, J.W. (2015). Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic Metabolism. J. Bacteriol. 197, 1690-1699.

Boyd, E.S., Hamilton, T.L., and Peters, J.W. (2011b). An alternative path for the evolution of biological nitrogen fixation. Front. Microbiol. 2. 205.

Boyd, E.S., and Peters, J.W. (2013). New insights into the evolutionary history of biological nitrogen fixation. Front Microbiol 4. 201.

Boyd, E.S., Schut, G.J., Adams, M.W., and Peters, J.W. (2014). Hydrogen metabolism and the evolution of biological respiration. Microbe 9, 361-367.

Buan, N.R., and Metcalf, W.W. (2010). Methanogenesis by *Methanosarcina acetivorans* involves two structurally and functionally distinct classes of heterodisulfide reductase. Mol. Microbiol. 75, 843-853.

Buckel, W., and Thauer, R.K. (2013). Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. Biochim. Biophys. Acta, Bioenerg. 1827, 94-113.

Buckel, W., and Thauer, R.K. (2018a). Flavin-based electron bifurcation, a new mechanism of biological energy coupling. Chem. Rev.

Buckel, W., and Thauer, R.K. (2018b). Flavin-based electron bifurcation, ferredoxin, flavodoxin, and anaerobic respiration with protons (Ech) or NAD+ (Rnf) as electron acceptors: A historical review. Front. Microbiol. 9, 401.

Calusinska, M., Happe, T., Joris, B., and Wilmotte, A. (2010). The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. Microbiol. 156, 1575-1588.

Chowdhury, N.P., Mowafy, A.M., Demmer, J.K., Upadhyay, V., Koelzer, S., Jayamani, E., Kahnt, J., Hornung, M., Demmer, U., Ermler, U., and Buckel, W. (2014). Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J. Biol. Chem .289, 5145-5157.

Colman, D.R., Poudel, S., Stamps, B.W., Boyd, E.S., and Spear, J.R. (2017). The deep, hot biosphere: Twenty-five years of retrospection. Proc. Natl. Acad. Sci. USA 114, 6895-6903.

Costa, K.C., Lie, T.J., Xia, Q., and Leigh, J.A. (2013). VhuD facilitates electron flow from H_2 or formate to heterodisulfide reductase in *Methanococcus maripaludis*. J. Bacteriol. 195, 5160-5165.

Costa, K.C., Wong, P.M., Wang, T.S., Lie, T.J., Dodsworth, J.A., Swanson, I., Burn, J.A., Hackett, M., and Leigh, J.A. (2010). Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. Proc. Natl. Acad. Sci. USA 107, 11050-11055.

D'hondt, S., Inagaki, F., Zarikian, C.A., Abrams, L.J., Dubois, N., Engelhardt, T., Evans, H., Ferdelman, T., Gribsholt, B., Harris, R.N., Hoppie, B.W., Hyun, J.H., Kallmeyer, J., Kim, J., Lynch, J.E., Mckinley, C.C., Mitsunobu, S., Morono, Y., Murray, R.W., Pockalny, R., Sauvage, J., Shimono, T., Shiraishi, F., Smith, D.C., Smith-Duque, C.E., Spivack, A.J., Steinsbu, B.O., Suzuki, Y., Szpak, M., Toffin, L., Uramoto, G., Yamaguchi, Y.T., Zhang, G.L., Zhang, X.H., and Ziebis, W. (2015). Presence of oxygen and aerobic communities from sea floor to basement in deep-sea sediments. Nat. Geosci. 8, 299-304.

Demmer, J.K., Chowdhury, N.P., Selmer, T., Ermler, U., and Buckel, W. (2017). The semiquinone swing in the bifurcating electron transferring flavoprotein/butyryl-CoA dehydrogenase complex from *Clostridium difficile*. Nat. Commun. 8. 1577.

Demmer, J.K., Huang, H., Wang, S.N., Demmer, U., Thauer, R.K., and Ermler, U. (2015). Insights into flavin-based electron bifurcation via the NADH-dependent reduced ferredoxin:NADP oxidoreductase structure. J. Biol. Chem. 290, 21985-21995.

Djordjevic, S., Pace, C.P., Stankovich, M.T., and Kim, J.-J.P. (1995). Three-dimensional structure of butyryl-CoA dehydrogenase from *Megasphaera elsdenii*. Biochem. 34, 2163-2171.

Earl, C., Ronson, C., and Ausubel, F. (1987). Genetic and structural analysis of the *Rhizobium meliloti fixA*, *fixB*, *fixC*, and *fixX* genes. J. Bacteriol. 169, 1127-1136.

Eddy, S. (2015). HMMER: biosequence analysis using profile hidden Markov models [Online]. Available: hmmer.org [Accessed].

Edgren, T., and Nordlund, S. (2004). The *fixABCX* genes in *Rhodospirillum rubrum* encode a putative membrane complex participating in electron transfer to nitrogenase. J. Bacteriol. 186, 2052-2060.

Edwards, K.J., Becker, K., and Colwell, F. (2012). The deep, dark energy biosphere: Intraterrestrial life on Earth. Annu. Rev. Earth. Plan. Sci. 40, 551-568.

Evans, P.N., Parks, D.H., Chadwick, G.L., Robbins, S.J., Orphan, V.J., Golding, S.D., and Tyson, G.W. (2015). Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. Science 350, 434-438.

Ferry, J.G. (1990). Formate dehydrogenase. FEMS Microbiol. Lett. 87, 377-382.

Forterre, P. (2015). The universal tree of life: an update. Front Microbiol 6. 717.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152.

Garcia Costas, A.M., Poudel, S., Miller, A.-F., Schut, G.J., Ledbetter, R.N., Fixen, K.R., Seefeldt, L.C., Adams, M.W., Harwood, C.S., and Boyd, E.S. (2017). Defining electron bifurcation in the electron transferring flavoprotein family. J. Bacteriol., e00440-17.

Garvie, E.I. (1980). Bacterial lactate dehydrogenases. Microbiol. Rev. 44, 106-139.

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., and Morales, S.E. (2016). Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J. 10, 761.

Griffith, D.M., Veech, J.A., and Marsh, C.J. (2016). cooccur: Probabilistic species co-occurrence analysis in R. J. Stat. Software. 69, 1-17.

Guenther, B.D., Sheppard, C.A., Tran, P., Rozen, R., Matthews, R.G., and Ludwig, M.L. (1999). The structure and properties of methylenetetrahydrofolate reductase from *Escherichia coli* suggest how folate ameliorates human hyperhomocysteinemia. Nat. Struct. Biol. 6, 359-365.

Hedderich, R., Koch, J., Linder, D., and Thauer, R.K. (1994). The heterodisulfide reductase from *Methanobacterium hermoautotrophicum* contains sequence motifs characteristic of pyridine-nucleotide-dependent thioredoxin reductases. FEBS J. 225, 253-261.

Herrmann, G., Jayamani, E., Mai, G., and Buckel, W. (2008). Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. J. Bacteriol. 190, 784-791.

Hess, V., Schuchmann, K., and Müller, V. (2013). The ferredoxin:NAD+ oxidoreductase (Rnf) from the acetogen *Acetobacterium woodii* requires Na+ and is reversibly coupled to the membrane potential. J. Biol. Chem. 288, 31496-31502.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman,

D.A., Finstad, K.M., Amundson, R., Thomas, B.C., and Banfield, J.F. (2016). A new view of the tree of life. Nat. Microbiol 1. 16048.

Jannasch, H.W., and Mottl, M.J. (1985). Geomicrobiology of deep-sea hydrothermal vents. Science 229, 717-725.

Jungermann, K., Thauer, R.K., Leimenstoll, G., and Decker, K. (1973). Function of reduced pyridine nucleotide-ferredoxin oxidoreductases in saccharolytic Clostridia. Biochim. Biophys. Acta 305, 268-280.

Kaster, A.K., Moll, J., Parey, K., and Thauer, R.K. (2011). Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. Proc. Natl. Acad. Sci. USA 108, 2981-2986.

Kieft, T.L. (2016). Microbiology of the deep continental biosphere. Ad.v Env. Microbiol. 1, 225-249.

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874.

Laso-Perez, R., Wegener, G., Knittel, K., Widdel, F., Harding, K.J., Krukenberg, V., Meier, D.V., Richter, M., Tegetmeyer, H.E., Riedel, D., Richnow, H.H., Adrian, L., Reemtsma, T., Lechtenfeld, O.J., and Musat, F. (2016). Thermophilic archaea activate butane via alkyl-coenzyme M formation. Nature 539, 396-401.

Ledbetter, R., Garcia-Costas, A.M., Lubner, C.E., Mulder, D.W., Tokmina-Lukaszewska, M., Artz, J.H., Patterson, A., Magnuson, T., Jay, Z.J., and Duan, H.D. (2017). The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: Generation of low-potential reducing equivalents for nitrogenase catalysis. Biochem. 56, 4177-4190.

Letunic, I., and Bork, P. (2016). Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242-W245.

Li, F., Hinderberger, J., Seedorf, H., Zhang, J., Buckel, W., and Thauer, R.K. (2008). Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-coA dehydrogenase/Etf complex from *Clostridium kluyveri*. J. Bacteriol 190, 843-850.

Li, W., and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659.

Lindsay, M.R., Anderson, C., Fox, N., Scofield, G., Allen, J., Anderson, E., Bueter, L., Poudel, S., Sutherland, K., Munson-Mcgee, J.H., Van Nostrand, J.D., Zhou, J., Spear,

J.R., Baxter, B.K., Lageson, D.R., and Boyd, E.S. (2017). Microbialite response to an anthropogenic salinity gradient in Great Salt Lake, Utah. Geobiology 15, 131-145.

Lubner, C.E., Jennings, D.P., Mulder, D.W., Schut, G.J., Zadvornyy, O.A., Hoben, J.P., Tokmina-Lukaszewska, M., Berry, L., Nguyen, D.M., Lipscomb, G.L., Bothner, B., Jones, A.K., Miller, A.F., King, P.W., Adams, M.W.W., and Peters, J.W. (2017). Mechanistic insights into energy conservation by flavin-based electron bifurcation. Nat. Chem. Biol. 13, 655-659.

Lyons, T.W., Reinhard, C.T., and Planavsky, N.J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. Nature 506, 307-315.

Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., and Williams, P. (2011). IMG: The integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40, D115-D122.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. (2008). Hydrothermal vents and the origin of life. Nat. Rev. Microbiol. 6, 805-814.

Martin, W.F. (2012). Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. FEBS Lett. 586, 485-493.

Martin, W.F., Bryant, D.A., and Beatty, J.T. (2017). A physiological perspective on the origin and evolution of photosynthesis. FEMS Microbiol. Rev.

Meyer, J. (2007). [FeFe] hydrogenases and their evolution: a genomic perspective. Cell Mol. Life Sci. 64, 1063-1084.

Mock, J., Wang, S.N., Huang, H.Y., Kahnt, J., and Thauer, R.K. (2014). Evidence for a gexaheteromeric methylenetetrahydrofolate reductase in *Moorella thermoacetica*. J. Bacteriol. 196, 3303-3314.

Mulder, D.W., Boyd, E.S., Sarma, R., Lange, R.K., Endrizzi, J.A., Broderick, J.B., and Peters, J.W. (2010). Stepwise [FeFe]-hydrogenase H-cluster assembly revealed in the structure of HydAΔEFG. Nature 465, 248-U143.

Müller, V., Chowdhury, N.P., and Basen, M. 2018. Electron bifurcation: A long-hidden energy-coupling mechanism. Annual Review of Microbiology 72:null.

Nguyen, D.M.N., Schut, G.J., Zadvornyy, O.A., Tokmina-Lukaszewska, M., Poudel, S., Lipscomb, G.L., Adams, L.A., Dinsmore, J.T., Nixon, W.J., Boyd, E.S., Bothner, B., Peters, J.W., and Adams, M.W.W. (2017). Two functionally distinct NADP(+)-dependent ferredoxin oxidoreductases maintain the primary redox balance of Pyrococcus furiosus. J. Biol. Chem. 292, 14603-14616.

Nitschke, W., and Russell, M.J. (2012). Redox bifurcations: Mechanisms and importance to life now, and at its origin. Bioessays 34, 106-109.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289-290.

Paul, K., Nonoh, J.O., Mikulski, L., and Brune, A. (2012). "Methanoplasmatales," Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens. Appl. Environ. Microbiol. 78, 8245-8253.

Peters, J.W., Lanzilotta, W.N., Lemon, B.J., and Seefeldt, L.C. (1998). X-ray crystal structure of the Fe-only hydrogenase (Cpl) from *Clostridium pasteurianum* to 1.8 angstrom resolution. Science 282, 1853-1858.

Peters, J.W., Miller, A.F., Jones, A.K., King, P.W., and Adams, M.W.W. (2016). Electron bifurcation. Curr. Opin. Chem. Biol. 31, 146-152.

Peters, J.W., Schut, G.J., Boyd, E.S., Mulder, D.W., Shepard, E.M., Broderick, J.B., King, P.W., and Adams, M.W.W. (2015). [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochim. Biophys. Acta, Mol Cell Res. 1853, 1350-1369.

Poudel, S., Colman, D.R., Fixen, K.R., Ledbetter, R., Zheng, Y., Pence, N., Seefeldt, L.C., Peters, J.W., Harwood, C.S., and Boyd, E.S. (2018). Electron transfer to nitrogenase in different genomic and metabolic backgrounds. J. Bacteriol. 200, e00757-17.

Poudel, S., Tokmina-Lukaszewska, M., Colman, D.R., Refai, M., Schut, G.J., King, P.W., Maness, P.C., Adams, M.W.W., Peters, J.W., Bothner, B., and Boyd, E.S. (2016). Unification of [FeFe]-hydrogenases into three structural and functional groups. Biochim. Biophys. Acta, Gen. Subj. 1860, 1910-1921.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F.O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-D596.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., and Woyke, T. (2013). Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431-437.

Sait, M., Hugenholtz, P., and Janssen, P.H. (2002). Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. Environ. Microbiol. 4, 654-666.

Sarkar, A., Köhler, J., Hurek, T., and Reinhold-Hurek, B. (2012). A novel regulatory role of the Rnf complex of Azoarcus sp. strain BH72. Mol. Microbiol. 83, 408-422.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., and Robinson, C.J. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541.

Schuchmann, K., and Müller, V. (2012). A bacterial electron-bifurcating hydrogenase. J. Biol. Chem 287, 31165-31171.

Schut, G.J., and Adams, M.W.W. (2009). The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. J. Bacteriol. 191, 4451-4457.

Schut, G.J., Boyd, E.S., Peters, J.W., and Adams, M.W.W. (2013). The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. FEMS Microbiol. Rev. 37, 182-203.

Schut, G.J., Zadvornyy, O., Wu, C.H., Peters, J.W., Boyd, E.S., and Adams, M.W.W. (2016). The role of geochemistry and energetics in the evolution of modern respiratory complexes from a proton-reducing ancestor. Biochim. Biophys. Acta, Bioenerg. 1857, 958-970.

Shen, L.D., Wu, H.S., Gao, Z.Q., Cheng, H.X., Li, J., Liu, X., and Ren, Q.Q. (2016). Distribution and activity of anaerobic ammonium-oxidising bacteria in natural freshwater wetland soils. Appl. Microbiol. Biotech. 100, 3291-3300.

Shock, E.L., Holland, M., Meyer-Dombard, D., Amend, J.P., Osburn, G.R., and Fischer, T.P. (2010). Quantifying inorganic sources of geochemical energy in hydrothermal ecosystems, Yellowstone National Park, USA. Geochim. Cosmochim. Acta 74, 4005-4043.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., Mcwilliam, H., Remmert, M., Söding, J., Thompson, J.D., and Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7, 539-544.

Sorokin, D.Y., Makarova, K.S., Abbas, B., Ferrer, M., Golyshin, P.N., Galinski, E.A., Ciordia, S., Mena, M.C., Merkel, A.Y., Wolf, Y.I., Van Loosdrecht, M.C.M., and Koonin, E.V. (2017). Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. Nat. Microbiol. 2.

Sousa, F.L., Preiner, M., and Martin, W.F. (2018). Native metals, electron bifurcation, and CO2 reduction in early biochemical evolution. Curr. Opion. Microbiol. 43, 77-83.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30, 1312-1313.

Stein, L.Y., La Duc, M.T., Grundl, T.J., and Nealson, K.H. (2001). Bacterial and archaeal populations associated with freshwater ferromanganous micronodules and sediments. Environ. Microbiol. 3, 10-18.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596-1599.

Team R, R.C. (2014). "R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014".).

Thauer, R.K., Kaster, A.-K., Goenrich, M., Schick, M., Hiromoto, T., and Shima, S. (2010). Hydrogenases from methanogenic archaea, nickel, a novel cofactor, and H2 storage. Ann. Rev. Biochem. 79, 507-536.

Thauer, R.K., Kaster, A.K., Seedorf, H., Buckel, W., and Hedderich, R. (2008). Methanogenic archaea: ecologically relevant differences in energy conservation. Nat. Rev. Microbiol. 6, 579-591.

Vanwonterghem, I., Evans, P.N., Parks, D.H., Jensen, P.D., Woodcroft, B.J., Hugenholtz, P., and Tyson, G.W. (2016). Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. Nat. Microbiol. 1. 16170.

Vignais, P.M., Billoud, B., and Meyer, J. (2001). Classification and phylogeny of hydrogenases. FEMS Microbiol. Rev. 25, 455-501.

Wang, S., Huang, H., Kahnt, J., Mueller, A.P., Köpke, M., and Thauer, R.K. (2013a). NADP-specific electron-bifurcating [FeFe]-hydrogenase in a functional complex with formate dehydrogenase in *Clostridium autoethanogenum* grown on CO. J. Bacteriol. 195, 4373-4386.

Wang, S., Huang, H., Kahnt, J., and Thauer, R.K. (2013b). Clostridium acidurici electron-bifurcating formate dehydrogenase. Appl. Environ. Microbiol. 79, 6176-6179.

Wang, S.N., Huang, H.Y., Moll, J., and Thauer, R.K. (2010). NADP+ reduction with reduced ferredoxin and NADP+ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. J. Bacteriol. 192, 5115-5123.

Weghoff, M.C., Bertsch, J., and Muller, V. (2015). A novel mode of lactate metabolism in strictly anaerobic bacteria. Environ. Microbiol. 17, 670-677.

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. 2016. The physiology and habitat of the last universal common ancestor. Nature Microbiology 1:16116.

Yan, Z., Wang, M.Y., and Ferry, J.G. (2017). A ferredoxin- and F420H2-dependent, electron bifurcating, heterodisulfide reductase with homologs in the domains Bacteria and Archaea. Mbio 8. e02285-16.

CHAPTER THREE

DEFINING ELECTRON BIFURCATION IN THE ELECTRON TRANSFERRING FLAVOPROTEIN FAMILY

Contribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Amaya M. Garcia Costas

Contributions: Co-first author. Designed and conducted the experiment. Contributed to the writing of the manuscript. Contributed equally to the manuscript.

Author: Saroj Poudel

Contributions: Co-first author. Designed and conducted the experiment. Contributed to the writing of the manuscript. Contributed equally to the manuscript.

Co-Author: Anne-Francis Miller

Contributions: Assisted with structural analysis. Contributed to the writing of the manuscript.

Co-Author: Gerrit J. Schut

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Rhesa N. Ledbetter

Contributions: Assisted with structural analysis.

Co-Author: Kathryn R. Fixen

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Lance C. Seefeldt

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Michael W.W. Adams

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Caroline S. Harwood

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Eric S. Boyd

Contributions: Provided feedback on the study design. Provided statistical advice and contributed to the writing of the manuscript.

Co-Author: John W. Peters

Contributions: Provided feedback on the study design. Provided statistical advice and contributed to the writing of the manuscript.

Manuscript Information Page

Amaya M. Garcia Costas, Saroj Poudel, Anne-Frances Miller, Gerrit J. Schut, Rhesa N. Ledbetter, Kathryn R. Fixen, Lance C. Seefeldt, Michael W. W. Adams, Caroline S. Harwood, Eric S. Boyd, and John W. Peters
Journal of Bacteriology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-review journal
____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published online August 2017, doi: 10.1128/JB.00440-17.

Defining Electron Bifurcation in the Electron Transferring Flavoprotein Family

Amaya M. Garcia Costas[1&], Saroj Poudel[2&], Anne-Frances Miller[3], Gerrit J. Schut[4], Rhesa N. Ledbetter[5], Kathryn R. Fixen[6], Lance C. Seefeldt[5], Michael W. W. Adams[4], Caroline S. Harwood[6], Eric S. Boyd[2#], and John W. Peters[1, 7#]

[1]Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717
[2]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717
[3]Department of Chemistry, University of Kentucky, Lexington, KY 40506
[4]Department of Biochemistry and Molecular Biology, Athens, GA 30602
[5]Department of Chemistry and Biochemistry, Utah State University, Logan, UT 84322
[6]Department of Microbiology, University of Washington, Seattle, WA 98195
[7]Institute of Biological Chemistry, Washington State University, Pullman, WA 99164

Running head: Energy Conserving Electron Transferring Flavoproteins

[#]Address correspondence to John W. Peters (jw.peters@wsu.edu) and Eric S. Boyd (eboyd@montana.edu)

[&]These authors contributed equally to this work.

ABSTRACT

Electron bifurcation is the coupling of exergonic and endergonic redox reactions to simultaneously generate (or utilize) low- and high-potential electrons. It is the third recognized form of energy conservation in biology and was recently described for select electron-transferring flavoproteins (Etfs). Etfs are flavin-containing heterodimers best known for donating electrons derived from fatty acid and amino acid oxidation to an electron transfer respiratory chain via Etf-quinone oxidoreductase. Canonical examples contain a flavin adenine dinucleotide (FAD) that is involved in electron transfer, as well as a non-redox-active AMP. However, Etfs demonstrated to bifurcate electrons, contain a second FAD in place of the AMP. To expand our understanding of the functional variety and metabolic significance of Etfs and to identify amino acid sequence motifs that potentially enable electron bifurcation, we compiled 1,314 Etf protein sequences from genome sequence databases and subjected them to informatic and structural analyses. Etfs were identified in diverse archaea and bacteria, and they clustered into five distinct well-supported groups, based on their amino acid sequences. Gene neighborhood analyses indicated that these Etf group designations largely correspond to putative differences in functionality. Etfs with the demonstrated ability to bifurcate were found to form one group, suggesting that distinct conserved amino acid sequence motifs enable this capability. Indeed, structural modeling and sequence alignments revealed that identifying residues occur in the NADH- and FAD-binding regions of bifurcating Etfs. Collectively, a new classification scheme for Etf proteins that delineates putative bifurcating versus nonbifurcating members is presented

and suggests that Etf-mediated bifurcation is associated with surprisingly diverse enzymes.

IMPORTANCE

Electron bifurcation has recently been recognized as an electron transfer mechanism used by microorganisms to maximize energy conservation. Bifurcating enzymes couple thermodynamically unfavorable reactions with thermodynamically favorable reactions in an overall spontaneous process. Here we show that the electron-transferring flavoprotein (Etf) enzyme family exhibits far greater diversity than previously recognized, and we provide a phylogenetic analysis that clearly delineates bifurcating versus nonbifurcating members of this family. Structural modeling of proteins within these groups reveals key differences between the bifurcating and nonbifurcating Etfs.

KEYWORDS electron-transferring flavoprotein, flavin, nitrogen fixation, nitrogenase, electron bifurcation

Electron bifurcation, or the coupling of exergonic and endergonic redox reactions to simultaneously generate (or utilize) low- and high-potential electrons, has been proposed as the third fundamental form of energy conservation (1,2). Bifurcating (Bf) enzymes play central roles in the energy metabolism of anaerobic bacteria and archaea, allowing them to reduce the low-potential [4Fe-4S] clusters of ferredoxin (Fd) using higher-potential electron donors. To date, a total of nine Bf enzymes have been identified. The nine currently demonstrated bifurcating enzymes employ diverse substrates and feature in a wide variety of pathways but share the common features of employing NAD(P)H and Fd as redox substrates/products and possessing at least one flavin, proposed to be the site of bifurcation (1,3).

The first bifurcating enzyme to be characterized in detail was electron-transferring flavoprotein (Etf)-butyryl coenzyme A (butyryl-CoA) dehydrogenase (Bcd) from *Clostridium kluyveri* and *Acidaminococcus fermentans* (4–6). Etf-Bcd was shown to couple the oxidation of NADH (E_m = -320 mV) to the simultaneous endergonic reduction of Fd (E_m = -500 mV) and exergonic reduction of crotonyl-CoA (E_m = -10 mV) during acetate or ethanol (*C. kluyveri*) or glutamate (*A. fermentans*) fermentation (6) (Fig. 1). Structural and biochemical analysis of the Etf complex from *A. fermentans* revealed that it consists of two subunits, with the subunit coordinating one flavin adenine dinucleotide (FAD) (α-FAD) and the subunit coordinating a second FAD (β-FAD) (6) (Fig. 2). The β-FAD was established to be the site of hydride acceptance from NADH and electron bifurcation, because NADH binds close to this cofactor (6).

Aside from proposals that a flavin is the site of electron bifurcation, key details of the mechanism of electron bifurcation were lacking until a recent biophysical study of $NADP^+$ oxidoreductase (Nfn) (3), which drives the simultaneous endergonic reduction of Fd and exergonic reduction of NAD^+ based on oxidation of NADPH (7). Therein, it was proposed that reduction of Fd is achieved by an unstable flavin anionic semiquinone generated by one-electron oxidation of the doubly reduced flavin hydroquinone formed by hydride transfer from NADPH. Thus, a favorable single-electron transfer from flavin hydroquinone "pays for" production of a much more strongly reducing species that is able to reduce Fd. Moreover, it was proposed that electron transfers from the Bf flavin are mediated by iron-sulfur clusters along two physically separate electrons transfer paths whose constituent electron carriers have potentials that span more than 1 V. Given that flavins are common to all known NAD(P)H-utilizing Bf enzymes and that they were proposed (1) and then shown (3) to be the site of bifurcation, it is plausible that unstable flavin anionic semiquinones and additional redox cofactors that differ markedly in their potentials are involved in other Bf systems as well. If so, it is possible that the protein environment of the Bf flavin will be found to produce similar flavin reactivities in diverse bifurcating enzymes, including destabilization of the semiquinone state of flavin and provision of efficient paths of electron transfer as requisites for bifurcation. Thus, the precedent of Nfn suggests that a Bf flavin should be positioned near a binding site for NAD(P)H as well as cofactors or amino acid side chains able to mediate efficient electron transfer (3). Therefore, we hypothesized that such features should be present in Bf Etfs but not necessarily in Etfs in general.

The Bf EtfAB module of the Bf Etf-Bcd complex belongs to a larger group of Etfs, many of which have been biochemically characterized and shown to execute single-electron transfer only, not bifurcation. Members of the larger group include the well-studied mitochondrial EtfAs and EtfBs that are involved in lipid and amino acid metabolism (8). This family of Etfs has been shown to accept electrons from at least nine different acyl-CoA dehydrogenases (9). The acyl-CoA dehydrogenases capture electrons extracted during oxidation of CoA-conjugated fatty acids or oxidation of amino acids. Etfs convey those electrons to another flavoprotein, Etf-quinone oxidoreductase (Etf-QO), which in turn passes them into the respiratory electron transport chain via the reduced quinone pool. Hence, Etfs function as an electron funnel in which electrons from a variety of fatty acid or amino acid substrates are channeled into the electron transport chain via Etf-QO (9).

In addition to the Etfs that work with Etf-QO, some bacterial genomes encode Etfs that are not involved in fatty acid or amino acid oxidation or quinone reduction (10–13). For example, Etfs involved in trimethylamine oxidation by a methylotrophic bacterium (10, 14, 15) and carnitine oxidation by Escherichia coli (11, 16) have been described. These Etfs are not associated with Etf-QO but rather are associated with a substrate specific dehydrogenase (16, 17). Other Etf homologs were identified in the nitrogen-fixing bacterium *Sinorhizobium meliloti*; the genes were termed *fixA* and *fixB* because mutations in them rendered *S. meliloti* unable to fix nitrogen (12, 18, 19). Regrettably, FixA and FixB correspond to Etf-β and Etf-α, respectively. These nitrogen fixation-associated Etfs have a conserved gene organization in which the genes coding

for Etf-α and Etf-β are immediately followed by the genes *fixC* and *fixX*, encoding an oxidoreductase and a Fd-like protein, respectively. FixC and FixX, when concatenated, are homologous to Etf-QO and have been shown to be essential for nitrogen fixation in *Rhodopseudomonas palustris* (20), where they presumably accept electrons from FixA and FixB (12) (Fig. 1). FixABCX was proposed (12, 21) and then shown (22) to oxidize NADH to generate reduced quinone, which feeds the respiratory chain, and reduced Fd, which is the reductant of nitrogenase for N_2 fixation.

Structural characterization of Etfs from human mitochondria as well as from the bacterium *Paracoccus denitrificans* revealed that the two Etf subunits, Etf-α and Etf-β, are tightly associated in a heterodimer composed of three different domains (Fig. 2) (23,24). Domain I is formed by the N-terminal portion of the subunit and does not appear to be involved in cofactor binding. Domain II is formed mostly by the C-terminal region of the subunit, with some contribution from the C-terminal region of the subunit. A single FAD cofactor is coordinated by residues found primarily within the C terminus of the subunit, in a region that is highly conserved among all known Etfs. Domain III is formed by most of the subunit and, in non-Bf Etfs, it coordinates an AMP moiety thought to stabilize the protein but not to play a redox function (10, 24). The Bf Etf-Bcd differs structurally and mechanistically from non-Bf Etfs (1), in that the bound AMP is replaced by a FAD that binds to domain III with its flavin at the interface, where it also interacts with residues in domain I (4–6, 25) (Fig. 1 and 2). Indeed, a previous analysis suggested that a peptide motif from domain III near the binding site of the Bf FAD might differentiate Bf and non-Bf Etfs (26).

Although it is clear from work carried out over the past 30 years that Etf-associated enzymes are widely prevalent in biology and play diverse roles in metabolism, a comprehensive perspective on the diversity of these enzymes is lacking. Here we analyze the phylogenetic and structural variations among Etf proteins in genome sequence databases to discern putative functional variations within this group of enzymes and to identify sequence motifs that potentially signify and even underlie the ability of members of this class of enzymes to bifurcate electrons. A new scheme that classifies Etfs into five phylogenetic and functionally coherent groups is presented. Structural analyses of representatives of these groups reveal amino acid motifs that may be diagnostic and predictive of the ability of Etfs to catalyze electron bifurcation reactions.

RESULTS AND DISCUSSION

Diverse archaea and bacteria encode Etf-α and Etf-β homologs. Identified Etf-α and Etf-β homologs were first examined at the level of the automated annotations specified in the gene ontology. In most groups, automated or manual annotation during genome curation correctly identified the sequences as Etfs. In some cases, however, Etf gene sequences were incorrectly identified or annotated as *fixA* or *fixB*. For example, the α subunit of Etfs (Etf-α) encoded by E. coli genomes and other enterobacteria was consistently annotated as "nitrogen fixation protein FixB" in NCBI databases, although these enterobacteria do not encode the other component needed to fix nitrogen. The annotated *fixA* and *fixB* genes in these taxa were often flanked by genes annotated as *fixC* or *fixX*. Homologs in actinobacteria were often annotated as *fixA* or *fixB* although homologs of *fixC* and *fixX* were not identified in the genome. Based on numerous

instances of inconsistency in genome annotation, we subjected homologs identified in genome sequence databases to thorough taxonomic and phylogenetic analyses, to develop a more consistent framework for describing and annotating Etf proteins.

A total of 1,314 homologous Etf sequences were identified from 890 bacterial and archaeal genomes (see Table S2 in the supplemental material). Of the 51 archaeal genomes that encoded Etfs, 45% (i.e., n = 23) belonged to the phylum Crenarchaeota, while 55% (n = 28) belonged to the phylum Euryarchaeota. Similarly, a total of 839 bacterial genomes encoded Etfs. The majority of Etfs were found in the phyla Proteobacteria (i.e., 52% of the total of 839 bacterial genomes), Firmicutes (i.e., 17% of the total bacterial genomes), and Actinobacteria (i.e., 15% of the total bacterial genomes) (Table S2). However, the identification of genes encoding Etf-α, Etf-β, FixA, and FixB homologs in numerous phyla previously not known to possess those genes points to their likely widespread importance in cellular metabolism and suggests that they may have a broader array of functions than currently known. In support of this notion, the numbers of copies of *etfα* and *etfβ* genes per genome varied from one in about one-third of the genomes to as many as eight or nine in the genomes of some species of Azoarcus, Desulfitobacterium, Geobacter, and Burkholderia (Table S2).

Phylogenetic analyses support five distinct Etf groups. Maximum-likelihood phylogenetic reconstruction of 1,314 concatenated Etf-α/β sequences and analysis of their relatedness resulted in five well-defined clades, which were designated group 1 (G1) to G5 (Fig. 3A and Table 1). G1 sequences include Etf-α and Etf-β from *P. denitrificans* and human mitochondria that have been biochemically characterized; these largely

correspond to the canonical housekeeping Etfs (23, 24). This group of proteins encompasses the largest number of sequences (473 of 1,314 sequences) and contains representatives of over 300 different Alphaproteobacteria, Betaproteobacteria, Deltaproteobacteria, and Gammaproteobacteria species. About 70 organisms in this group possess multiple G1 Etfs. These organisms also have genes corresponding to *etfQO*, but they are often separate from the *etf* genes. Furthermore, there is no biochemical evidence to suggest involvement of any of these enzymes in electron bifurcation; informatic analyses (presented below) support this proposal.

G2 sequences are of interest since they include Etf-α and Etf-β from *A. fermentans*, *Megasphaera elsdenii*, *Acetobacterium woodii*, and *C. kluyveri*, all of which have been shown to function in electron bifurcation and are proposed to coordinate two flavins (4, 6, 26, 27). All of the Etfs known to bifurcate electrons cluster in this group, which nonetheless exhibits substantial phylogenetic diversity, as its 311 included Etf sequences are derived from 15 different phyla (Table 1; also see Table S2).

Functional diversity among G2 enzymes is indicated by the emergence of five subgroups based on phylogenetic clustering and gene neighborhood analyses (described below), with several subgroups being further divided into subsets (Fig. 3B and Table 1; also see Table S2). G2A and G2B consist primarily of Etfs from Firmicutes (i.e.,85% and 56%, respectively, of the total homologous sequences) and include all of the biochemically characterized Etfs that are known to bifurcate. G2C1 consists exclusively of archaeal Etfs, all affiliated with the phylum Crenarchaeota. Interestingly, genes coding for the Etf-α and Etf-β homologs in the genus Sulfolobus are fused and form a single

open reading frame, which may indicate that these proteins form a functional complex in a manner similar to that described for several maturase proteins involved in nitrogen fixation (28) and hydrogen metabolism (29). G2C2 Etf genes were identified in 52% of the Bacteroidetes genomes; these proteins represent 50% of the total G2C2 sequences. G2D1 genes primarily include sequences from Firmicutes genomes (i.e., 66% of the total G2D1 sequences), and many of the organisms in this group also encode G1 Etfs. G2D2 Etfs are predominantly found in phylum Proteobacteria (i.e., 98% of the total G2D2 sequences). These G2D2 Etfs are classified as Fix and were proposed (12, 21) and then shown (22) to function in coupling the oxidation of NADH to the formation of quinol and reduced Fd as a reductant for nitrogen fixation through the process of electron bifurcation (21). In support of a role for G2D2 Etfs (Fix) in delivering reduced Fd to nitrogenase, all of the genomes that encode G2D2 Etfs also encode the minimal set of proteins required to fix nitrogen via molybdenum-dependent nitrogenase (i.e., NifHDKENB) (28, 30). Lastly, all G2E Etfs are affiliated with the phylum Proteobacteria. A representative of G2E Etf (i.e., Gmet_2067/2066 from *Geobacter metallireducens*) has been proposed to be capable of electron bifurcation (31). These Etfs also harbor a peptide motif in the β-FAD cofactor environment that has been proposed to allow for bifurcation (26), which is described in more detail below. A small number of genes encoding G2E Etf-α and Etf-β homologs are flanked by genes encoding homologs of FixC and FixX (described below), suggesting that they may be involved in nitrogen fixation, based on the observation that several of these genomes also encode NifHDKENB.

G3 Etfs include sequences belonging to a variety of bacterial and archaeal phyla (Table 1; also see Table S2). This group forms four subgroups (i.e., G3A to G3D). G3A and G3B Etfs are primarily found in members of Proteobacteria (100% and 70%, respectively, of the total sequences in each group). G3B includes the characterized Etf involved in carnitine oxidation in *E. coli* (10, 14, 15). Interestingly, many enterobacteria have a G3B Etf coding sequence that has been annotated as *ydiQRST* (represented as *etfABCX* in Table 1) and is associated with *fadK* (*ydiD*). Genetic studies have shown that *ydiQRST* and *ydiD* are required for the capacity to oxidize fatty acids under anoxic conditions in *E. coli* (11, 16). G3C-like Etfs were enriched in members of the phylum Euryarchaeota (i.e., 54% of the total G3C sequences), although they were also identified in bacteria; this group includes a protein from a methylotrophic bacterium that is involved in trimethylamine oxidation (10, 14, 15). Lastly, all G3D Etfs belonged to members of the phylum Actinobacteria. Based on a lack of conserved residues correlated with the capability to bifurcate in other groups (described in more detail below), we surmise that these enzymes are unlikely to bifurcate.

G4 Etfs and G5 Etfs are confined to the bacterial phyla Bacteroidetes and Firmicutes, respectively. Representative enzymes from G4 and G5 have yet to be biochemically characterized. However, based on the similarity of the proteins encoded in their gene neighborhoods (described below), it is possible that their functions are analogous to those of Etfs in G1 (Table 1). Moreover, 40 Etf homologs could not be readily classified into any of these phylogenetic groups and hence were left unclassified

(Table S2). These unclassified Etfs primarily belonged to the phylum Deinococcus-Thermus.

Gene neighborhood analyses associate metabolic processes with Etf function. To investigate the potential functions of the diverse Etfs uncovered here and to identify potential determinants that define Bf and non-Bf enzymes, we examined the gene neighborhoods flanking Etf genes for each group. Unfortunately, no genes were consistently identified in the flanking regions of G1 Etf genes, precluding a prediction of the functionality of these Etfs based on gene neighborhood analysis. However, all G1 Etf-encoding genomes encoded ubiquinone oxidoreductase (QO) elsewhere in the genome (data not shown); Etf-QO is involved in fatty acid metabolism (32). Therefore, we speculate that G1 Etfs are involved in fatty acid metabolism.

G2 Etfs include characterized bifurcating Etfs (3, 5, 19, 33). Each G2 subgroup was accompanied by a unique suite of proteins encoded by genes in its gene neighborhood (Fig. 4). The gene for acyl-CoA dehydrogenase, which is involved in fatty acid oxidation (9), was identified in the flanking region of 76% and 26% of the genomes that encode G2A and G2B Etfs, respectively. This indicates potential involvement of these Etfs in electron transfer reactions involving fatty acid metabolism. In support of this hypothesized role, G2A and G2B Etfs have been found to be involved in electron transfer reactions during the oxidation of butyrate and lactate (4, 25). The neighborhoods of G2C1 Etf genes included genes for homologs of FixX, FixC-like, and MoaD-like enzymes. MoaD homologs have been shown to be involved in molybdenum/tungsten cofactor biosynthesis or thiamine biosynthesis (34, 35). Thus, it is possible that G2C1

Etfs participate in electron transfer reactions contributing to cofactor or vitamin biosynthesis. All genes for subgroups of G2D Etfs encoded FixC and FixX-like enzymes in their gene neighborhoods. The neighborhoods of 27% of the G2D1 Etf genes included genes for homologs of carbon monoxide dehydrogenase and acetyl-CoA synthase, which raises the intriguing possibility of involvement in CO or CO_2 metabolism (36). G2D2 Etfs were all derived from putatively diazotrophic genomes (i.e., the genomes also encoded NifHDKENB). Mutations in Etf gene subunits suppressed growth under nitrogen-fixing conditions, which led to the Etf genes being named *fix* (*fixABCX*) in nitrogen-fixing microbes (12, 21, 33, 37, 38). Lastly, the neighborhoods of 26% of the G2E Etf genes encoded acetyl-CoA dehydrogenase, acetyl-CoA acetyltransferase, the nitrogenase co-factor biosynthesis proteins NifE and NifN, and aldehyde oxidoreductase. It is not clear from gene neighborhood analysis what functional role these Etfs may have.

G3 Etfs were divided into four groups, with G3A deriving from aerobic and facultatively anaerobic organisms. The gene environments for these Etfs included genes for NADH oxidase and multiple proteins containing iron-sulfur clusters (Fig. S1). G3B Etf genes were found to be accompanied by genes for FixC-like and FixX-like enzymes, carnitine dehydratase, and acyl-CoA dehydrogenase in the gene neighborhood, supporting the idea that they are involved in carnitine metabolism (11, 14). The neighborhoods of 32% of G3C Etf genes encoded homologs of acetyl-CoA dehydrogenase and acetyl-CoA acetyltransferase, both of which are involved in fatty acid oxidation. In addition, the neighborhoods of 29% of G3C Etf genes encoded homologs of ubiquinone methyltransferase, which is involved in menaquinone biosynthesis (39); thus,

these Etfs may support biosynthesis of menaquinone. Genes that flanked G3D Etf genes commonly encoded cysteine desulfurase (73% of G3D Etfs), Asn/Gln amidotransferase (62% of G3D Etfs), and Gln amidotransferase (60% of G3D Etfs), suggesting a role for these flavoproteins in electron transfer during amino acid biosynthesis (Fig. S1).

The neighborhoods of G4 Etf genes did not reveal any conserved (>50% relative frequency) genes. However, among the more commonly observed genes flanking G4 Etf genes were those encoding pyruvate dehydrogenase (41% of G4 Etfs) and thymidylate synthase (41% of G4 Etfs). Lastly, the neighborhood of G5 Etf genes included homologs of genes for fumarate reductase (52% of G5 Etfs) and the cytochrome b_{558} subunit of succinate dehydrogenase (52% of G5 Etfs) (Fig. S1).

Structural modeling reveals key differences between bifurcating and nonbifurcating Etfs. Regardless of their putative roles in metabolism (Table 1), Etf proteins participate in two mechanistically distinct sets of redox reactions, namely, reactions that bifurcate electrons and reactions that do not bifurcate electrons (4–6, 26, 40). From the recent analysis of the mechanism of the flavoenzyme Nfn (3), it was established that FAD was the site of bifurcation, and we refer to this as the Bf FAD. This flavin accepts a pair of electron equivalents from NADPH and donates one to each of two different electron acceptors.

We interrogated our sequence alignments for motifs and individual residues that could differentiate Bf from non-Bf enzymes. Toward these ends, we carried out structural homology analyses with sequences belonging to G1 and G2 Etfs. Available evidence from biochemical characterization of selected members of these two groups suggests that

G1 Etfs do not bifurcate (23, 24), whereas several G2 Etfs do (4, 6, 26, 27), making comparisons between these two types of Etfs especially attractive. A crucial difference is that the Etfs that bifurcate have been demonstrated to contain two FADs, whereas Etfs that do not bifurcate have been found to contain only one. Moreover, bifurcating Etfs can accept electrons from NADH, whereas those that do not bifurcate cannot. Indeed, our analyses revealed conserved amino acid differences affecting residues that interact with NAD(H), based on a recent crystal structure (6) (motif 1), and a motif that coordinates with β-FAD (motif 2), which is the proposed site of bifurcation in Etfs (1, 4).

We examined conserved differences at each position, looking for instances in which sequences from G1 or G2 share a unique residue within their group that is distinct from the type present in the other group (Fig. 5). Figure 5A depicts the functional distinctions between nonbifurcating (G1) and bifurcating (G2) Etfs. The structural comparison reveals that the motifs identified previously are clustered near the β-FAD binding site and indeed are differentially conserved in G1 versus G2 Etfs (Fig. S2). For example, β-Gly^{127} (using *R. palustris* numbering) represents a conserved difference in amino acid type; it is a small residue that minimizes steric conflict with the flavin in motif 2 in all G2 proteins, but it is replaced by a larger and generally anionic residue in G1. This negative charge would be expected to repel the phosphate of β-FAD (Bf FAD in G2 enzymes) and thus disfavor binding. In addition, G1 proteins tend to have bulkier residues overall around this residue, further predicting a diminished capacity to bind the flavin mononucleotide (FMN) portion of β-FAD. Thus, our approach

reproduces previous findings but adds residues that are nearby in the 3-dimensional protein structure but not necessarily in the amino acid sequence.

Other conserved distinctions in the region in which the β-FAD binds to G2 Etfs are observed in the Etf-α subunit, where two strands of the α chain interface with the β chain and are predicted to participate in the binding of the flavin moiety of β-FAD (Fig. 5C). Thus, a structurally coherent pattern of amino acid conservation supports a tendency for FAD binding in G2 versus AMP binding in G1. Figure 5B depicts a second comparison, in which bifurcating G2 Etfs are compared with predicted bifurcating G2D2 Etfs (also known as Fix). This structural comparison shows that G2D2 Etfs share the motif identified in bifurcating Etfs in G2A and G2B (Fig. 5B), strongly suggesting that G2D2 Etfs (or Fix), like the rest of the G2 subgroups, are capable of electron bifurcation. The region displaying the most blue-colored residues and the most residues with intermediate constancy (shades of purple) coincides with the recognition loop that has been found to interact with partner proteins (41). This is consistent with G2D2 Etfs interacting with FixC and FixX while the genes of some other bifurcating Etfs (G2A and G2B) are not accompanied by genes for FixC or Etf-QO homologs, suggesting that these Etfs have different partners.

A region found to interact with NAD^+ in the crystal structure of the *A. fermentans* Bf Etf (6) (motif 1) also exhibits a contiguous string of conserved differences between Bf and non-Bf Etfs (Fig. 5A). This region includes the peptide bond between β-Phe^{89} and β-Ala^{90}, which π stacks with the NAD^+ adenine ring, the backbone amide of β-Asp^{86}, which forms hydrogen bonds with the adenine ring, and the backbone amide of β-Gly^{91},

which forms hydrogen bonds with the α-phosphate of NAD^+ (Fig. 5C). These combined conserved differences in the region of the β-FAD flavin-binding pocket all suggest that NADH is a substrate for G2 but not G1 Etfs.

To investigate whether motifs 1 and 2 that are conserved in G2A and G2B Etfs are conserved in other Etf groups, we aligned the corresponding consensus sequences from each of the subgroups that emerged from our database (Fig. 5D). As expected, nearly all of the G2 sequences exhibited conservation in the residues that were identified as being important in distinguishing Bf G2 sequences from non-Bf G1 sequences, whereas the same residues were not conserved in the sequences belonging to the other groups. Importantly, our informatic analyses suggested that more members of G2A and G2B may be found to be capable of electron bifurcation, providing the impetus for biochemical studies of Etfs from G2C (primarily found in Archaea) and G2D (including the enzymes that were hypothesized [12, 21] and then shown [22] to be involved in providing reductant for nitrogen fixation), to test for bifurcation activity. Importantly, there are likely exceptions or limitations to the predictions of which enzymes can bifurcate, based on conservation of the aforementioned amino acid sequence motifs. For example, *Clostridium propionicum* has been shown to have an insertion in Etf- that prevents bifurcation, although the sequence for this Etf has the identified conserved motifs that predict bifurcation capability (26).

Our structural comparison predicts that only G2 Etfs have the capacity to bifurcate, which raises the questions of how and when bifurcating Etfs evolved. Since our

phylogenetic tree is not rooted (i.e., it is a protein family tree consisting of all paralogs), we cannot use it to predict the origin of bifurcating Etfs. However, the Etf protein family phylogeny does suggest that, once the ability to bifurcate electrons arose (as either an ancestral or derived characteristic), this activity was maintained (i.e., G2 Etf all share an apparent ability to bifurcate). Both archaeal and bacterial genomes encode putative bifurcating G2 Etfs; however, the archaeal versions known so far all occur in G2B and G2C. These lineages branch closer to the crown of the G2 lineage, suggesting that G2 Etfs diverged from a bacterial ancestor. Furthermore, the majority of G2C and G2D Etfs, which have genes encoding FixC- and FixX-like proteins in their neighborhoods, diverged after G2A and G2B. Parsimony would suggest that fixC- and fixX-like genes were acquired during the evolution of G2 Etfs. Importantly, the genomes of several members of G3B also encode FixC- and FixX-like subunits, and these are nested among G3 sequences that lack these subunits; this suggests that these genes were acquired independently multiple times during the evolution of Etfs. These collective observations, including those showing that Etf-α and Etf-β can form complexes with other enzymes, such butyryl-CoA dehydrogenase (4), crotonyl-CoA dehydrogenase (4), and pyruvate dehydrogenase (42), underscore the versatility of the Etf-α/β protein "chassis" in the emergence and evolution of new metabolic processes in microorganisms.

Overall, the Etfs provide a unique opportunity to compare many amino acid sequences for proteins that are structural homologs but represent two distinct functional capacities, i.e., single-electron transfer via a single FAD versus electron bifurcation within a system containing two FADs in the Etf plus additional cofactors in the

associated enzyme complex. The diverse genomic contexts in which Etf genes are found may reflect their versatility and the essential nature of the energy and electron transactions they enable. However, the distinction between single-electron transfer and electron bifurcation is critical in energy conservation. Therefore, the identified protein characteristics that demarcate Bf and non-Bf groups of Etfs provide exciting predictions regarding which Etfs should be tested for their ability to bifurcate and which residues and types of residues provide the protein environment allowing for this capability.

Concluding remarks. Our analysis shows that Etf enzymes are phylogenetically diverse and widely distributed in archaea and bacteria, consistent with the essential roles of these enzymes in the metabolism and fitness of numerous strains (43, 44). Many strains encode more than one Etf, which is also consistent with the roles of Etfs in diverse metabolic processes. These Etf homologs (or paralogs, if they are not functionally redundant) sometimes belong to the same phylogenetic group, e.g., multiple G3B Etfs in *E. coli* and multiple G2E Etfs in *Geobacter* spp., but in other instances belong to distinct groups, e.g., many proteobacteria with G2D2 Etfs also encode G1 Etfs and Crenarchaeota with G2C1 Etfs also encode G3B Etfs. The vast phylogenetic diversity and numerous copies of Etfs in microbial genomes suggest that they have unexplored functional diversity, which could shed light on the metabolism of these organisms. In several cases, predictions can be made regarding the potential functions of these uncharacterized Etfs based on proteins encoded by adjacent genes, as discussed above. These predictions provide a roadmap for future biochemical studies aimed at characterizing the functional diversity of this enzyme family.

The demonstration of bifurcation for Etfs of anaerobic bacteria (4–6) significantly advanced our understanding of the interplay among processes that contribute to energy conservation in fermentative metabolism. Perhaps our most intriguing findings here are the identification of residues that apparently enable bifurcation capability in Etfs in phylogenetically and physiologically diverse organisms that extend beyond those recognized previously (26) and residues dispersed in the sequences of both subunits. Indeed, ours are the first such analyses to integrate information from putative Bf Etfs obtained from a comprehensive database of sequences from nonfermentative archaeal and bacterial taxa as well as those putatively associated with nitrogen fixation. The identification of bifurcation-associated residues provides a foundation for future biochemical studies aimed at elucidating the specific mechanism by which FAD transfers electrons to acceptors with differing midpoint potentials. Bifurcating Etfs need to provide a protein environment that potentiates the states of FAD likely required during bifurcation, such as the proposed unstable anionic semiquinone. Our study identifies residues coordinating FAD that could play such a role

MATERIALS AND METHODS

Generation of an Etf/Fix database. Homologs of Etf-α were identified using a BLASTp search (cutoff E value of 10^{-5}) against completed bacterial and archaeal genomes available in the Department of Energy Integrated Microbial Genomes (IMG) database (45), with FixB (Etf-α) from *Azotobacter vinelandii* (STRING network identifier: Avin_10530) serving as the query. A total of 1,314 Etf-α homologs across 890

genomes were identified. Homologs of Etf-β (compiled from the IMG database for each Etf-α sequence) and Etf-α were aligned individually using Clustal Omega (46), and the two resultant alignment blocks were concatenated.

Protein clustering and phylogenetic analyses. The concatenated Etf-α and Etf-β sequences were subjected to maximum-likelihood phylogenetic reconstruction with RaxML (version 7.3.0) (47), specifying the LG substitution matrix, the PROTGAMMA option, and 100 bootstrap iterations. The tree was projected with FigTree. In addition, a custom Python (version 2.7.6) script was used to extract gene sequences (20 upstream and 20 downstream) that flanked each Etf-β. The 40 genes extracted were translated in silico, and the protein sequences were clustered using identity thresholds of 90%, 60%, and 30% while holding the pairwise sequence coverage threshold constant at >60%. The clusters generated by the three-step clustering method were later combined to yield a final "averaged" cluster identity.

Network analysis. Protein sequence clusters obtained from the earlier step were used to generate a binary matrix describing the presence or absence of genes. The binary matrix generated was then used to identify the abundant proteins (i.e., present in >=50% of the genomes) in each group to create a network plot, with Cytoscape specifying the force-directed organic layout (58). Each unique protein or group identified was denoted as a node, and the edges between the nodes represent the abundance of the genes in the group.

Structural modeling. Amino acid diversity was scored as the number of different amino acids found in an alignment position. Functional diversity at each aligned position was described in terms of five defined types of residues, i.e., negatively charged residues, including Asp and Glu; positively charged residues, including Lys and Arg; hydrophobic residues, including Ile, Leu, Val, Met, and Phe; and polar uncharged residues, including Gln, Asn, Ser, His, Thr, Cys, Ala, and Gly (48). Trp, Tyr, and Pro were considered to be special, based on their ability to mediate electron transfer within proteins (49) and their ability to influence protein dynamics and folding (50).

Structural models were generated by using the consensus sequence and online tools provided by the SwissModeler suite (51, 52) (see the supplemental material for the consensus sequences used in this study). Etf-α and Etf-β protein templates were chosen from the same complex to optimize preservation of the interface between monomers, based on metrics of structure quality captured by Qmean4 (53, 54). Monomer coordinates were combined with cofactor coordinates from the template, and the resulting holoproteins were assessed on the basis of the number of clashes between protein and FAD molecules and between monomers, using tools provided within the Chimera modeling software package (55). Holoprotein dimers with the fewest clashes and clashes derived mainly from side chain interactions were subjected to energy minimization after the addition of H atoms and the assignment of charges (56), as implemented in Chimera (55). The resulting minimized holoprotein dimer models were again evaluated with respect to clashes, and only models free of clashes were used for further work. The

templates used for each model are reported in Table S1 in the supplemental material, along with metrics of the quality of the models.

To identify conserved differences between groups of sequences, we compared the prevalence of amino acid types at analogous positions. At each position, we constructed a vector in the 5-dimensional space defined by the five functional types of amino acids. The five components of the vector represented the prevalence of each of the five functional types of amino acids, as outlined above. To determine the extent of similarity at each position of the two groups being compared, we calculated the scalar product of the corresponding vectors. Thus, conservation of the same functional type at an alignment position (parallel vectors) yields the maximum value of 1, whereas conservation of different functional types in each of the two groups yields a minimum value of 0. Positions characterized by similar variations among functional types in the two groups of sequences yield high values for the scalar product because the two vectors point in the same direction in the 5-dimensional space, even though no one type is conserved. To highlight positions where a single functional type is conserved in the reference group, we displayed the degree of type conservation using worm width, with wide worm segments for positions at which the type is conserved and thinner worm segments for positions at which the type varies.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JB.00440-17

.

SUPPLEMENTAL FILE 1, PDF file, 1.0 MB.

SUPPLEMENTAL FILE 2, XLSX file, 1.6 MB

ACKNOWLEDGMENTS

This work was supported as part of the Biological Electron Transfer and Catalysis Energy Frontier Research Center, funded by the U.S. Department of Energy, Office of Science, Basic Energy Sciences, under award DE-SC0012518.

Table 1. Designation, taxonomic distribution, and proposed functions of Etf groups.

GROUP	SUBGROUP	DOMINANT TAXA (% total)	NO. HOMOLOGs/ NO. GENOMES	GENE NAME	PROPOSED FUNCTION†
Group 1	1	Proteobacteria (100)	473/360	*etfAB*	Fatty acid oxidation
Group 2	2A	Firmicutes (85)	81/63	*etfAB*	Fatty acid oxidation / Fermentation (butyrate)‡
	2B	Firmicutes (56)	53/49	*etfAB*	Fatty acid oxidation / Fermentation (lactate‡)
	2C1	Crenarchaeota (100)	15/15	*etfABCX**	Mo/W cofactor biosynthesis / Vitamin B synthesis
	2C2	Bacteroidetes (50)	50/48	*etfABCX/ etfAB*	Unknown
	2D1	Firmicutes (66)	35/35	*etfABCX*	CO_2/CO metabolism
	2D2	Proteobacteria (98)	55/53	*fixABCX*	Nitrogen fixation‡
	2E	Proteobacteria (100)	23/10	*etfABCX**	Nitrogen fixation / Fatty acid-aldehyde metabolism
Group 3	3A	Proteobacteria (100)	76/76	*etfAB*	Amino acid/Redox metabolism
	3B	Proteobacteria (71)	66/51		Carnitine metabolism‡ / Anaerobic fatty acid metabolism
	3C	Euryarchaeota (54)	76/61	*etfAB*	Fatty acid oxidation/ Menaquinone metabolism
	3D	Actinobacteria (100)	137/130	*etfAB*	Amino acid metabolism
Group 4	4	Bacteroidetes (100)	51/51	*etfAB*	Pyruvate metabolism / Nucleotide metabolism
Group 5	5	Firmicutes (100)	83/70	*etfAB*	Fatty acid oxidation / Amino acid metabolism
Unclassified	NA	Deinococcus-Thermus (45)	40/39	*etfAB*	N.A.

[a] The number in the parentheses represents the percentage of the total sequences within a designated group or subgroup that is within the specified phylum. For simplicity, only the dominant phyla are shown. See Table S2 in the supplemental material for the remaining phyla that ecodes Etfs in each group.
[b] A given genome may encode multiple Etf homologs and thus be counted multiple times in this table. The total number of genomes encoding Etf homologs was 890.
[c] Based on the annotated function of genes colocalized (±20 open reading frames) with *etfAB*, *etfABCX*, or *fixABCX*.
[d] The link of Etf to this metabolic function is supported by genetic and/or biochemical studies.
[e] *etfA* and *etfB* are fused.

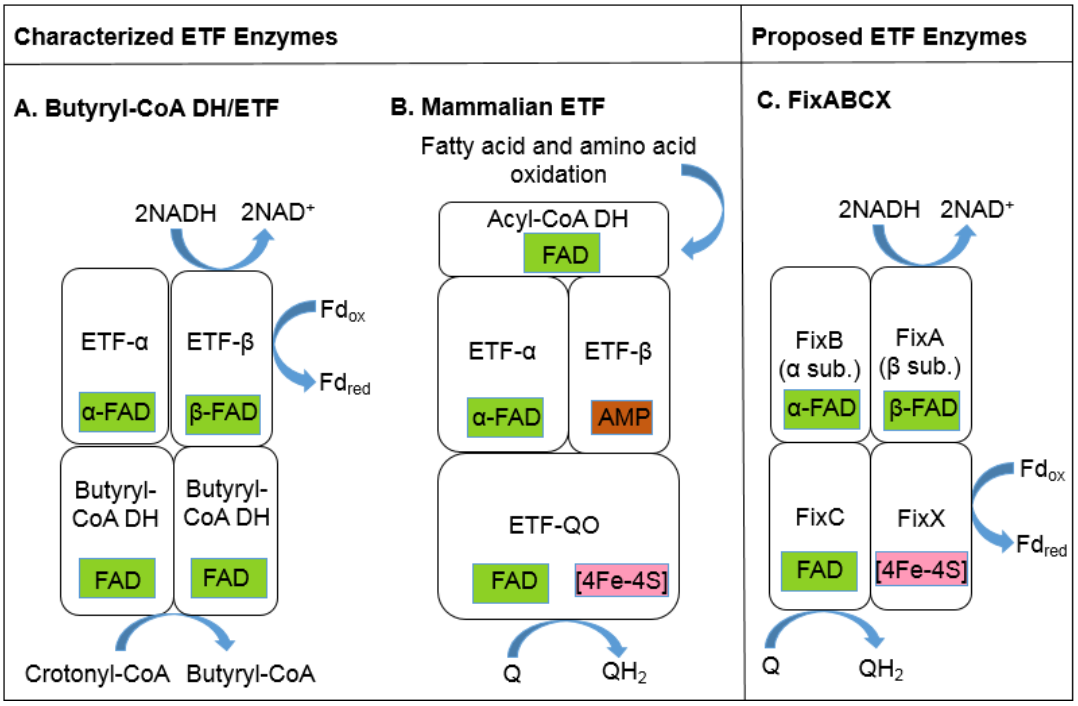


FIG 1 Proposed subunit and cofactor composition and reaction scheme for Etfs. (A) Proposed electronbifurcating electron transfer mechanism for Etfs from anaerobic fermentative firmicutes (6), in which the β-FAD in Etf-β accepts electrons from NADH and bifurcates them to Fd and to the α-FAD in Etf-α. Etf-α donates electrons to Bcd, which reduces crotonyl-CoA to butyryl-CoA. Notice the absence of a FixX equivalent in this model. (B) Non-Bf electron transfer mechanism in mammalian Etfs. Electrons from the oxidation of fatty acids or amino acids are donated to acyl-CoA dehydrogenase (DH), which passes them to the FAD in Etf-α. Etf-α transfers those electrons to the [4Fe-4S] cluster in Etf-QO, which then reduces a FAD cofactor also in Etf-QO (57). Lastly, this FAD reduces quinone (Q). Note that Etf-β contains a redox-inert AMP cofactor. (C) Electron-bifurcating electron transfer mechanism proposed for FixABCX (2). NADH has been proposed to donate electrons to the β-FAD in FixA, which bifurcates electrons to an [4Fe-4S] cluster in FixX and to the α-FAD in FixB. α-FAD is proposed to subsequently reduce another flavin cofactor (FAD) located in FixC. Lastly, FixX and FixC are proposed to reduce Fd and quinone, respectively.

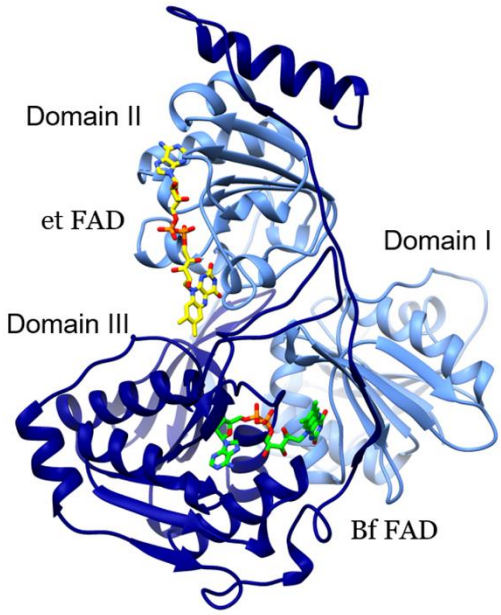


FIG 2 Ribbon diagram of EtfA and EtfB, showing the model of G2 Etfs generated from the consensus sequence based on sequences provided Table S1 in the supplemental material. Light blue indicates Etf-α and dark blue indicates Etf-β; the domains are also labeled. Stick models depict the Bf FAD (green C atoms) and the electron transfer (et) FAD (yellow C atoms). The figure was generated using Chimera (55).

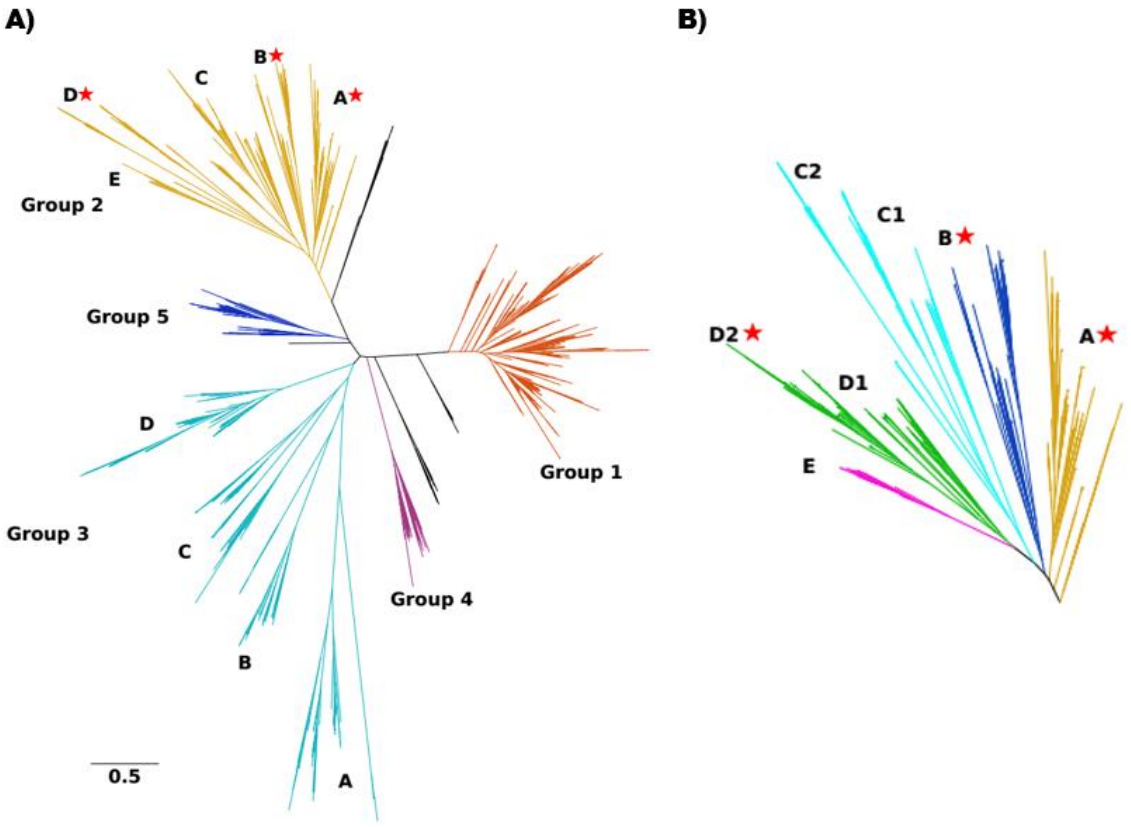


FIG 3 Maximum-likelihood phylogenetic tree depicting relationships among 1,314 concatenated Etf-α/β sequences. (A) Five distinct phylogenetic groups (shown in distinct colors) that emerged from phylogenetic reconstruction of Etf-β and Etf-α. Sequences that did not belong to any of those groups were left black and are not discussed further (see Table S2 in the supplemental material). Sequences in G2 and G3 were further classified into subgroups (denoted by uppercase letters). (B) Closeup of G2 subgroups, showing further diversity among lineages putatively involved in electron bifurcation; stars denote the presence of biochemically characterized bifurcating Etfs. Sequences in G2D2 have been annotated as FixAB because of their proposed role in nitrogen fixation.

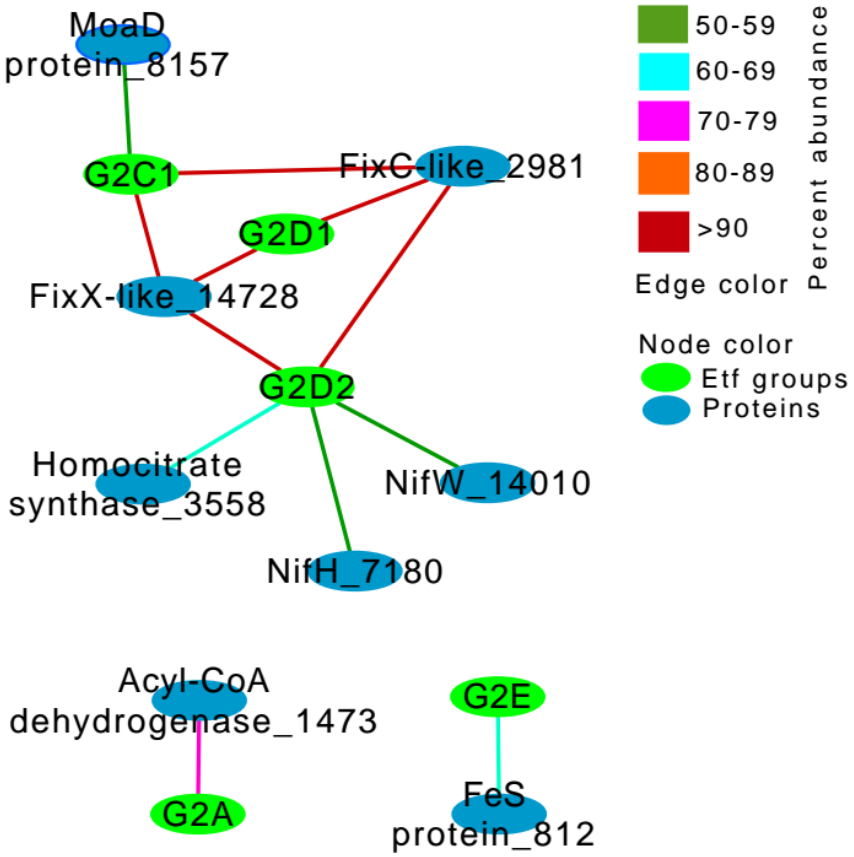


FIG 4 Network depiction of the covariation of proteins encoded in the flanking regions (±20 genes) of *etf* and specific group designations. Only proteins (n = 8) encoded by >50% of the G2 Etf-encoding genomes (i.e., relative frequency of >50%) were considered in this analysis. Here, nodes represent either the Etf group designation (green) or proteins in the flanking gene environment of Etfs (blue), while the edge color represents the abundance of the proteins in the group.

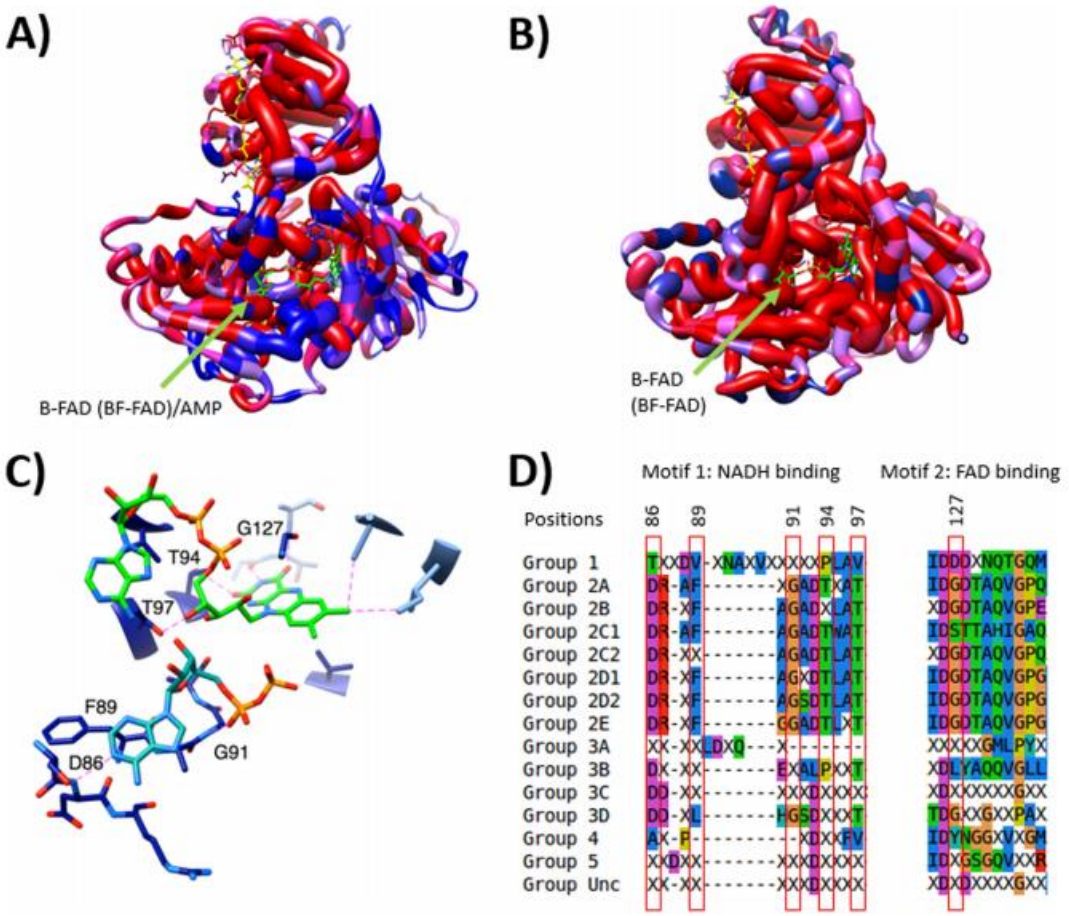


FIG 5 Structural analyses of Bf and non-Bf Etfs showing differences near the -FAD site. (A) Homology model of Etf-α and Etf-β with mapped conserved differences in amino acid type between G1 Etfs, with representatives that do not bifurcate and contain AMP in the subunit, and G2 Etfs, with representatives that have been demonstrated to catalyze electron-bifurcating reactions. Red indicates equivalent residues at a position, whereas blue indicates a position at which the distribution of residues characterizing one group differs from the distribution of residues found in the other group. The thickness of the worms indicate whether amino acids at a position share the same functional property (wide worm) or not (thin worm) within G2. (B) Homology model of Etf-α/β with mapped conserved differences in amino acid use between Etfs in G2A and G2B combined, which have representatives known to bifurcate, and Etfs in G2D2, which includes the FixAB Etfs. Note the high degree of conservation (thick red worms) around the putative -FAD binding site, leading us to predict that G2D2 Etfs also bifurcate. Thick worms indicate high levels of conservation within G2D2. (C) Closeup depiction of the NADH-binding (blue green) and FAD-binding (bright green) residues in bifurcating Etfs. Key residues proposed to coordinate these two molecules are numbered with their positions in the sequence of *R. palustris* Etf (GenBank accession no. YP_783418). Only a portion of the

NADH molecule, where the quality of the electron density was good enough to permit unambiguous refinement in the structure upon which our model is based (PDB accession no. 4L2I), is shown. Only residues associated with the proposed ability to bifurcate are shown. (D) Alignment of residues predicted to coordinate NADH and FAD in bifurcating Etfs. Groups in the first column correspond to the groups mentioned in Table 1. Unc indicates the uncharacterized group. Residues highlighted in panel C are shown in red boxes.

Table S1: Figures of merit of models generated based on group consensus sequences.[a]

Group	# sequences	Template	Clashes[b] (percentile)	Geometry[b] (percentile)	Ramachandran, favored[b]
Fix (2D2)	53	4kpu	88th %	85th %	93%
Bf (2A & 2B, non-basal[c])	118	4kpu	88th %	92 %	96%
Non-Bf (group 1)	449	1efv	99th %	98th%	96%
Group 2 (non-basal[c])	310	4kpu	86th%	90th %	96%
Group 2 (non-basal[c])•NAD	310	4L2I	80th%	86th%	95%

[a] FAD, AMP and NAD^+ moieties were grafted into the models from the related template files (listed below).

[b] Clashes, amino acid geometries, and Ramachandran allowed-ness were evaluated via the MolProbity Server (http://molprobity.biochem.duke.edu)

[c]Non-basal sequences represent more than 95% of all of the group sequences and exclude isolated sequences that cluster close to the branching point in each group. Omission of such sequences is performed to increase the signal-to-noise of the analyses that are directed within the group. We employed this precaution for those analyses whose objective was to learn what residues were strongly correlated within the group and to construct a consensus sequence for the group. This practice is based on the understanding that basal sequences are those that are least similar to the consensus sequence and carry the highest statistical risk of possessing residues that could randomly be those characteristic of another group (they also tend to have poor bootstrap support).

Consensus Amino Acid Sequences used to generate models

Group 2AB, Includes known Bifurcating Etfs

>EtfB-G2AB

MKIVVCIKQVPDTTEVKIDPVTGTLIRDGVPSIMNPDDKNALEEALRLKEEYGGK
VTVITMGPPQAKAALREALAMGADEAILLSDRAFAGADTLATSYTLAAAIKKLG
DYDLIICGRQAIDGDTAQVGPQIAEHLGIPQVTYVEKIEVVGDDSLTVKRALEDG
YEVIEVKTPCLLTVIKELNEPRYPSVKGIFEAYDKDKEIKVWSADDIEVDESKLGL
KGSPTQVKKSFTPEAKGAGEILEGLSPEEAADKLVEKLKEKHII

>EtfA-G2AB

MSIDKSDYKGVWVFAEQREGKIQPVSLELLGKGRELADKLGVEVTAVLLGSNV
KDLAKELIAYGADKVYVVDDPELKDYTTEPYTKAICDLINEYKPEIVLVGATTIG
RDLAPRVAARLRTGLTADCTSLDIDEETKLLLMTRPAFGGNIMATIICPNHRPQM
ATVRPGVMKKLERDESRKGEIIKVEVDLTESDIRTKVLEIVKKAKETVDIEEADII
VSGGRGVGSPENFELLEELADLLGGEVAASRAAVDAGWIDADHQVGQTGKTVR
PKLYIACGISGAIQHLAGMQDSDYIIAINKDPDAPIFKVADYGIVGDLYKVVPELIE
KIKANSLK

Group 2D2, Diazotrophs

>EtfB-G2D

MHIVVCIKQVPDSAQIRVHPVTNTIMRQGVPTIINPYDLFALEEALRLRDRFGGEV
TVLTMGPPMAEDALRKALSYGADRAVLLTDRAFAGSDTLATSYALAAAIRKIGE
EFPVDIVFTGKQTIDGDTAQVGPGIAKRLGLQQLTYVSKIVSIDLAAREITVERRA
EGGVQVLKTKLPCLITMLEGTNEIRRGSMDDALRAARAEIVKWSAADAGIEDVS
KCGLKGSPTVVKKVFAPTPRAEKAEMIETADKTPRDLAEALIAKIFTRQPKLEAE
LAFRAAA

>EtfA-G2D

MSTANKEPAAPAKGRAGMKKELPEHFKAYKHVWVFIELERGQVHPVSWELLGE
GRKLADKLGVELAGVVLGPPGEALEAAAAEAFAYGADLAYLVEDPVLADYRNE
PYTKALTDLVNTYKPEILLLGATTLGRDLAGSVATTLLTGLTADCTELDIDADRS
LAATRPTFGGSLLCTIYTLNYRPQMATVRPRVMAMPERDESRTGRIIEHKLGMVE
EDIVTKVLDFIPDRQSNKANLAYADVVVAGGLGLGNAENFQLVKDLARVLGAE
VGCSRPLVQKGWVPADRQIGQTGKTIRPKLYIAAGISGAIQHRVGVEGADLIVAI
NTDPNAPIFDFAHYGIVGDAIRLLPALTEAFRRRLSPHSRDRLAS

Group 1 non-basal (non-Bf)

>EtfB-G1nb

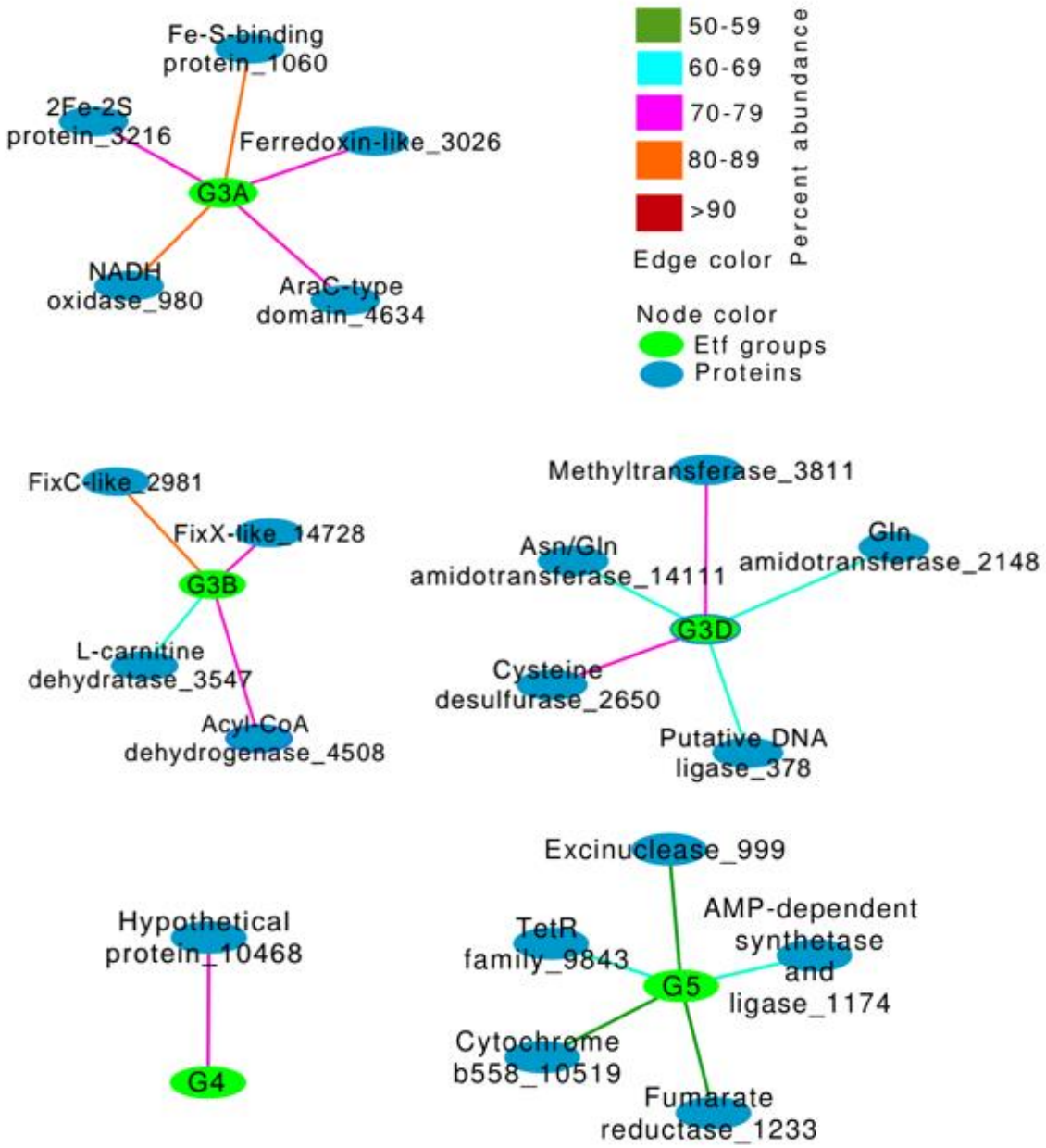
MKVLVPVKRVVDYNVKVRVKADGSGVDLANVKMSMNPFDEIAVEEAVRLKEA
GVATEVVAVSIGPAQAQETLRTALAMGADRAILVETDEELEPLAVAKLLKAVVD
KEQPQLVILGKQAIDDDSNQTGQMLAALLGWPQATFASKVEVADGKATVTREV
DGGLETLSLKLPAVVTTDLRLNEPRYASLPNIMKAKKKPLDTVTPADLGVDVAP
RLKTLKVEEPAKRSAGVKVADVAELVEKLKNEAKVI

>EtfA-G1nb
MTILVIAEHDNASLKAATLNTVTAAAKIGGDVHVLVAGSGAGAAAEAAAKIAG
VSKVLLADAAAYAHGLAENVAALVVSLAGDYSHILAPATATGKNVLPRVAALL
DVAQISDITAVVSADTFERPIYAGNAIATVQSSDAKKVITVRTTAFDAAAAEGGS
AAVEAVAAAADAGLSSFVGRELAKSDRPELTSAKIIVSGGRGLGSGENFTKVLEP
LADKLGAAVGASRAAVDAGYVPNDWQVGQTGKIVAPQLYIAVGISGAIQHLAG
MKDSKVIVAINKDEEAPIFQVADYGLVADLFEAVPELEKAL

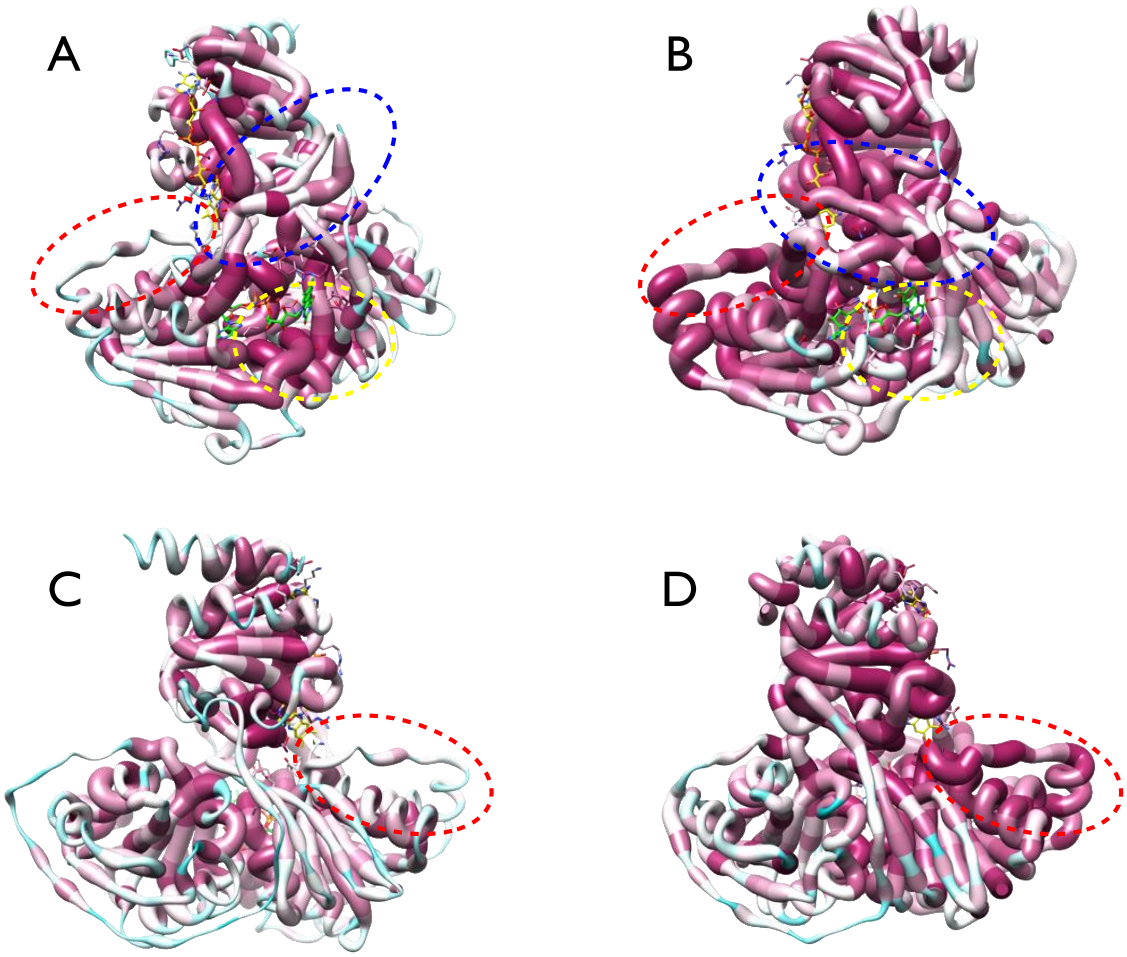
Group 2 non-basal
>EtfB-G2nb
MHIVVCIKQVPDTTQVRIDPVTGTLIREGVPSIINPYDLHALEEALRLKDKFGGKV
TVLTMGPPQAEEALREALAMGADEAILLSDRAFAGADTLATSYALAAAIRKIGEE
DVDLIFCGKQAIDGDTAQVGPGIAERLGIPQVTYVEKIEEVDLDKTITVKRRLEGG
YEVVEVKLPCLITVLKELNEPRYPSLPGKLRAARAEIKVWSAADLGDVDPSKIGL
KGSPTKVKKVFTPEARKEGEIIEGGDPEEADDDAAELLVEKLKEKPILEAKCKGC
GKCVKECPEDLAD

>EtfA-G2nb
YKGVWVFIEQEEGEVHPVSLELLGKGRKLADKLGVELAAVLLGSVVEDLAKELF
AYGADKVYVVDDPVLKDYRTEPYTKALTDLINKYKPEIILLGATTIGRDLAPRVA
TRLKTGLTADCTELDIDPETRLLLQTRPAFGGNIMATIVCPNHRPQMATVRPGVM
KKPERDEGRTGEIIEEEVDLTEEDILTKVLEVIKDREKEKVNLAEADIIVAGGRGL
GSKENFKLLEELADVLGGEVGASRAAVDAGWIPHDRQVGQTGKTVRPKLYIAC
GISGAIQHLVGMQDSDLIIAINKDPNAPIFDVADYGIVGDLFEVVPALTEALKKRL
AAKAK

Supplemental Table S2. Group designation and taxonomy of organisms that encode for *etfAB* sequences in their genomes identified in available genome sequences (see the excel file in the supplemental online documents)



Supplemental Figure S1. Network analysis of proteins encoded by flanking genes (+/- 20 genes) of *etf*. Only proteins (n=20) encoded by ≥50% of the Etf G3, G4 and G5 encoding genomes (i.e., relative frequency of ≥50%) were considered in this analysis. Here, a node represents either a group or a subgroup designation (denoted by green color) or an identified protein (denoted by blue color) within the gene neighborhood, while edge color represents the abundance of the proteins in the group.



Supplemental Figure S2. G2 is marked by greater amino acid conservation in the region of the Bf flavin and NAD binding, whereas G1 is distinguished by greater conservation in the recognition loop. Models of the consensus Etf of G2 (A and C) and G1 (B and D) are shown, based on consensus amino acid sequences determined excluding basal sequences in both cases (see Table S1). Panels A and B present a 'front' view whereas panels C and D present the 'back'. Amino acid identify conservation is color-coded as burgundy for positions at which a single identity is found, transitioning to teal for positions at which the maximum diversity of amino acid identity is found. Amino acid functional type conservation is depicted via worm widths with wide worms indicating positions where amino acids of a single functional type are retained and narrow worms indicating positions where diverse functional types are found, within each group. The FAD shared by both groups of Etf is in yellow sticks with heteroatoms in CPK colours. Green sticks depict the presumed Bf FAD of G2 Etfs, which has also been modeled into the G1 model

structure, where it can be accommodated without steric clashes with surrounding amino acids. The red dashed oval identifies the recognition loop (2) which is more conserved in G1 than in G2, suggesting different interaction modes with partner proteins or more diverse partner proteins in group 2. The blue dashed oval identifies the ‘thumb’ loop of EtfB (residues 12-30) that interacts with domain I (EtfA) and in group 1 drapes over the site in which the Bf flavin binds. This loop appears to occupy different locations in G1 vs. G2 Etfs, however this is based on only two different crystal structures for G2 and four from G1 Etfs. The yellow dashed oval identifies the region in which the Bf flavin and NAD bind in the Etf structure 4L2I.pdb of Chowdhury et al (3). High conservation is seen in the yellow oval in G2 but not G1.

REFERENCES

1. Chen, V. B., W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson. 2010. MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12-21.

2. Toogood, H. S., D. Leys, and N. S. Scrutton. 2007. Dynamics driving function: new insights from electron transferring flavoproteins and partner complexes. Febs j 274:5481-504.

3. Chowdhury, N. P., A. M. Mowafy, J. K. Demmer, V. Upadhyay, S. Koelzer, E. Jayamani, J. Kahnt, M. Hornung, U. Demmer, U. Ermler, and W. Buckel. 2014. Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J. Biol. Chem. 289:5145-57.

REFERENCES

1. Buckel W, Thauer RK. 2013. Energy conservation via electron bifurcatingferredoxin reduction and proton/Na^+ translocating ferredoxin oxidation. Biochim Biophys Acta Bioenerg 1827:94 –113. https://doi.org/10.1016/j.bbabio.2012.07.002.

2. Peters JW, Miller AF, Jones AK, King PW, Adams MW. 2016. Electron bifurcation. Curr Opin Chem Biol 31:146 –152. https://doi.org/10.1016/j.cbpa.2016.03.007.

3. Lubner CE, Jennings DP, Mulder DW, Schut GJ, Zadvornyy OA, Hoben JP, Tokmina-Lukaszewska M, Berry L, Nguyen DM, Lipscomb GL, Bothner B, Jones AK, Miller AF, King PW, Adams MWW, Peters JW. 2017. Mechanistic insights into energy conservation by flavin-based electron bifurcation. Nat Chem Biol 13:655– 659. https://doi.org/10.1038/nchembio.2348.

4. Herrmann G, Jayamani E, Mai G, Buckel W. 2008. Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. J Bacteriol 190:784 –791. https://doi.org/10.1128/JB.01422-07.

5. Li F, Hinderberger J, Seedorf H, Zhang J, Buckel W, Thauer RK. 2008. Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. J Bacteriol 190:843– 850. https://doi.org/10.1128/JB.01417-07.

6. Chowdhury NP, Mowafy AM, Demmer JK, Upadhyay V, Koelzer S, Jayamani

E, Kahnt J, Hornung M, Demmer U, Ermler U, Buckel W. 2014. Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J Biol Chem 289:5145–5157. https://doi.org/10.1074/jbc.M113.521013.

7. Demmer JK, Huang H, Wang S, Demmer U, Thauer RK, Ermler U. 2015. Insights into flavin-based electron bifurcation via the NADH-dependent reduced ferredoxin:NADP oxidoreductase structure. J Biol Chem 290:21985–21995. https://doi.org/10.1074/jbc.M115.656520.

8. Crane KL, Beinert H. 1956. On the mechanism of dehydrogenation of fatty acyl derivatives of coenzyme A. II. The electron-transferring flavoprotein. J Biol Chem 218:717–731.

9. Ghisla S, Thorpe C. 2004. Acyl-CoA dehydrogenases: a mechanistic overview.Eur J Biochem 271:494 –508. https://doi.org/10.1046/j.1432-1033.2003.03946.x.

10. Chen D, Swenson RP. 1994. Cloning, sequence analysis, and expression of the genes encoding the two subunits of the methylotrophic bacterium W3A1 electron transfer flavoprotein. J Biol Chem 269:32120 –32130.

11. Eichler K, Buchet A, Bourgis F, Kleber HP, Mandrand-Berthelot MA. 1995.The *fix Escherichia coli* region contains four genes related to carnitine metabolism. J Basic Microbiol 35:217–227. https://doi.org/10.1002/jobm.3620350404.

12. Earl CD, Ronson CW, Ausubel FM. 1987. Genetic and structural analysis of the *Rhizobium meliloti fixA, fixB, fixC*, and *fixX* genes. J Bacteriol 169:1127–1136. https://doi.org/10.1128/jb.169.3.1127-1136.1987.

13. Tsai MH, Saier MH, Jr. 1995. Phylogenetic characterization of the ubiquitous electron transfer flavoprotein families ETF-α and ETF-β. Res Microbiol 146:397– 404. https://doi.org/10.1016/0923-2508(96)80285-3.

14. Davidson VL, Husain M, Neher JW. 1986. Electron transfer flavoprotein from *Methylophilus methylotrophus*: properties, comparison with other electron transfer flavoproteins, and regulation of expression by carbon source. J Bacteriol 166:812– 817. https://doi.org/10.1128/jb.166.3.812-817.1986.

15. DuPlessis ER, Rohlfs RJ, Hille R, Thorpe C. 1994. Electron-transferring flavoprotein from pig and the methylotrophic bacterium W3A1 contains AMP as well as FAD. Biochem Mol Biol Int 32:195–199.

16. Walt A, Kahn ML. 2002. The *fixA* and *fixB* genes are necessary for anaerobic carnitine reduction in Escherichia coli. J Bacteriol 184:4044 – 4047. https://doi.org/10.1128/JB.184.14.4044-4047.2002.

17. Huang L, Rohlfs RJ, Hille R. 1995. The reaction of trimethylamine dehydrogenase with electron transferring flavoprotein. J Biol Chem 270:23958 –23965. https://doi.org/10.1074/jbc.270.41.23958.

18. Pühler A, Aguilar MO, Hynes M, Müller P, Klipp W, Priefer U, Simon R, Weber G. 1984. Advances in the genetics of free-living and symbiotic nitrogen fixing bacteria, p 609 – 619. In Veeger C, Newton WE (ed), Advances in nitrogen fixation research: proceedings of the 5th International Symposium on Nitrogen Fixation, Noordwijkerhout, The Netherlands, August 28 –September 3, 1983. Springer Netherlands, Dordrecht, Netherlands.

19. Ruvkun GB, Sundaresan V, Ausubel FM. 1982. Directed transposon Tn5 mutagenesis and complementation analysis of *Rhizobium meliloti* symbiotic nitrogen fixation genes. Cell 29:551–559. https://doi.org/10.1016/0092-8674(82)90171-4.

20. Huang JJ, Heiniger EK, McKinlay JB, Harwood CS. 2010. Production of hydrogen gas from light and the inorganic electron donor thiosulfate by *Rhodopseudomonas palustris*. Appl Environ Microbiol 76:7717–7722. https://doi.org/10.1128/AEM.01143-10.

21. Edgren T, Nordlund S. 2004. The *fixABCX* genes in *Rhodospirillum rubrum* encode a putative membrane complex participating in electron transfer to nitrogenase. J Bacteriol 186:2052–2060. https://doi.org/10.1128/JB.186.7.2052-2060.2004.

22. Ledbetter RN, Garcia Costas AM, Lubner CE, Mulder DW, Tokmina Lukaszewska M, Artz JH, Patterson A, Magnuson TS, Jay ZJ, Duan HD,Miller J, Plunkett MH, Hoben JP, Barney BM, Carlson RP, Miller A-F,Bothner B, King PW, Peters JW, Seefeldt LC. 2017. The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: generation of low-potential reducing equivalents for nitrogenase catalysis. Biochemistry https://doi.org/10.1021/acs.biochem.7b00389.

23. Roberts DL, Frerman FE, Kim JJ. 1996. Three-dimensional structure of human electron transfer flavoprotein to 2.1-Å resolution. Proc Natl Acad Sci U S A 93:14355–14360. https://doi.org/10.1073/pnas.93.25.14355.

24. Roberts DL, Salazar D, Fulmer JP, Frerman FE, Kim JJ. 1999. Crystal structure of *Paracoccus denitrificans* electron transfer flavoprotein: structural and electrostatic analysis of a conserved flavin binding domain. Biochemistry 38:1977–1989. https://doi.org/10.1021/bi9820917.

25. Sato K, Nishina Y, Shiga K. 2003. Purification of electron-transferring flavoprotein from *Megasphaera elsdenii* and binding of additional FAD with an unusual absorption spectrum. J Biochem 134:719 –729. https://doi.org/10.1093/jb/mvg199.

26. Chowdhury NP, Kahnt J, Buckel W. 2015. Reduction of ferredoxin or oxygen by flavin-based electron bifurcation in *Megasphaera elsdenii*. FEBS J 282:3149 –3160. https://doi.org/10.1111/febs.13308.

27. Weghoff MC, Bertsch J, Muller V. 2015. A novel mode of lactate metabolism in strictly anaerobic bacteria. Environ Microbiol 17:670 – 677. https://doi.org/10.1111/1462-2920.12493.

28. Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW. 2011. A latemethanogen origin for molybdenum-dependent nitrogenase. Geobiology 9:221–232. https://doi.org/10.1111/j.1472-4669.2011.00278.x.

29. Posewitz MC, King PW, Smolinski SL, Zhang L, Seibert M, Ghirardi ML. 2004. Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. J Biol Chem 279:25711–25720. https://doi.org/10.1074/jbc.M403206200.

30. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13:162. https://doi.org/10.1186/1471-2164-13-162.

31. Estelmann S, Boll M. 2014. Glutaryl-coenzyme A dehydrogenase from *Geobacter metallireducens*: interaction with electron transferring flavoprotein and kinetic basis of unidirectional catalysis. FEBS J 281: 5120 –5131. https://doi.org/10.1111/febs.13051.

32. Zhang J, Frerman FE, Kim JJ. 2006. Structure of electron transfer flavoprotein-ubiquinone oxidoreductase and electron transfer to the mitochondrial ubiquinone pool. Proc Natl Acad SciUSA 103:16212–16217. https://doi.org/10.1073/pnas.0604567103.

33. Gubler M, Zürcher T, Hennecke H. 1989. The *Bradyrhizobium japonicum fixBCX* operon: identification of fixX and of a 5'mRNA region affecting the level of the fixSCX transcript. Mol Microbiol 3:141–148. https://doi.org/10.1111/j.1365-958.1989.tb01803.x.

34. Hochstrasser M. 2009. Origin and function of ubiquitin-like protein conjugation. Nature 458:422– 429. https://doi.org/10.1038/nature07958.

35. Makarova KS, Koonin EV. 2010. Archaeal ubiquitin-like proteins: functional versatility and putative ancestral involvement in tRNA modification revealed by comparative genomic analysis. Archaea 2010:710303. https://doi.org/10.1155/2010/710303.

36. Tan X, Volbeda A, Fontecilla-Camps JC, Lindahl PA. 2006. Function of the tunnel in acetylcoenzyme A synthase/carbon monoxide dehydrogenase. J Biol Inorg Chem 11:371–378. https://doi.org/10.1007/s00775-006-0086-9.

37. Grönger P, Manian SS, Reiländer H, O'Connell M, Priefer UB, Pühler A. 1987. Organization and partial sequence of a DNA region of the *Rhizobium leguminosarum* symbiotic plasmid pRL6JI containing the genes *fixABC*, *nifA*, *nifB* and a novel open reading frame. Nucleic Acids Res15:31– 49. https://doi.org/10.1093/nar/15.1.31.

38. Kaminski AP, Norel F, Desnoues N, Kush A, Salzano G, Elmerich C. 1988. Characterization of the fixABC region of *Azorhizobium caulinodans* ORS571 and identification of a new nitrogen fixation gene. Mol Gen Genet 214:496 –502. https://doi.org/10.1007/BF00330486.

39. Lee PT, Hsu AY, Ha HT, Clarke CF. 1997. A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the *Escherichia coli ubiE* gene. J Bacteriol 179:1748 –1754. https://doi.org/10.1128/jb.179.5.1748-1754.1997.

40. Jang M-H, Scrutton NS, Hille R. 2000. Formation of W3A1 electron transferring flavoprotein (ETF) hydroquinone in the trimethylamine dehydrogenase·ETF protein complex. J Biol Chem 275:12546 –12552.https://doi.org/10.1074/jbc.275.17.12546.

41. Toogood HS, Leys D, Scrutton NS. 2007. Dynamics driving function: new insights from electron transferring flavoproteins and partner complexes. FEBS J 274:5481–5504. https://doi.org/10.1111/j.1742-4658.2007.06107.x.

42. Scott JD, Ludwig RA. 2004. *Azorhizobium caulinodans* electron transferring flavoprotein N electrochemically couples pyruvate dehydrogenase complex activity to N_2 fixation. Microbiology 150:117–126. https://doi.org/10.1099/mic.0.26603-0.

43. Luo H, Lin Y, Gao F, Zhang C-T, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 42:D574 –D580. https://doi.org/10.1093/nar/gkt1131.

44. Pechter KB, Gallagher L, Pyles H, Manoil CS, Harwood CS. 2015. Essential genome of the metabolically versatile alphaproteobacterium *Rhodopseudomonas palustris*. J Bacteriol 198:867– 876. https://doi.org/10.1128/JB.00771-15.

45. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122. https://doi.org/10.1093/nar/gkr1044.

46. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.

47. Silvestro D, Michalak I. 2012. raxmlGUI: a graphical front-end for RAxML. Org Divers Evol 12:335–337. https://doi.org/10.1007/s13127-011-0056-0.

48. Wimley WC, Creamer TP, White SH. 1996. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry 35:5109 –5124. https://doi.org/10.1021/bi9600153.

49. Gray HB, Winkler JR. 2015. Hole hopping through tyrosine/tryptophan chains protects proteins from oxidative damage. Proc Natl Acad SciUSA 112:10920 –10925. https://doi.org/10.1073/pnas.1512704112.

50. Pradeep L, Shin HC, Scheraga HA. 2006. Correlation of folding kinetics with the number and isomerization states of prolines in three homologous proteins of the RNase family. FEBS Lett 580:5029 –5032. https://doi.org/10.1016/j.febslet.2006.08.024.

51. Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195–201. https://doi.org/10.1093/bioinformatics/bti770.

52. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T. 2014. SWISSMODEL:modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258. https://doi.org/ 10.1093/nar/gku340.

53. Benkert P, Biasini M, Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343–350. https://doi.org/10.1093/bioinformatics/btq662.

54. Benkert P, Tosatto SC, Schomburg D. 2008. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277. https://doi.org/10.1002/prot.21715.

55. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera: a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084.

56. Wang J, Wang W, Kollman PA, Case DA. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.

57. Watmough NJ, Kiss J, Frerman FE. 1992. Structural and redox relationships between *Paracoccus denitrificans*, porcine and human electrontransferring flavoproteins. Eur J Biochem 205:1089 –1097. https://doi.org/10.1111/j.1432-1033.1992.tb16877.x.

58. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431– 432.

CHAPTER FOUR

ELECTRON TRANSFER TO NITROGENASE IN DIFFERENT GENOMIC AND METABOLIC BACKGROUNS

Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Saroj Poudel

Contributions: Designed and conducted the experiment. Contributed to the writing of the manuscript.

Author: Daniel R. Colman

Contributions: Contributed to the writing of the manuscript.

Co-Author: Kathryn R. Fixen

Contributions: Provided field expertise and feedback on drafts of the manuscript.

Co-Author: Rhesa N. Ledbetter

Contributions: Contributed to the writing of the manuscript.

Co-Author: Yanning Zheng

Contributions: Provided field expertise and feedback on drafts of the manuscript.

Co-Author: Natasha Pence

Contributions: Contributed to the writing of the manuscript.

Co-Author: Lance C. Seefeldt

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: John W. Peters

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Eric S. Boyd

Contributions: Principal investigator and corresponding author. Provided feedback on the study design. Provided statistical advice and contributed to the writing of the manuscript.

Manuscript Information Page

Saroj Poudel, Daniel R. Colman, Kathryn R. Fixen, Rhesa N. Ledbetter, Yanning Zheng, Natasha Pence, Lance C. Seefeldt, John W. Peters, Caroline S. Harwood, Eric S. Boyd, and
Journal of Bacteriology

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-review journal
____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published online February, 2018, doi: 10.1128/JB.00757-17.

Electron transfer to nitrogenase in different genomic and metabolic backgrounds

Saroj Poudel[1], Daniel R. Colman[1], Kathryn R. Fixen[2], Rhesa N. Ledbetter[3], Yanning Zheng[4], Natasha Pence[5], Lance C. Seefeldt[3], John W. Peters[5], Caroline S. Harwood[3], Eric S. Boyd[1#]

[1]Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717
[2]Department of Plant and Microbial Biology, University of Minnesota, St. Paul. Minnesota
[3]Department of Chemistry and Biochemistry, Utah State University, Logan, UT 84322
[4]Department of Microbiology, University of Washington, Seattle, WA 98195
[5]Institute of Biological Chemistry, Washington State University, Pullman, WA 99164

Running Head (limit 54 characters and spaces): Electron delivery to nitrogenase

[#]Author of correspondence: Eric S. Boyd (eboyd@montana.edu)
Department of Microbiology & Immunology
109 Lewis Hall
Montana State University
Bozeman, MT 59717
Phone: (406) 994-7046
Fax: (406) 994-4926

Keywords.

Nitrogenase, nitrogen fixation, ferredoxin, flavodoxin, Fix, Rnf, bifurcation

ABSTRACT

Nitrogenase catalyzes the reduction of dinitrogen (N_2) using low-potential electrons from ferredoxin (Fd) or flavodoxin (Fld) through an ATP-dependent process. Since its emergence in an anaerobic chemoautotroph, this oxygen (O_2)-sensitive enzyme complex has evolved to operate in a variety of genomic and metabolic backgrounds, including those of aerobes, anaerobes, chemotrophs, and phototrophs. However, whether pathways of electron delivery to nitrogenase are influenced by these different metabolic backgrounds is not well understood. Here, we report the distribution of homologs of Fds, Flds, and Fd-/Fld-reducing enzymes in 359 genomes of putative N_2 fixers (diazotrophs). Six distinct lineages of nitrogenase were identified, and their distributions largely corresponded to differences in the host cells' ability to integrate O_2 or light into energy metabolism. The predicted pathways of electron transfer to nitrogenase in aerobes, facultative anaerobes, and phototrophs varied from those in anaerobes at the levels of Fds/Flds used to reduce nitrogenase, the enzymes that generate reduced Fds/Flds, and the putative substrates of these enzymes. Proteins that putatively reduce Fd with hydrogen or pyruvate were enriched in anaerobes, while those that reduce Fd with NADH/NADPH were enriched in aerobes, facultative anaerobes, and anoxygenic phototrophs. The energy metabolism of aerobic, facultatively anaerobic, and anoxygenic phototrophic diazotrophs often yields reduced NADH/NADPH that is not sufficiently reduced to drive N_2 reduction. At least two mechanisms have been acquired by these taxa to overcome this limitation and to generate electrons with potentials capable of reducing Fd. These

include the bifurcation of electrons or the coupling of Fd reduction to reverse ion translocation.

IMPORTANCE

Nitrogen fixation supplies fixed nitrogen to cells from a variety of genomic and metabolic backgrounds, including those of aerobes, facultative anaerobes, chemotrophs, and phototrophs. Here, using informatics approaches applied to genomic data, we show that pathways of electron transfer to nitrogenase in metabolically diverse diazotrophic taxa have diversified primarily in response to host cells' acquired ability to integrate O_2 or light into their energy metabolism. The acquisition of two key enzyme complexes enabled aerobic and facultatively anaerobic phototrophic taxa to generate electrons of sufficiently low potential to reduce nitrogenase: the bifurcation of electrons via the Fix complex or the coupling of Fd reduction to reverse ion translocation via the *Rhodobacter* nitrogen fixation (Rnf) complex.

KEYWORDS: fix, nitrogenase, Rnf, ferredoxin, flavodoxin, hydrogen, nitrogen fixation, oxygen, photosynthesis, pyruvate, bifurcation

Nitrogenase is an oxygen-sensitive enzyme that catalyzes the reduction of dinitrogen (N_2) to ammonia (NH_3), accounting for nearly two-thirds of fixed nitrogen (N) on earth today (1,2) and thereby modulating the global fixed N supply (3,4). Nitrogenase is comprised of two components, a homodimeric iron protein (H subunit) and a heterotetrameric dinitrogenase reductase protein complex (D and K subunits) (5,6). Homologs of nitrogenase are widely distributed among Bacteria but have only been identified in one group of taxa within the archaeal *Euryarchaeota* phylum, methanogens. Nitrogenase homologs have not been identified among eukarya (1, 7–12). Three types of nitrogenase have been described that are evolutionarily and structurally related but differ in the metallic composition of their active site: molybdenum (Mo)-containing nitrogenases (Nif), vanadium (V)-containing nitrogenases (Vnf), and iron (Fe)-containing nitrogenases (Anf) (13–16).

Phylogenetic evidence indicates that Nif emerged prior to Anf and Vnf and that the earliest-evolving Nif lineages are from anaerobic, hyperthermophilic methanogens, implicating a chemotrophic origin for N_2 fixation in an anoxic environment (7, 9, 13, 16, 17). N_2-fixing microorganisms (i.e., diazotrophs) have since diversified to function in a wide variety of genomic and metabolic backgrounds, including those of aerobes, facultative anaerobes, and phototrophs. Given the O_2 sensitivity of Nif, aerobic, facultatively anaerobic, and oxygenic phototrophic diazotrophs have had to evolve or acquire one of several mechanisms to mitigate the toxicity of O_2 to this enzyme during N_2 fixation (17). These mechanisms include fixing N_2 only during dark hours when oxygenic photosynthesis ceases and when heterotrophic respiration keeps O_2 tensions

low (18), using specialized cells called heterocysts in filamentous *Cyanobacteria* to spatially localize nitrogenase in anaerobic cellular compartments (19), maintaining an anoxic environment in the cytoplasm through increased O_2-dependent respiratory activity (20), or fixing N_2 only during anaerobic growth. A recent study showed that the diversification of Nif in aerobes and facultative anaerobes was also accompanied by the recruitment and loss of genes that are involved in regulating and protecting Nif against oxidative stress (17).

The primary electron donors to Nif are reduced ferredoxin (Fd) (21–24) and flavodoxin (Fld) (25–27). The transfer of eight electrons from Fd or Fld to NifH and ultimately to NifDK is an ATP-dependent process, requiring at a minimum 16 mol ATP per mol N_2 reduced (28–30). Less is known of the stoichiometry of ATP hydrolysis per mol N_2 reduced by Anf or Vnf (31–34). The reduction of Fd or Fld in diazotrophs that occupy anoxic environments (e.g., clostridia and methanogens) occurs via the activity of pyruvate-flavodoxin oxidoreductase (PFOR) (35–38), a select subset (groups 3c, 3d, 4d, and 4e) of [NiFe]-hydrogenases, or [FeFe]-hydrogenases (39–45). However, the reduction of Fd (standard electrode potential [E_o'] ~ -420 mV) in aerobic and some anoxygenic phototrophic diazotrophs that inhabit less-reducing environments is more of a challenge. Rather than producing reduced Fd during their primary energy metabolism, aerobic bacteria and some anoxygenic phototrophs, in particular facultatively anaerobic purple sulfur and nonsulfur bacteria, generate reduced NADH or NADPH (E_o' = -320 mV) (46, 47), which are not of low enough potential to drive N_2 reduction (48–50). Anaerobic purple sulfur and facultatively anaerobic nonsulfur

anoxygenic phototrophic bacteria utilize photosystems to drive cyclic electron transfer. The reduction of $NAD^+/NADP^+$ is typically accomplished with electrons supplied by the oxidation of an inorganic substrate (e.g., sulfide, thiosulfate, or hydrogen [H_2]) or by the oxidation of organic compounds and energy from reverse electron transport if the inorganic electron donor is not of low-enough potential to reduce $NAD^+/NADP^+$ (51,52). It has been proposed that these taxa acquired the Fix complex (encoded by *fixABCX*) and/or the *Rhodobacter* nitrogen fixation [Rnf, encoded by rnfABCDEG (H)] complex in order to generate reduced Fd from NADH/NADPH (46). Fix catalyzes the oxidation of two NADH to generate a reduced quinone, which fuels the respiratory chain, and a reduced Fd (27, 53, 54), whereas Rnf catalyzes the NADH-dependent reduction of Fd by coupling it to the depletion of the electrochemical gradient (55–58).

Oxygenic phototrophic *Cyanobacteria* utilize photosystem I to energize electrons to potentials negative enough to drive the reduction of Fd (51, 52). However, this Fd is not available to Nif, since it must be temporally or spatially separated from oxygenic photosynthesis due to inhibition of Nif by O_+. Rather than coding for [FeFe]-hydrogenase, [NiFe]-hydrogenase, Fix, or Rnf and using these enzymes to reduce Fd, *Cyanobacteria* encode ferredoxin-$NADP^+$ oxidoreductase (FNR), which can function in reverse to reduce Fd or Fld with NADPH generated by carbohydrate oxidation in heterocysts or when O_2 tensions are low (35, 59–62). Some *Cyanobacteria* also encode PFOR, which might be expected to contribute to Fd reduction in these cells, and this might be used by Nif. Like *Cyanobacteria* anaerobic anoxygenic green sulfur bacteria use a type I photosystem that is distantly related to photosystem I to generate reduced Fd

as a component of photosystem-driven cyclic electron transfer (51, 52); however, unlike *Cyanobacteria* it is possible that this serves as a reductant for N_2 fixation (63). Green sulfur bacteria also encode an FNR, which is phylogenetically and structurally unrelated to conventional FNR found in *Cyanobacteria* and this can also be used to drive reduction of Fd (64, 65). Thus, at least seven enzyme complexes have evolved to provide reduced Fd for N_2 fixation: PFOR, [NiFe]-hydrogenase, [FeFe]-hydrogenase, Rnf, Fix, and both forms of FNR (35); however, their distribution in the genomes of diazotrophs is not known.

Different Fds and Flds are also likely to be involved in the delivery of electrons to nitrogenase in aerobes, anaerobes, and phototrophs, and these Fds and Flds may vary alongside the primary enzymes that are involved in reducing these electron carriers. Fds are sensitive to O_2 due to the lability of their iron sulfur (FeS) clusters (66–69). In contrast, Flds contain flavin mononucleotide as the prosthetic group involved in electron transfers instead of FeS clusters and, hence, are thought to be less sensitive to O_2 (67, 70, 71). Previous bioinformatics analyses have shown that NifF (17), an Fld that can donate electrons to Nif (24, 72, 73), was recruited to nif operons during the diversification of Nif from anaerobic to aerobic taxa (17), which may point to the use of Flds as an adaptive strategy to fix N_2 in oxic environments. Moreover, under the iron-deficient conditions that characterize most circumneutral oxic environments, diazotrophs tend to synthesize Flds preferentially as primary electron donors to Nif (22–24, 26, 74). Studies have also shown that electron delivery by an Fd or Fld can be complemented by other Fds or Flds that are encoded in the genomes of diazotrophs (24, 35, 74–76). These observations

suggest that pathways that mediate electron flow to nitrogenase are flexible and vary according to the genomic and metabolic background of taxa.

To better define the systems of electron transfer to nitrogenase in diverse microbes, we compiled all Fd and Fld homologs in Nif-encoding genomes and classified them using homology-based methods. In addition, we compiled homologs of all enzymes that have been shown to reduce Fd or Fld for use in the reduction of N_2 by nitrogenase. These include PFOR, [NiFe]-hydrogenase, [FeFe]-hydrogenase, Fix, Rnf, and both forms of FNR. Since Nif is encoded in all genomes that encode Anf and Vnf (1), we also examined the systems of electron transfer to these alternative nitrogenase isoforms. Organisms encoding Nif, Anf, and Vnf were classified phylogenetically and physiologically as aerobes, anaerobes, or facultative anaerobes and as chemotrophs, anoxygenic phototrophs, or oxygenic phototrophs based on published data. Statistical analyses were then applied to this curated data set to identify patterns of cooccurrence between the distribution of nitrogenase lineages/isoforms, enzymes that putatively reduce Fd/Fld, and Fds/Flds. The results are discussed in the context of the metabolism of the cells, specifically the influence of O_2 and light on putative pathways of electron delivery to nitrogenase.

RESULTS AND DISCUSSION

Taxonomic distribution and phylogeny of Nif/Anf/Vnf HDK homologs. Nitrogenases were identified in the genomes of diverse *Bacteria* and *Archaea* that included obligate aerobes, facultative anaerobes, obligate anaerobes, phototrophs and chemotrophs, and

autotrophs and heterotrophs. The identification of Nif in organisms with diverse metabolisms is consistent with a complex evolutionary history that has been described previously for N_2 fixation (see Table S1A in the supplemental material) (1, 7, 8, 13, 16). Of the total 4,588 publicly available genomes in our database, 359 genomes (7.8% of the total) encoded the minimal set of proteins for nitrogen fixation (i.e., homologs of NifHDKENB) (7). Forty-six of the 359 taxa with genomes that encode nitrogenase complements have been experimentally shown to grow with atmospheric N_2 as their sole N source (Table S1B). Of these 359 nitrogenase-encoding genomes, 48 belonged to obligately chemotrophic and anaerobic *Archaea*, all of which were affiliated with methanogens within the phylum *Euryarchaeota*. The remaining 311 diazotrophic genomes were identified in the bacterial domain, with the majority identified as members of the *Proteobacteria* (n = 191) and Firmicutes (n = 68) (Table S1A). Of the 311 diazotrophic bacterial genomes, 31% were from aerobes, 40% were from facultative anaerobes, and 28% were from obligate anaerobes. Furthermore, 79% of the diazotrophic bacterial genomes were from chemotrophs and 21% were from phototrophs (both oxygenic and anoxygenic).

Phylogenetic reconstruction of a concatenation of HDK proteins revealed six distinct lineages of nitrogenases (Fig. 1). These included the four lineages that have been identified in previous studies (1, 8, 13, 14, 16, 77), which include two Nif sublineages (designated Nif-A and Nif-B) and two sublineages designated Anf and Vnf. In addition, previous combined informatics and structural analyses have suggested that two biochemically uncharacterized lineages likely harbor a molybdenum cofactor (9), and

organisms that encode them have experimentally been shown to fix N_2 (11, 78); hence, we refer to these lineages as Nif-C and Nif-D in this study.

A total of 224 genomes encoded Nif-A, and these were from taxa that are primarily from the *Proteobacteria* (n = 169), *Firmicutes* (n = 26), and *Cyanobacteria* (n = 20) (Table S1A). Thirty of these taxa have been experimentally shown to fix N_2 (Table S1B). However, the cyanobacterium *Microcoleus chthonoplastes* (which is not included in our database since a complete genome is not available) does encode Nif-A and the minimal set of proteins that allow for N_2 fixation (i.e., homologs of NifHDKENB) (7). And yet, cultivation studies suggest that this taxon cannot grow with N_2 as its sole N source (79), indicating that the presence of a full nif gene complement does not guarantee the ability to fix N_2. The majority of the taxa that encode Nif-A are aerobes (42%) or facultative anaerobes (38% of total taxa) (Fig. 2A). In addition, 23% of Nif-A-encoding genomes were from phototrophs and 77% were from chemotrophs (Fig. S1). Among the 23% of Nif-A homologs that are encoded in the genomes of phototrophs, 9% are from *Cyanobacteria*, 11% are from facultatively anaerobic anoxygenic purple nonsulfur bacteria, and 2% are from anaerobic anoxygenic purple bacteria. However, it is not clear how rigorously O_2 usage has been characterized in these anaerobic anoxygenic purple bacteria, which include the following strains: *Allochromatium vinosum*, *Thiocystis violascens*, *Thioflavicoccus mobilis*, and *Halorhodospira halophila*.

The majority of the 106 genomes that encode Nif-B were from obligate anaerobes (93% of total taxa) and chemotrophs (87% of the total taxa) and anaerobic anoxygenic green phototrophic bacteria (11% of total). Nif-B-encoding taxa were primarily affiliated

with *Firmicutes* (n = 34), *Euryarchaeota* (n = 29), and *Proteobacteria* (n = 22). Of these 106 Nif-B-encoding genomes, 13 were from taxa (primarily from the genus Clostridium) that have been experimentally shown to fix N_2 (Table S1B). All of the anaerobic anoxygenic phototrophs were affiliated with the *Chlorobi* or *Chloroflexi* (n = 12). Members of the *Firmicutes* (n = 8) and *Euryarchaeota* (n = 6) encoded Nif-C (n = 15), while all (n = 13) Nif-D-encoding genomes were methanogens (*Euryarchaeota*). None of the taxa that encode Nif-C have been shown to fix N_2, while three organisms that encode Nif-D (all of which are methanogens) have been experimentally shown to fix N_2 (Table S1B). Taxa that encode Nif-C were primarily anaerobes (88% of the total taxa), while taxa that encode Nif-D were all anaerobes. All the Nif-C- and Nif-D-encoding genomes were chemotrophs.

A total of 32 genomes encoded Anf, and these were primarily from the *Proteobacteria* (n = 16) and *Firmicutes* (n = 11), while a total of 23 genomes encoded Vnf. Most of the Vnf-encoding genomes were from the *Euryarchaeota* (n = 11) and *Firmicutes* (n = 8). A separate lineage comprising HDK proteins from taxa that have not yet been shown to fix N_2 were nested among Nif sublineages. These proteins were termed "uncharacterized," as previously described (13), and were affiliated with members of the *Chloroflexi* (i.e., *Roseiflexus* spp.).

Multiple forms of nitrogenase were often detected in the same genome (Fig. 2B). However, the genomes that coded for Nif-C or Nif-D did not code for any other forms of nitrogenases. The genome of the cyanobacterium *Pleurocapsa* sp. strain PCC 7327 was found to encode both Nif-A and Nif-B. All Anf- and Vnf-encoding genomes also

encoded Nif, which is consistent with previous observations (1, 7, 13). Four genomes were identified that coded for Anf, Vnf, and Nif-B, while two genomes encoded Anf, Vnf, and Nif-A. Interestingly, Anf was more commonly detected in Nif-A-encoding genomes (65% of total Anf-encoding genomes), while Vnf was more commonly detected in Nif-B-encoding genomes (77% of total Vnf-encoding genomes). If the root of the nitrogenase phylogeny is within the Nif-D lineage, as has been suggested previously (7, 13, 16), our phylogenetic reconstruction indicates that Anf diverged from a Vnf-like ancestor, both of which emerged from a Nif-C- or Nif-D-like ancestor. Likewise, our phylogenetic reconstruction indicates that Nif-A diversified from a Nif-B-like ancestor.

Taxonomic distribution of ferredoxin and flavodoxin homologs in the genomes of putative diazotrophs. (i) Ferredoxin. The taxonomic distribution of all 47 Fds identified in the genomes of putative diazotrophs is detailed in Table S2 in the supplemental material. A total of 36 Fds were detected in Nif-A-encoding genomes, of which FdxB and FdxA were the most common; they were present in >40% of those genomes (Fig. S2A). *fdxB* is encoded near *nifQ* and *nifB* in the genome of *Azotobacter vinelandii* (80, 81) and is expressed at a similar level to these genes under N_2-fixing conditions (82). Mutational studies have shown involvement of FdxB in active-site metallocluster biosynthesis in *A. vinelandii* (83, 84). Furthermore, FdxB was shown to be incapable of serving as an electron donor to nitrogenase in vitro using protein from *Rhodobacter capsulatus* (83). In contrast, in the cyanobacterium *Anabaena* sp. strain PCC 7120, FdxB has been shown to complement *fdxH* mutant strains; FdxH is the preferred donor of electrons to nitrogenase in this taxon (85). These observations suggest

that FdxB has multiple roles in diazotrophic cells. The prevalence of FdxA in Nif-A-encoding genomes is consistent with its colocalization near major nif gene clusters, such as in the case of *A. vinelandii* (24), or within the nif gene cluster itself, such as in *Herbaspirillum seropedicae* (86). This agrees with findings from multiple studies that have documented the ability of FdxA to donate electrons to nitrogenase (87–90). Several less commonly detected Fds (identified in <40% of the Nif-A-encoding genomes) have also been shown or predicted to be involved in electron transfer to nitrogenase. These include FdxN in *Rhodospirillum rubrum* and *Rhizobium meliloti* (91, 92), FdxE in *Rhodobacter capsulatus* (93), FdxH (85, 94), FdxI in *Anabaena* sp. PCC 7120 (95, 96), and FdxC in *R. capsulatus* (Fig. S2A) (97). Together, these results suggest that FdxA and FdxB may have a role in electron delivery to Nif-A-like nitrogenases; however, other Fds could potentially complement the functionality of these Fds.

Two Fds found in *Clostridium pasteurianum*, CpFd1 (65% of the total Nif-B-encoding genomes) and CpFd4 (48% of the total genomes), correlated with the distribution of Nif-B in diazotrophic genomes (Fig. S2A), suggesting a role for these Fds in electron delivery to Nif-B-like nitrogenases. FdxA (47% of the total Nif-C-encoding genomes) and three Fds found in *Caldicellulosiruptor saccharolyticus*, CsFd3 (47% of the total genomes), CsFd6 (47% of the total genomes), and CsFd1 (41% of the total genomes), were predominant in Nif-C-encoding genomes, while Nif-D-encoding genomes encoded multiple Fds found in *Methanocaldococcus vulcanius* (MvFds) (Fig. S2A). The distribution of Anf in genomes was only moderately correlated with the distribution of Fds, with FdxA and FdxB yielding the highest correlations (i.e., >45% of

the genomes) (Fig. S2A). Finally, the distribution of Vnf was highly correlated with the distribution of CpFd4 (80% of the total genomes) and, to lesser extents, four Fds found in *Methanosarcina barkeri*, MbFd1, MbFd2, MbFd3, and MbFd4 (Fig. S2A). While electron transfer to Nif-A-like nitrogenases has been studied extensively, we are unaware of experimental data on the role of Fds in electron transfer to Nif-B, Nif-C, Nif-D, Anf, or Vnf, which precludes a comparison of our informatics-based predictions to experimental data.

The Fds identified in our study varied markedly among genomes coding for the various isoforms of nitrogenase (Fig. S2A). Importantly, genomes encoding Nif-A nitrogenases coded for unique Fds that were either absent or present but rarely detected in the genomes of other diazotrophs (Fig. S2A). For example, of 14 abundant Fds (i.e., present in >40% of the genomes) encoded by Nif-A-encoding genomes, 12 were unique to Nif-A-encoding genomes, while only two were identified in the genomes of other diazotrophs (Fig. S2A). In contrast, considerable overlap was observed in the composition of Fds/Flds encoded in genomes that also encoded Nif-B, Nif-C, or Nif-D.

(ii) Flavodoxin. We identified a total of five phylogenetically distinct Flds among nitrogenase-encoding genomes. The taxonomic distribution of all five Flds identified is detailed in Table S2 in the supplemental material. Of the five Flds detected in Nif-Aencoding genomes, FldA was the most common and was identified among 26% of the genomes (Fig. S2B). The distribution of CpFld2 and CpFld3 cooccurred with Nif-B in >40% of the genomes, while CpFld3 was detected in >40% of the Nif-C- and Nif-D encoding genomes (Fig. S2B). Like Nif-A-encoding genomes, FldA was the dominant

Fld in Anf-encoding genomes (Fig. S2B). Lastly, CpFld3 and CpFld2 were the dominant Flds in Vnf-encoding genomes (Fig. S2B). There is no experimental evidence that these Flds are involved in electron transfer to nitrogenase, with the exception of NifF (present in a few Nif-A-encoding genomes), which has been shown to transfer electrons to nitrogenase in *Klebsiella pneumoniae* (98, 99) and *A. vinelandii* (24, 27).

The distribution of the five Flds identified in this study varied in diazotrophs with different metabolic backgrounds. Of the five Flds identified among the genomes of diazotrophs, NifF and FldA were only detected in the genomes of aerobes, facultative anaerobes, or phototrophs that encoded Nif-A (Fig. S2B). The sole exception to this observation was in the genome of the anaerobic spirochete *Spirochaeta smaragdinae* strain DSM 11293, which was found to code for FldA and Nif-B (Table S2). The remaining three Flds were common in the genomes of anaerobic chemotrophs that encoded Nif-B, Nif-C, and Nif-D. This observation is like that made for Fds, where the distribution of Flds in genomes that encode Nif-A were distinct from those that encoded Nif-B, Nif-D, and Nif-D. These collective observations indicate that the Fds/Flds in aerobic/facultatively anaerobic diazotrophs differ from those in anaerobic diazotrophs, leading to the hypothesis that O_2 played a role in the diversification of systems for the delivery of electrons to nitrogenase. Moreover, the Fds/Flds are different among phototrophs and chemotrophs, leading to the hypothesis that the integration of light into the energy metabolism of diazotrophs played a key role in the diversification of systems of electron delivery to nitrogenase.

Distribution of ferredoxin- and flavodoxin-reducing protein homologs in the

genomes of putative diazotrophs. The taxonomic distributions of PFOR, [NiFe]-hydrogenase, [FeFe]-hydrogenase, Rnf, Fix, and FNR homologs are detailed in Table S2. Briefly, genomes that encoded Nif-A and Nif-B coded for homologs of six of the seven putative Fd- or Fld-reducing enzymes (Fig. S3); FNR homologs identified in green sulfur bacteria were not identified in Nif-A-encoding genomes. Nif-A-encoding genomes tended to code for Fix, PFOR, and Rnf homologs, while Nif-B-encoding genomes tended to code for [FeFe]-hydrogenase, [NiFe]-hydrogenase (primarily group 4e) (41), and PFOR (Fig. S3). Genomes that coded for Nif-C were found to code for homologs of four of the seven Fd- or Fld-reducing enzymes, with [NiFe]-hydrogenase (primarily group 4e) (41) and PFOR being the most abundant among these. Nif-D-encoding genomes were found to code for only group 4d [NiFe]-hydrogenase (41) and PFOR. Anf- and Vnf-encoding genomes coded for homologs of five of the seven reducing enzymes. Fix, [FeFe]-hydrogenases, and PFOR were enriched in Anf-encoding genomes, while group 4e [NiFe]-hydrogenases (41), [FeFe]-hydrogenases, and PFOR were enriched in Vnf-encoding genomes.

The shift in the distribution of enzyme homologs capable of reducing Fd or Fld in diazotrophic genomes generally corresponded to the diversification of Nif-A from Nif-B/Nif-C/Nif-D (Fig. 1 and 3). Chemotrophic anaerobic diazotrophs that coded for Nif-B/Nif-C/Nif-D also encoded PFOR, [FeFe]-hydrogenase, or [NiFe]-hydrogenase. PFOR couples the oxidation of pyruvate to the reduction of Fd, and its expression is regulated by the availability of N in a variety of diazotrophs (36, 73, 100, 101), whereas specific

lineages of hydrogenase (both [FeFe] and [NiFe]) have been shown to couple reversible H_2 oxidation to the reduction of Fd (Fig. 3) (41, 102, 103).

Like chemotrophic anaerobes, anaerobic anoxygenic phototrophic green (both sulfur and nonsulfur) bacteria and anaerobic anoxygenic purple (both sulfur and nonsulfur) bacteria tended to encode PFOR (92% and 60% of total genomes, espectively). Anaerobic anoxygenic purple bacteria also encoded Fix, Rnf, [NiFe]-hydrogenase, and [FeFe]-hydrogenase, while green sulfur bacteria also encoded a functionally similar but phylogenetically distinct FNR isoform that is found in *Cyanobacteria* (Fig. 3). Like anaerobic diazotrophs, PFOR was identified in the genomes of aerobic/facultatively anaerobic diazotrophs. However, unlike anaerobic diazotrophs, the genomes of aerobic/facultatively anaerobic diazotrophs also tended to encode Fix, Rnf, and FNR (Fig. 3). Fix and Rnf generate reduced Fd with a more-reduced potential than its substrate NADH via electron bifurcation (27, 56) or by coupling the reaction with reverse ion translocation (either proton or sodium dependent) (55–57), respectively. Fix has been suggested to play a role in supplying reductant to Nif in aerobic diazotrophs (27, 104–106). Consistent with this suggestion, Fix and Rnf are encoded in the genomes of aerobic chemotrophs (50% of these genomes) and anoxygenic phototrophs (68% of these genomes), specifically purple sulfur and nonsulfur bacteria (Fig. 3). Furthermore, homologs of FNR were only identified in the genomes of oxygenic phototrophic and anoxygenic green sulfur phototrophic bacteria, although the FNR homologs identified in green sulfur phototrophic bacteria were phylogenetically distinct from those cyanobacterial FNR. In *Cyanobacteria* it is thought that FNR catalyzes the reduction of

Fd with NADPH generated during carbohydrate oxidation, which is then used to fix N_2 (61); it is not yet known if Fd produced by FNR in green sulfur bacteria is involved in N_2 fixation. Together, these observations suggest that the enzymes that diazotrophs use to generate reduced Fd or Fld to drive N_2 reduction vary in organisms with different metabolisms, in particular those that are able to integrate O_2 or light into their energy metabolism compared to those that are not able to take advantage of these metabolic strategies.

Predicted electron transfer pathways to nitrogenase. The covariation among the distributions of specified nitrogenase lineage homologs, Fd/Fld homologs, and Fd-/Fldreducing-enzyme homologs was examined to identify putative systems of electron delivery to Nif in different metabolic backgrounds (Fig. 4). The primary Fd-/Fldreducing-enzyme homologs in genomes that code for Nif-A are Fix, PFOR, and Rnf; FNR homologs were enriched in the genomes of Nif-A-encoding diazotrophic *Cyanobacteria* (n = 20). The source of electrons to Nif-A is likely either pyruvate in the case of PFOR (107), NADH generated from central metabolism in the case of Fix and Rnf (55, 56, 58), or NADPH in the case of FNR (60, 108).

Genomes that code for Nif-B were associated with different Fd/Fld homologs than Nif-A-encoding genomes and were associated with homologs of three different Fd-/Fld-reducing enzymes: PFOR, [FeFe]-hydrogenase, and [NiFe]-hydrogenase (Fig. 4). FNR was detected in all Nif-B-encoding green sulfur bacterial anoxygenic phototrophs (n = 11). To this end, the source of electrons to reduce Nif-B is predicted to be H_2 in the case of [FeFe]- and [NiFe]-hydrogenases (109, 110), pyruvate in the case of PFOR, or

NADPH in the case of FNR. Genomes that encoded Nif-C and Nif-D coded for a different suite of Fd/Fld homologs potentially involved in electron delivery to Nif than did Nif-A- and Nif-B encoding genomes (Fig. 4). Very little variation was observed in the distribution of putative Fd-/Fld-reducing-enzyme homologs and Fd/Fld homologs in Nif-C-encoding genomes compared to their distribution in genomes that encoded Nif-D. However, unlike Nif-C-encoding genomes, Nif-D-encoding genomes, which are derived from methanogens, lack evidence for homologs of [FeFe]-hydrogenase. This is consistent with the absence of genes encoding [FeFe]-hydrogenase in Archaea (102, 103). Thus, like Nif-B, the source of electrons to reduce Nif-C and Nif-D is likely H_2 or pyruvate.

Anf- and Vnf-encoding diazotrophic genomes also encode Nif, which in the case of Anf was typically Nif-A and for Vnf was typically Nif-B. As such, patterns in the distribution of Fds/Flds and Fd/Fld-reducing enzymes in Anf- and Vnf-encoding genomes are like Nif-A and Nif-B, respectively. The organisms that encode Anf are likely to rely on three putative Fd-/Fld-reducing enzymes: [FeFe]-hydrogenase, Fix, or PFOR (Fig. S4). As in Nif-A-encoding organisms, the source of electrons to Anf is predicted to be from pyruvate or from central metabolism in the form of NADH. The organisms that encode Vnf are likely to rely on three putative Fd-/Fld-reducing enzymes: PFOR, [FeFe]-hydrogenase, and [NiFe]-hydrogenase (Fig. S4). As such, the source of electrons used to reduce Vnf is likely from H_2 or pyruvate.

Intriguingly, the genomes of several putative diazotrophs (n = 25) lacked a homolog of PFOR, [NiFe]-hydrogenase, [FeFe]-hydrogenase, Rnf, Fix, or either isoform of FNR (Table S2). The inability to detect homologs of these seven enzymes was not due

to these genomes being incomplete, since all of them were closed. Of the 25 genomes that lacked a homolog of these enzymes, 14 encoded Nif-A-like nitrogenases, while 11 encoded Nif-B-like nitrogenases. All 14 genomes that encoded Nif-A-like nitrogenases were classified as either aerobic or facultatively anaerobic, while all 11 genomes that encoded Nif-B-like nitrogenases were classified as strictly anaerobic. These data suggest the presence of at least one additional enzymatic mechanism for generating Fds or Flds that can be used to reduce N_2.

Possible drivers in the diversification of pathways for electron transfer to nitrogenase. Previous studies have shown that Nif diversified in large part in response to O_2, in particular the integration of O_2 into cellular metabolism (17, 111). By extension, this indicates that O_2 would have been at least temporarily available in the local habitat of the ancestors of aerobic/facultatively anaerobic diazotrophs, which would increase the oxidation state of their local habitat. Previous studies have noted a strong correlation between the average oxidation state of carbon in archaeal and bacterial proteomes (inferred from metagenomic data) and the oxidation state of the local environment (112). Likewise, extracellular proteins from yeast have been shown to have a higher average oxidation state of carbon than cytoplasmic proteins, an observation that was attributed to a higher oxidation state on the exterior of the cell than in the interior (113). As has been suggested previously, a shift in the oxidation state of a system to a more-oxidizing potential favors the formation of products (e.g., proteins) that themselves are more oxidized (113). Since the oxidation state of the cytoplasm of a cell should be related to that of the external environment of a cell, and cells that are under selection should be

under selection to minimize the cost of protein synthesis to the extent that those proteins remain functional, we hypothesized that the integration of O_2 into the metabolism of diazotrophs would be accompanied by an overall shift in the oxidation state of the proteome of those cells that may in turn influence the functionality or stability of Fd/Flds, the enzymes that function to reduce these cofactors, or the availability of substrates for these enzymes. Observations supporting this hypothesis would serve as evidence that O_2 had a fundamental role in the evolution of diazotrophs, beyond what has been recognized previously, which includes physiological mechanisms of spatial/temporal decoupling of nitrogenase with O_2 metabolism (18, 19), increased consumption of O_2 through respiratory activity (20), and recruitment and loss of genes involved in regulating nitrogenase expression and activity (17).

To test this hypothesis, we calculated the average oxidation state of carbon in inferred proteomes for the taxa that comprised organisms encoding each of the four Nif sublineages (Fig. 5). Welch's *t* test showed that the average oxidation state of carbon in inferred proteomes was significantly ($P < 2.2 \times 10^{-16}$) more positive (-0.024 ± 0.018) for Nif-A-encoding genomes than for Nif-B-encoding genomes (-0.043 ± 0.019), the most closely related lineage (Fig. 1). Likewise, the average oxidation state of carbon in inferred proteomes was significantly ($P = 1.9 \times 10^{-8}$ and 1.8×10^{-8}, respectively) more positive for Nif-A-encoding genomes than for genomes encoding Nif-C (-0.078 ± 0.033) and Nif-D (-0.058 ± 0.016). These observations are consistent with the observation that most of the Nif-A-encoding genomes were from aerobic or facultatively anaerobic taxa, whereas the majority of Nif-B-, Nif-C-, and Nif-D-encoding genomes were from strict

anaerobes and, to a lesser extent, facultative anaerobes (Fig. 2A). These observations may be due to Nif-A-encoding diazotrophs inhabiting, on average, a more oxidized environment than Nif-B-, Nif-C-, and Nif-D-encoding diazotrophs. If true, this finding would suggest that Nif-A-encoding cells also may harbor a more oxidized cytoplasm than strictly anaerobic diazotrophs. A caveat to this analysis is that it assumes that all proteins in a diazotrophic cell are expressed at the same level, which is unlikely to be the case. Further experiments are needed to examine whether the average oxidation state of a cellular proteome is sensitive to growth conditions (e.g., aerobic versus anaerobic) and whether such a response reflects acclimatization to minimize energetic costs associated with protein biosynthesis.

We hypothesized that differences in the oxidation state of the external environment of cells and their cytoplasm would affect the functionality of proteins involved in electron transfer to nitrogenase and might account for the differences in the Fds/Flds used by aerobic/facultative anaerobic and strictly anaerobic diazotrophs. To test this hypothesis, we reconstructed the evolutionary history of several Fds/Flds that appear to be commonly used as electron donors to the various lineages of Nif. The Fds FdxA and FdxB, which are commonly identified in Nif-A-encoding genomes, branch among (are nested within) a variety of Fd lineages commonly identified in anaerobic diazotrophs (i.e., Nif-B-, Nif-C-, and Nif-D-encoding taxa). These include MvFd4, MvFd5, CsFd6, MvFd2, and CpFd4 (Fig. S5). Parsimony would suggest that FdxA and FdxB are recently evolved and that these Fds evolved from Fds that likely functioned in an anaerobe. Similarly, NifF and FldA were nested among Flds commonly identified in anaerobic

diazotroph genomes, including CpFld1, CpFld3, and CpFld2 (Fig. S6). Among Flds, NifF and FldA formed a monophyletic lineage and were sister to CpFld2, suggesting that they diverged from an ancestor of these proteins and that this ancestral protein most likely functioned in an anaerobe.

A previous study attempted to identify structural features associated with Fds from *Anabaena variabilis* that lend stability in oxic versus anoxic environments (114). The authors showed that slight changes in residues near the active-site FeS cluster protected it from reactive oxygen species. This indicates that the Fds that are active in aerobes inhabiting oxic environments likely have different amino acid compositions than Fds found in anaerobes, a finding that is supported by our informatics and phylogenetics analyses. As such, the results presented here provide a suite of Fds/Flds that appear to have evolved to function optimally in aerobic versus anaerobic diazotrophs and, thus, provide a template for downstream studies aimed at further elucidating the structural features that enable Fd/Fld function in more-oxidized environments.

The more-positive oxidation state of carbon in the inferred proteomes of Nif-A encoding diazotrophs, which are largely from aerobic or facultatively anaerobic taxa, may indicate that the cytoplasm or local environment of those cells is more oxidizing. In general, H_2 (which itself has a low oxidation-reduction potential) would be expected to be at low abundance in an oxidized environment (112, 113, 115, 116). It follows that the diversification of Nif into more-oxidizing environments may have been accompanied by a decrease in available H_2, which in turn may have represented a selective pressure to evolve mechanisms to generate reduced Fd/Fld other than those that are dependent on H_2.

In response, Nif-A-encoding diazotrophs would have continued to use PFOR, if available, to meet these demands. In other aerobic or facultatively anaerobic and anoxygenic phototrophic Nif-A-encoding diazotrophs, the primary electron carrier involved in central metabolism likely switched to NADH (51, 52). Under such conditions, diazotrophic cells would have been under selective pressure to evolve mechanisms to drive the formation of reduced Fd from NADH, such as Fix and Rnf.

Concluding remarks. Diazotrophs have evolved elegant mechanisms to overcome O_2 toxicity to nitrogenase during the transition from anaerobic to aerobic metabolism, including spatial and temporal decoupling of N_2 fixation activity from O_2 respiration (18, 19), increased O_2 respiration to maintain an anoxic cytoplasm (20), and recruitment and loss of genes involved in regulating nitrogenase expression and activity (17). Our data show that the average oxidation state of carbon in the inferred proteomes of diazotrophs also likely increased during the transition from anaerobic to aerobic metabolism, which would have driven large-scale changes in the composition of proteins and enzymes that drive the energy metabolism of diazotrophic cells, as well as the availability of substrates for those enzymes. Indeed, our analysis of the inferred systems of electron delivery to Nif reveals that substantial changes took place during the diversification of diazotrophs. These changes include those at the level of the primary electron donors that provide reductant to Nif, with early-evolving chemotrophic anaerobic diazotrophs likely supporting N_2 reduction with electrons derived from oxidation of H_2 or pyruvate. Later-evolving aerobic/facultatively anaerobic diazotrophs likely support this activity with electrons primarily in the form of NADH/NADPH

derived from central metabolism or cyclic electron pathways in anoxygenic phototrophs. Similarly, a shift in the enzymes used to generate reduced Fd was observed, with chemotrophic anaerobic taxa inferred to be primarily dependent on [NiFe]-hydrogenase and PFOR and aerobic, facultative anaerobic or anoxygenic phototrophic taxa largely dependent on Fix and, to a lesser extent, PFOR and Rnf. Finally, a nearly complete turnover in the putative Fds/Flds in aerobic/facultatively anaerobic versus anaerobic diazotrophic taxa was observed. Collectively, these data are consistent with previous reports suggesting that O_2 had a profound influence on the evolution of nitrogenase (7, 13, 16, 17, 111) and further suggest that these changes may have been a global adaptive response to an increased oxidation state of the local environment.

The ability of cells to harness light energy likely also impacted systems of electron delivery to Nif. Due to the O_2 sensitivity of Nif, oxygenic phototrophs likely utilize FNR to generate reduced Fd from NADPH produced during carbohydrate fermentation at night or in specialized cells. The photosystems of anoxygenic green sulfur bacteria that also encode FNR energize electrons and shuttle them through a cyclic electron transport chain that involves Fd. It is possible that this Fd can be used to reduce Nif directly. Alternatively, these cells may reduce Fd with PFOR, which utilizes pyruvate produced from the oxidation of glycogen. In contrast, anoxygenic phototrophic bacteria, in particular purple sulfur and nonsulfur bacteria, generate reduced quinones during light-driven electron transport, with those electrons supplied by the oxidation of exogenous inorganic or organic electron donors. The potential of these electrons can be further

reduced through reverse electron transfer such that it is low enough to reduce NADH. These phototrophs often then utilize Fix or Rnf to drive the reduction of Fd with NADH.

The collective insights described herein suggest that the genomic and metabolic backgrounds of diazotrophs are associated with wholesale changes in the source of electron donors, enzymes involved in coupled oxidation of electron donors to the reduction of Fds/Flds, and the Fds/Flds used to reduce Nif. These changes were associated with differences in the ability to integrate O_2 and light into the energy metabolism of the cells. While the informatics-based data that are presented here are predictions of potential systems of electron delivery to Nif, in the case of Nif-Aencoding taxa, they are largely consistent with available biochemical data. However, the validity of the predictions made for pathways of electron delivery to Nif-B-, Nif-C-, and Nif-D-encoding taxa is unknown, since biochemical data have yet to be compiled for model taxa encoding homologs of these enzymes. These include diazotrophs that represent the most ancestral forms of Nif in both chemotrophic and phototrophic genomic and metabolic backgrounds.

MATERIALS AND METHODS

Identification and compilation of nitrogenase homologs. Genomes that encode the alpha (D) subunit of nitrogenase (i.e., NifD, AnfD and VnfD) were compiled as previously described (17, 117). Briefly, NifD homologs that were extracted using BLASTp were first aligned along with paralogs using Clustal Omega (118). The Nif-/Anf-/VnfD paralogs identified included ChlN from *Chlamydomonas reinhardtii*

(ACJ50143) and *Synechocystis* sp. strain PCC 6803 (WP_010874227), BchN from *Acidiphilium rubrum* (BAA76536) and *Chloroflexus aurantiacus* (WP_012258416), and NflD from *Methanocaldococcus jannaschii* (WP_010870941) and *Methanosarcina mazei* (AAM30211). The aligned sequences were subjected to maximum-likelihood phylogenetic reconstruction with RAxML (version 7.3.0) (119), specifying the LG substitution matrix, the PROTGAMMA option, and 1,000 bootstrap iterations. Only Anf-/Vnf-/NifD protein homologs that clustered with previously characterized Nif/-Anf-/VnfD lineages (1, 13, 16) were extracted for downstream analysis. Furthermore, only genomes that encode the minimum set of proteins hypothesized to be required for N_2 fixation via Nif (i.e., *nifHDKENB*) or Anf/Vnf (i.e., *anfHDK*/*vnfHDK*) were retained, as previously described (7). In addition, genomes that encode homologs of just *nifHDKEB* were also retained, given that physiological studies suggest that organisms with this minimum gene complement can assimilate N2 (11, 78). A total of 359 genomes from putative diazotrophs were compiled. To retrieve corresponding homologs of the iron protein (i.e., NifH, AnfH, and VnfH) and the beta subunit (i.e., NifK, AnfK, and VnfK), we performed a BLASTp search with NifH (ACO76403) and NifK (ACO76405) from *Azotobacter vinelandii* as queries, specifying cutoffs of 30% percent amino acid sequence identities and 60% sequence coverage. Only subunits that colocalized (e.g., *nifHDK* in an apparent operon) were retained. Each of the three nitrogenase subunits (i.e., H, D, and K subunits of Anf, Vnf, and Nif) was aligned individually with Clustal Omega (118), using the default settings, and the resultant alignment blocks were concatenated. The

concatenated HDK sequences were subjected to phylogenetic reconstruction as described above.

Identification of homologs of Fd, Fld, and putative Fd- or Fld-reducing enzymes in the genomes of diazotrophs. To identify Fds and Flds in the genomes of putative diazotrophs, we first identified all the Fds and Flds in representative organisms for each Nif sublineage. These included *Azotobacter vinelandii, Klebsiella pneumoniae, Rhodopseudomonas palustris,* and *Rhodobacter capsulatus* for Nif-A, *Clostridium pasteurianum* and *Methanosarcina barkeri* for Nif-B, *Caldicellulosiruptor saccharolyticus* and *Methanocaldococcus vulcanius* for Nif-C, and *Methanococcus maripaludis* for Nif-D. All protein sequences that were annotated as Fds or Flds in the genomes of the above-named taxa were compiled. These protein sequences were subjected to Conserved Domain Database (120) BLAST to identify the type of iron-sulfur clusters present in the Fd and to confirm that the extracted Fld contained a flavin binding motif(s). All the compiled and curated Fds and Flds identified in these representative genomes were separately aligned using Clustal Omega (118). The aligned Fd or Fld sequences were subjected to maximum-likelihood phylogenetic reconstruction with RaxML (version 7.3.0) (119), specifying the LG substitution matrix and the PROTGAMMA option to empirically cluster the sequences into unique groups. In total, 48 Fds (Table 1) and five unique Flds (Table 2) were identified among diazotrophic genomes. The amino acid sequences of all the Fds and Flds were then used as bait sequences to identify homologs in the genomes of putative diazotrophs using BLASTp.

Putative enzymes that have been implicated in providing reductant to nitrogenase in the form of reduced Fd and/or Fld include the following: (i) pyruvate-flavodoxin oxidoreductase (PFOR) (35–38), (ii) ferredoxin-$NADP^+$ oxidoreductase (FNR) (35, 62), (iii) *Rhodobacter* nitrogen fixation protein (Rnf) (81), and (iv) electron transfer flavoprotein (Fix) (27, 53, 54). Homologs of these enzymes were compiled from genomes of putative diazotrophs using BLASTp. The amino acid sequences of PFOR from *Klebsiella pneumoniae* (WP_064371580) and *Rhodopseudomonas palustris* (CAE30161), FNR from *Anabaena azollae* (WP_013193003) and *Chlorobium tepidum* strain TLS (AAM72739), Rnf from *Azotobacter vinelandii* (ACO81179 to ACO81185), and Fix from *Rhodopseudomonas palustris* (WP_011665892 to WP_011665895) were used as query sequences. In addition, [FeFe]-hydrogenase has been suggested to provide reductant for N_2 fixation in *Clostridium pasteurianum* (39). Thus, the catalytic subunit of [FeFe] hydrogenase (HydA) from *Chlamydomonas reinhardtii* (AAL23572) was used as the query in a BLASTp search. Finally, group 3c, 3d, 4d, and 4e [NiFe]-hydrogenases catalyze the reversible reduction of Fd with H_2 (41, 42, 102). A group 3c representative ([NiFe]-hydrogenase from *Methanothermobacter marburgensis* [WP_013296467]) and a group 4 representative ([NiFe]-hydrogenase from *Pyrococcus abyssi* [WP_010867842.1]) were used as BLASTp queries to identify all group 3 and 4 [NiFe]-hydrogenase large-subunit sequences. Compiled sequences were aligned as described above and manually curated to include only those enzymes that had proximal and distal Cys-Cys pairs, the ligands for the [NiFe] active site (121, 122). Extracted sequences homologous to [NiFe] were then subjected to phylogenetic reconstruction, as described above, and assigned to

their respective groups based on phylogenetic coherence, as we have described previously (121, 122). Only [NiFe]-hydrogenase homologs belonging to groups 3c, 3d, 4d, and 4e that have been proposed to be capable of reducing Fd were considered in this study (40–44).

Identifying cooccurrence patterns in the distribution of Fds/Flds and putative Fd/Fld-reducing enzymes. A binary table was created based on the presence or absence of identified Fds, Flds, and Fd or Fld-reducing enzymes in diazotrophic genomes. To identify the frequent patterns in this data set, the Apriori algorithm was used as implemented with the arules package in R (123, 124). For simplicity, only the Fds, Flds, and Fd- or Fld-reducing enzymes present in >20% of the genomes were considered for this analysis. We applied the Apriori algorithm to our binary database while specifying the following parameters: a confidence threshold of 0.60 and a support threshold of 0.2. The confidence threshold is the measure of the statistical significance of an identified pattern within a data set, while the support threshold is indicative of the relative abundance of a given pattern in a data set. The Apriori algorithm works in two phases, with the first phase consisting of a scan of the entire database to objectively identify and extract frequently observed proteins. In the second phase, the algorithm uses those frequently occurring proteins to identify the dominant patterns in the distribution of proteins or combinations of those proteins. In doing so, the algorithm identifies correlations (within the user-defined confidence and support thresholds) between and among proteins in the database.

Oxidation state of carbon. The oxidation state of carbon for each amino acid in each protein sequence encoded in 359 putative diazotrophic genomes was calculated following the algorithm mentioned in reference 112 using a custom python script (version 2.7.12). The calculated oxidation number for each amino acid was then normalized to the total number of amino acids encoded in a genome to determine the average oxidation state of the proteome. This number was determined for each genome that encoded a specific Nif isoform. These numbers were summed and then divided by the total number of genomes to arrive at the average oxidation state of carbon in a proteome inferred for a genome that encodes a given nitrogenase lineage. Welch's t test was conducted to determine the statistical significance of differences between the average oxidation states of carbon in proteomes from different lineages using the stats package in R (125). Using R, box plots were produced to show the distribution of the oxidation state of carbon in proteomes inferred for genomes that encoded specific isoforms of nitrogenases (125).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/JB.00757-17.

SUPPLEMENTAL FILE 1, PDF file, 0.5 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.1 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported as part of the Biological Electron Transfer and Catalysis Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, and Basic Energy Sciences under award number DE-SC0012518.

TABLE 1 Ferredoxin homologs identified in the genomes of putative diazotrophs[a]

Annotation	Protein accession no.	Type of cluster	Reference(s) or source
asl2914	WP_013190310	2[4Fe-4S]	126
FdxH	WP_013190616	[2Fe-2S]	127
PetF	WP_013189829	[2Fe-2S]	128
FdxA	ACO80005	[3Fe-4S] [4Fe-4S]	129
FdxI	ACO76607	[4Fe-4S]	130
FdxN	ACO81189	2[4Fe-4S]	82, 131
VnfF	ACO76526	[4Fe-4S]	132
XylT	ACO77112	CXXXXCXXXXCXXC	130
AvFd1	AGK16019	[2Fe-2S]	This study
AvFd2	AGK16622	[2Fe-2S]	This study
AvFd3	AGK13276	2[4Fe-4S]	This study
CsFd1	ABP66178	2[4Fe-4S]	This study
CsFd2	ABP66040	2[4Fe-4S]	This study
CsFd3	ABP66582	2[4Fe-4S]	This study
CsFd4	ABP67884	2[4Fe-4S]	This study
CsFd5	ABP67132	2[4Fe-4S]	This study
CsFd6	ABP67141	2[4Fe-4S]	This study
CpFd1	AJA47502	[2Fe-2S]	This study
CpFd2	AJA47129	2[4Fe-4S]	This study
CpFd3	AJA49513	2[4Fe-4S]	This study
CpFd4	AJA46278	[4Fe-4S]	This study
CpFd5	AJA49585	2[4Fe-4S]	This study
KpFd1	BAS34286	[2Fe-2S]	This study
KpFd2	BAS37351	[2Fe-2S]	This study
KpFd3	BAS33241	[2Fe-2S]	This study
KpFd4	BAS35880	CXXCXXCC	This study
MvFd1	ACX73019	2[4Fe-4S]	This study
MvFd2	ACX72493	2[4Fe-4S]	This study
MvFd3	ACX73480	2[4Fe-4S]	This study
MvFd4	ACX72328	2[4Fe-4S]	This study
MvFd5	ACX73314	2[4Fe-4S]	This study
MvFd6	ACX73502	2[4Fe-4S]	This study
MvFd7	ACX72560	2[4Fe-4S]	This study
MvFd8	ACX72349	2[4Fe-4S]	This study
MmFd1	CAF29654	2[4Fe-4S]	This study
MbFd1	AKB59153	2[4Fe-4S]	This study
MbFd2	AKB57181	2[4Fe-4S]	This study
MbFd3	AKB57361	2[4Fe-4S]	This study
MbFd4	AKB58903	2[4Fe-4S]	This study
MbFd5	AKB59009	2[4Fe-4S]	This study
FdxC	ADE87009	[2Fe-2S]	97, 133
FdxD	ADE84338	[2Fe-2S]	80
FdxE	ADE86134	[2Fe-2S]	93
FdxB	ABJ08445	2[4Fe-4S]	81
Fer1	ADU46354	[4Fe-4S]	134
FerN	ACF03599	[4Fe-4S]	134
RpFd1	ABJ04561	2[4Fe-4S]	This study

[a]The protein annotation, accession number for a representative protein homolog, inferred type of [FeS] cluster in a given protein homolog (as predicted via the conserved domain database [120]), and a literature reference, if available, are provided for each Fd identified. Av, *A. vinelandii*; Cs, *Caldicellulosiruptor saccharolyticus*; Cp, *Clostridium pasteurianum*; Kp, *Klebsiella pneumoniae*; Mv, *Methanocaldococcus vulcanius*; Mm, *Methanococcus maripaludis*; Mb, *Methanosarcina barkeri*; Rp, *Rhodopseudomonas palustris.*

TABLE 2 Flavodoxin homologs identified in the genomes of putative diazotrophs

Annotation	Accession Number	Source
NifF	ACO76434	(Nieva-Gomez *et al*, 1980)
CpFld1	AJA46461	This study
CpFld2	AJA47463	This study
CpFld3	AJA47660	This study
FldA	WP_011157672	(Armengaud *et al*, 1997)

The protein annotation, accession number for a representative protein homolog, and a literature reference, if available, is provided for each Fld identified. Cp, *Clostridium pasteurianum*.

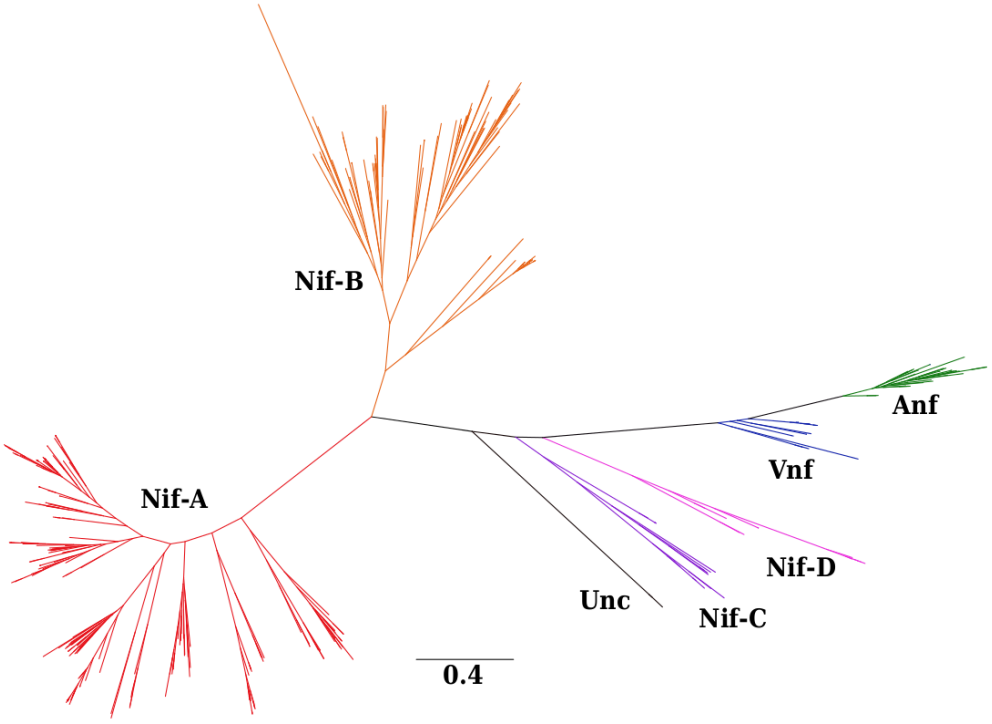


FIG 1 Phylogenetic reconstruction of a concatenation of H, D, and K subunits of nitrogenase and uncharacterized nitrogenase paralogs (*n* = 420 concatenated protein sequences). All nodes shown exhibited bootstrap supports of >90% (out of 1,000 bootstrap replicates), except where a black box (>70%) is shown. Nif, molybdenum (Mo) nitrogenase; Anf, iron-only (Fe) nitrogenase; Vnf, vanadium (V) nitrogenase; Unc, uncharacterized nitrogenase-like proteins.

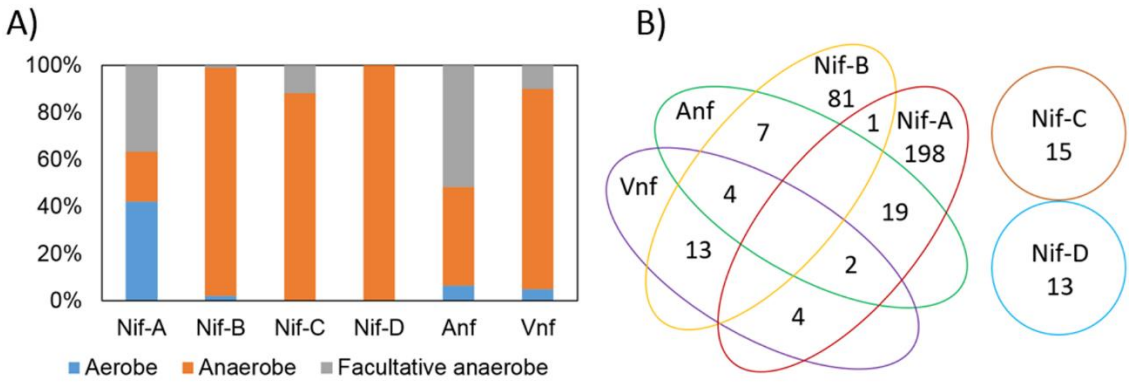


FIG 2 Homologs of nitrogenase identified in the genomes of putative diazotrophs. (A) Histogram depicting the percentage of diazotrophs within each specified Nif lineage (see Fig. 1) that are aerobic, anaerobic, or facultatively anaerobic, as determined from surveys of the literature for cultivated organisms. (B) Venn diagram representing the numbers of genomes that encode one or more specified lineages of nitrogenase (see Fig. 1). Genomes that encode Nif-C and Nif-D do not encode other isoforms of nitrogenase and are thus depicted as separate.

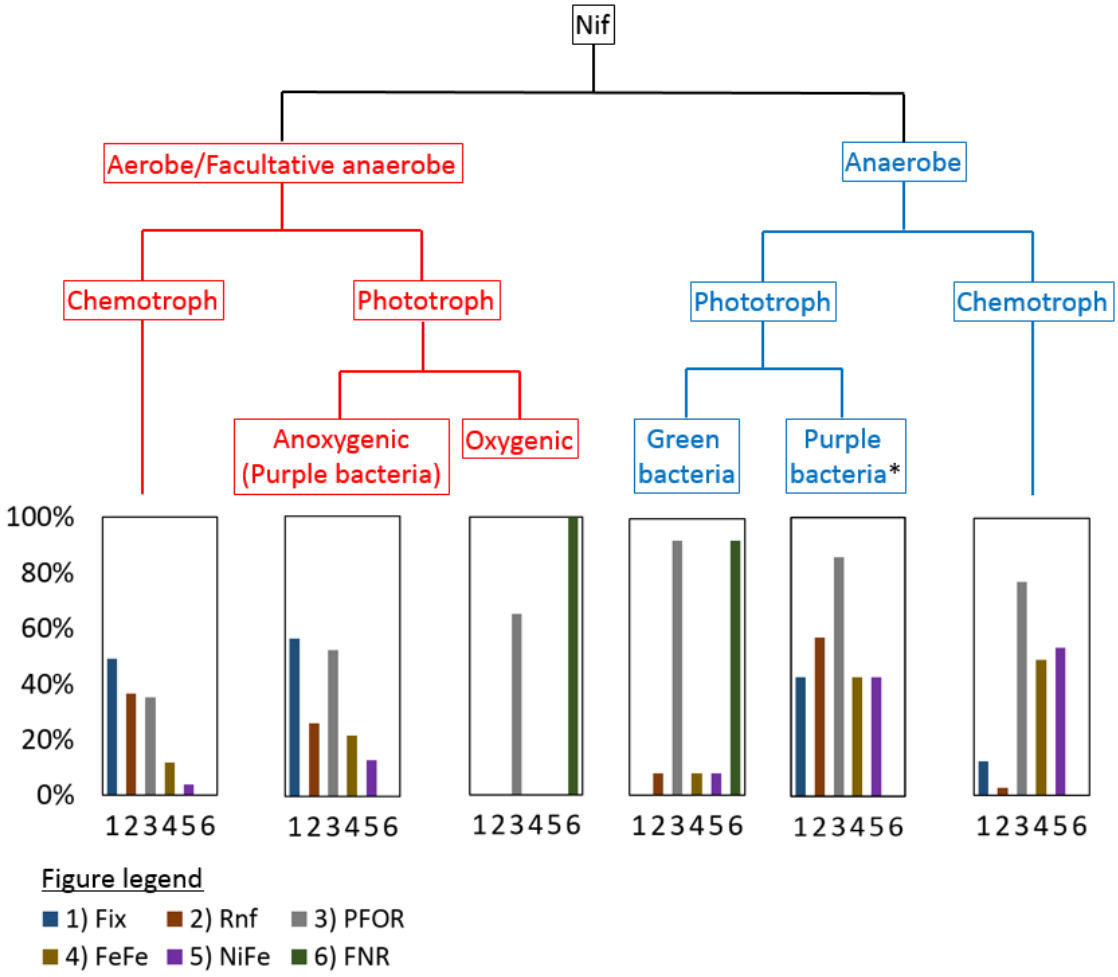


FIG 3 Flow chart of nitrogenase in different metabolic backgrounds and histograms depicting the percentage of enzyme homologs putatively involved in Fd/Fld reduction identified in the genomes comprising a specified group. Aerobic/facultative anaerobic and anaerobic organisms are further classified as phototrophs or chemotrophs. Phototrophs in aerobes/facultative anaerobes are further classified based on whether they are oxygenic or anoxygenic phototrophs, while anaerobic phototrophs are classified as either purple sulfur/nonsulfur bacteria or green sulfur/nonsulfur bacteria. Importantly, it is not clear from surveys of the literature that anaerobic purple bacterial strains (denoted by an asterisk) were robustly tested for their ability to use O_2. FNR, ferredoxin-$NADP^+$ oxidoreductase; PFOR, pyruvate-flavodoxin oxidoreductase; Rnf, *Rhodobacter* nitrogen fixation protein; FeFe, iron-only hydrogenase; NiFe,nickel-iron hydrogenase; Fix, electron transfer flavoprotein involved in nitrogen fixation.

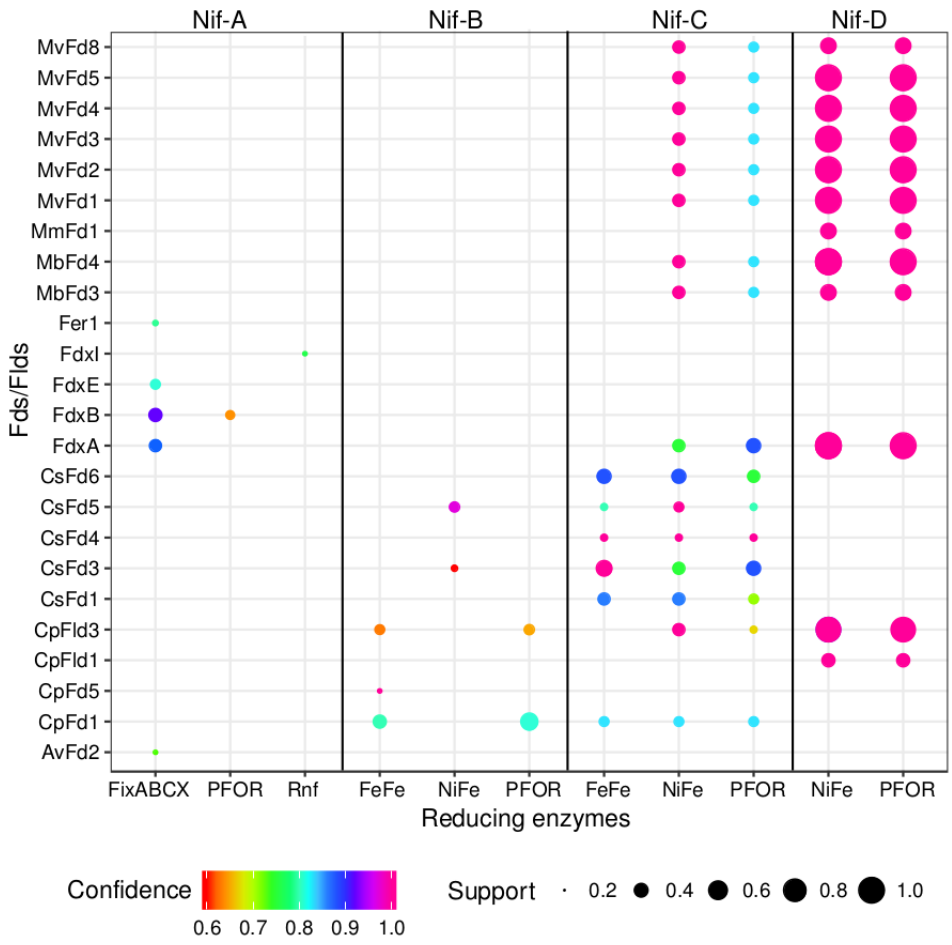


FIG 4 Bubble plot depicting the dominant patterns in the distribution of putative electron carrier protein homologs (Fds/Flds), enzyme homologs that putatively reduce Fd/Fld, and specified Nif lineages, as determined by the Apriori algorithm (124). Each unique pattern is given as a bubble, and the color represents the confidence value, or statistical significance, in the codistribution of specified proteins (only confidence values of 0.6 are presented). The size of the bubble represents the support value (0.2), or the frequency with which two proteins are identified in the same genome. For simplicity, only the proteins that were present in >20% of the diazotrophic genomes for each specified nitrogenase lineage were considered in this analysis. PFOR, pyruvate-Fld oxidoreductase; Rnf, *Rhodobacter* nitrogen fixation protein; FeFe, iron-only hydrogenase; NiFe, nickel-iron hydrogenase; FixABCX, electron transfer flavoprotein that is involved in nitrogen fixation. Names of source organisms for Fds and Flds are abbreviated as follows: Mv, *Methanocaldococcus vulcanius*; Mm, *Methanococcus maripaludis*; Mb, *Methanosarcina barkeri*; Cs, *Caldicellulosiruptor saccharolyticus*; Cp, *Clostridium pasteurianum*; Av, *A. vinelandii*. Abbreviations for Fds and Flds are presented in Tables 1 and 2, respectively.

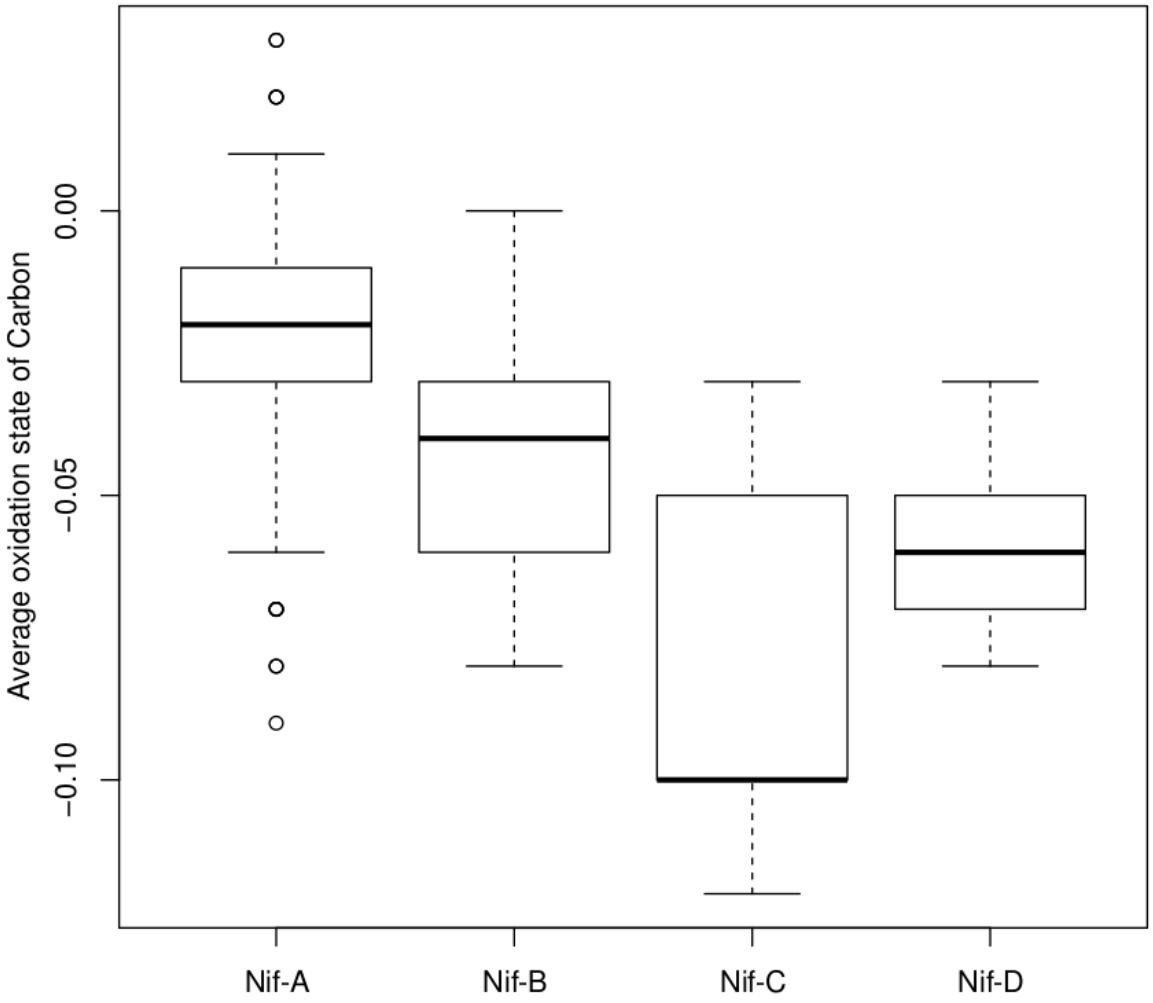
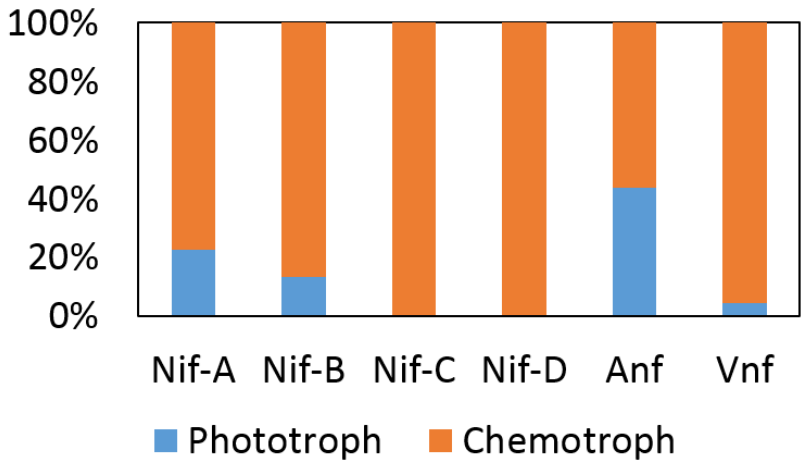
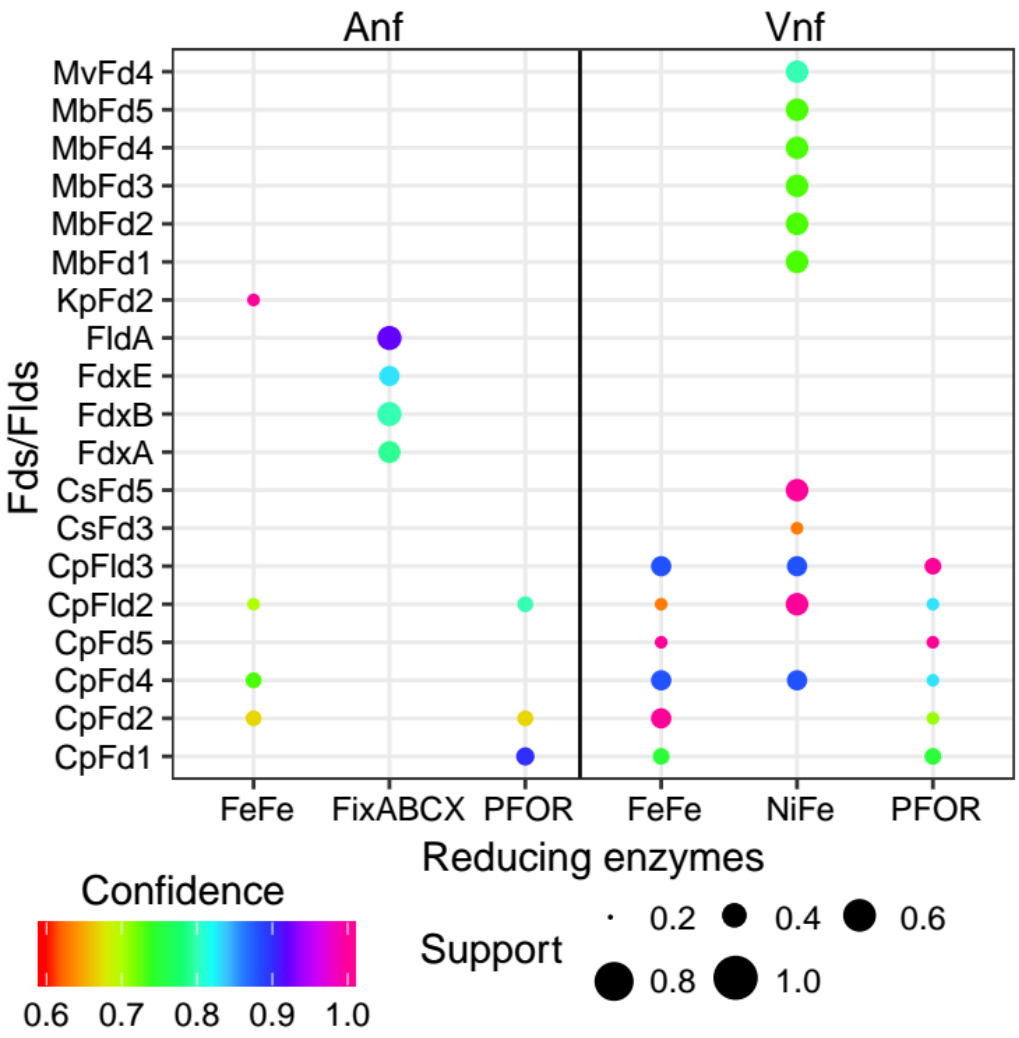


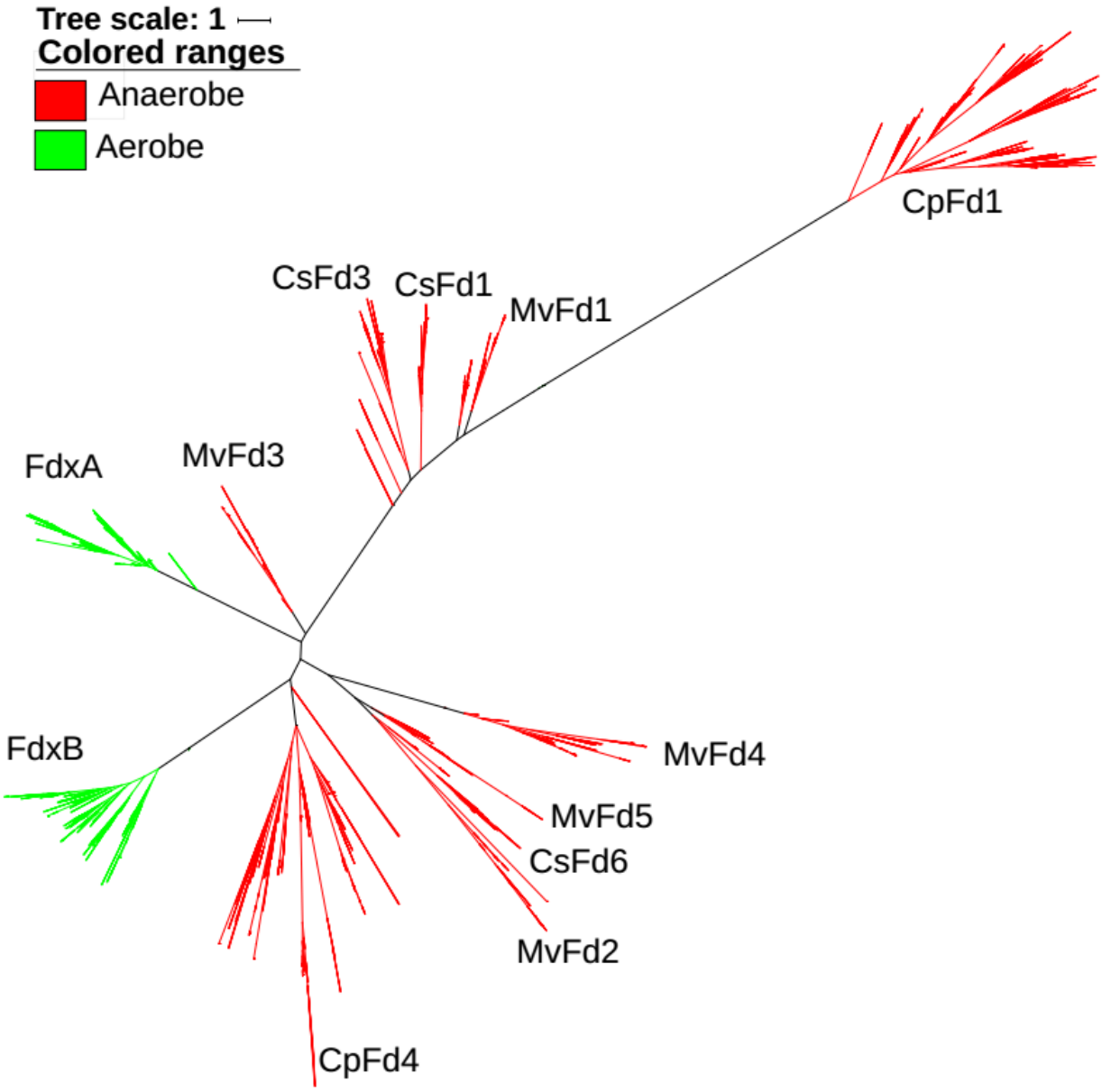
FIG 5 Box plot representing the average oxidation state of carbon in proteomes (inferred from genomes) for the taxa comprising organisms encoding the specified nitrogenase isoform. Here, the box represents the interquartile range and the whiskers show the full range of the data. Outlier values are represented as circles, and the line within the box represents the median value.



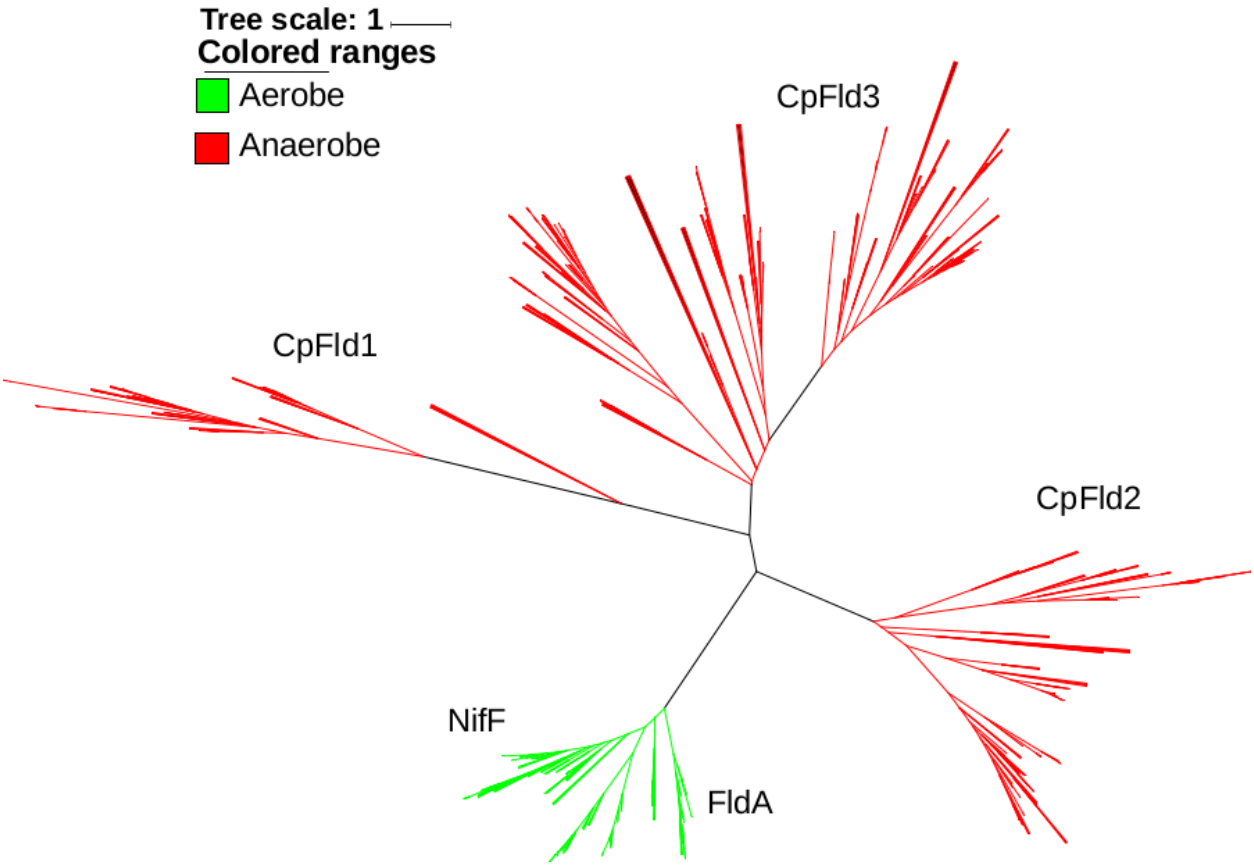
Supplemental Figure 1. Histogram depicting the percentage of diazotrophs within each specified Nif lineage (see **Fig. 1**) that are phototrophic or chemotrophic, as determined from literature surveys of cultivated organisms.



Supplemental Figure 2. Bubble plot depicting the dominant patterns in the distribution of putative electron carrier proteins (Fds, Flds), enzyme homologs putatively involved in the reduction of Fds/Flds, and Anf and Vnf in genomes coding for specified Anf and Vnf lineages, as determined by the Apriori algorithm (Agrawal and Srikant, 1994). Each unique pattern is given as a bubble and the color represents the confidence value or the statistical significance in the co-distribution of specified proteins (only confidence values of ≥ 0.6 are presented). The size of the bubble represents the support value (>=0.2) or the frequency of that two proteins are identified in the same genome. For simplicity, only the proteins that were present in >20% of the diazotrophic genomes for each specified nitrogenase lineage were considered in this analysis. Abbreviations: PFOR, pyruvate-flavodoxin oxidoreductase; Rnf, *Rhodobacter* nitrogen fixation protein; FeFe, iron-only hydrogenase; NiFe, nickel-iron hydrogenase; FixABCX, electron transfer flavoprotein that is involved in nitrogen fixation. Abbreviations for Fds and Flds are presented in **Tables 1** and **2**, respectively.



Supplemental Figure 3. Maximum likelihood phylogenetic reconstruction of abundant (present in >40% of the genomes) homologs of Fd (n=917) shown in **Fig. 3A** with an overlay of the aerobicity of the diazotrophic organism that the Fd is primarily associated with: Aerobe (green) and anaerobe (red).



Supplemental Figure 4. Maximum likelihood phylogenetic reconstruction of 405 homologs of Flds given in **Table 2** and **Fig. 3B** with an overlay of the aerobicity of the diazotrophic organism that the Fld is primarily associated with: Aerobe (green) and anaerobe (red).

REFERENCES

1. Agrawal R, Srikant R. Fast algorithms for mining association rules, p 487-499. *In* (ed),

REFERENCES

1. Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The natural history of nitrogen fixation. Mol Biol Evol 21:541–554. https://doi.org/10.1093/molbev/msh047.

2. Vitousek PM, Menge DNL, Reed SC, Cleveland CC. 2013. Biological nitrogen fixation: rates, patterns and ecological controls in terrestrial ecosystems. Philos Trans R Soc B Biol Sci 368:20130119. https://doi.org/10.1098/rstb.2013.0119.

3. Burk D, Lineweaver H, Horner CK. 1934. The specific influence of acidity on the mechanism of nitrogen fixation by *Azotobacter*. J Bacteriol 27:325–340.

4. Falkowski PG. 1997. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. Nature 387:272–275. https://doi.org/10.1038/387272a0.

5. Bulen W, LeComte J. 1966. The nitrogenase system from *Azotobacter*: two-enzyme requirement for N_2 reduction, ATP-dependent H_2 evolution, and ATP hydrolysis. Proc Natl Acad SciUSA 56:979 –986. https://doi.org/10.1073/pnas.56.3.979.

6. Peters JW, Boyd ES, Hamilton T, Rubio LM (ed). 2011. Biochemistry of Mo-nitrogenase. Caister Academic Press, Poole, United Kingdom.

7. Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW. 2011. A late methanogen origin for molybdenum-dependent nitrogenase. Geobiology 9:221–232. https://doi.org/10.1111/j.1472-4669.2011.00278.x.

8. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13:162. https://doi.org/10.1186/1471-2164-13-162.

9. McGlynn SE, Boyd ES, Peters JW, Orphan VJ. 2013. Classifying the metal dependence of uncharacterized nitrogenases. Front Microbiol 3:419. https://doi.org/10.3389/fmicb.2012.00419.

10. Leigh JA. 2000. Nitrogen fixation in methanogens: the archaeal perspective. Curr Issues Mol Biol 2:125–131.

11. Mehta MP, Baross JA. 2006. Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. Science 314:1783–1786. https://doi.org/10.1126/science.1134772.

12. Nishizawa M, Miyazaki J, Makabe A, Koba K, Takai K. 2014. Physiological and isotopic characteristics of nitrogen fixation by hyperthermophilic methanogens: key insights into nitrogen anabolism of the microbial communities in archean hydrothermal

systems. Geochim Cosmochim Acta 138:117–135. https://doi.org/10.1016/j.gca.2014.04.021.

13. Boyd ES, Hamilton TL, Peters JW. 2011. An alternative path for the evolution of biological nitrogen fixation. Front Microbiol 2:205. https://doi.org/10.3389/fmicb.2011.00205.

14. Joerger RD, Bishop PE, Bishop PE. 1988. Bacterial alternative nitrogenfixation systems. Crit Rev Microbiol 16:1–14. https://doi.org/10.3109/10408418809104465.

15. Rubio LM, Ludden PW. 2008. Biosynthesis of the iron-molybdenumcofactor of nitrogenase. Annu Rev Microbiol 62:93–111. https://doi.org/10.1146/annurev.micro.62.081307.162737.

16. Boyd ES, Peters JW. 2013. New insights into the evolutionary history of biological nitrogen fixation. Front Microbiol 4:201. https://doi.org/10.3389/fmicb.2013.00201.

17. Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW. 2015. Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. J Bacteriol 197:1690 –1699. https://doi.org/10.1128/JB.02611-14.

18. Stal LJ, Krumbein WE. 1985. Nitrogenase activity in the nonheterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. Arch Microbiol 143:67–71. https://doi.org/10.1007/BF00414770.

19. Fay P. 1992. Oxygen relations of nitrogen fixation in cyanobacteria. Microbiol Rev 56:340 –373.

20. Poole RK, Hill S. 1997. Respiratory protection of nitrogenase activity in *Azotobacter vinelandii*: roles of the terminal oxidases. Biosci Rep 17:303–317. https://doi.org/10.1023/A:1027336712748.

21. Egener T, Martin DE, Sarkar A, Reinhold-Hurek B. 2001. Role of a ferredoxin gene cotranscribed with the *nifHDK* operon in N_2 fixation and nitrogenase "switch-off" of *Azoarcus* sp. strain BH72. J Bacteriol 183:3752–3760. https://doi.org/10.1128/JB.183.12.3752-3760.2001.

22. Sandmann G, Peleato ML, Fillat MF, Lazaro MC, Gomezmoreno C. 1990. Consequences of the iron-dependent formation of ferredoxin and flavodoxin on photosynthesis and nitrogen fixation on *Anabaena* strains. Photosynth Res 26:119 –125. https://doi.org/10.1007/BF00047083.

23. Yakunin AF, Gennaro G, Hallenbeck PC. 1993. Purification and properties of a nif-specific flavodoxin from the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 175:6775–6780. https://doi.org/10.1128/jb.175.21.6775-6780.1993.

24. Martin AE, Burgess BK, Iismaa SE, Smartt CT, Jacobson MR, Dean DR. 1989. Construction and characterization of an *Azotobacter vinelandii* strain with mutations in the genes encoding flavodoxin and ferredoxin I. J Bacteriol 171:3162–3167. https://doi.org/10.1128/jb.171.6.3162-3167.1989.

25. Yang Z-Y, Ledbetter R, Shaw S, Pence N, Tokmina-Lukaszewska M, Eilers B, Guo Q, Pokhrel N, Cash VL, Dean DR. 2016. Evidence that the Pi release event is the rate-limiting step in the nitrogenase catalytic cycle. Biochemistry 55:3625–3635. https://doi.org/10.1021/acs.biochem.6b00421.

26. Gennaro G, Hubner P, Sandmeier U, Yakunin AF, Hallenbeck PC. 1996. Cloning, characterization, and regulation of *nifF* from *Rhodobacter capsulatus*. J Bacteriol 178:3949 –3952. https://doi.org/10.1128/jb.178.13.3949-3952.1996.

27. Ledbetter RN, Garcia Costas AM, Lubner CE, Mulder DW, Tokmina Lukaszewska M, Artz JH, Patterson A, Magnuson TS, Jay ZJ, Duan HD, Miller J, Plunkett MH, Hoben JP, Barney BM, Carlson RP, Miller AF, Bothner B, King PW, Peters JW, Seefeldt LC. 2017. The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: generation of low-potential reducing equivalents for nitrogenase catalysis. Biochemistry 56:4177–4190. https://doi.org/10.1021/acs.biochem.7b00389.

28. Simpson FB, Burris RH. 1984. A nitrogen pressure of 50 atmospheres does not prevent evolution of hydrogen by nitrogenase. Science 224: 1095–1097. https://doi.org/10.1126/science.6585956.

29. Tso M-YW, Ljones T, Burris R. 1972. Purification of the nitrogenase proteins from *Clostridium pasteurianum*. Biochim Biophys Acta 267:600 – 604. https://doi.org/10.1016/0005-2728(72)90193-4.

30. Seefeldt LC, Hoffman BM, Dean DR. 2009. Mechanism of Mo-dependent nitrogenase. Annu Rev Biochem 78:701–722. https://doi.org/10.1146/annurev.biochem.78.070907.103812.

31. Bergström J, Eady RR, Thorneley RN. 1988. The vanadium- and molybdenum-containing nitrogenases of *Azotobacter chroococcum*. Comparison of mid-point potentials and kinetics of reduction by sodium dithionite of the iron proteins with bound magnesium adenosine 5’-diphosphate. Biochem J 251:165–169.

32. Dilworth MJ, Eldridge ME, Eady RR. 1992. Correction for creatine interference with the direct indophenol measurement of NH_3 in steadystate nitrogenase assays. Anal Biochem 207:6 –10. https://doi.org/10.1016/0003-2697(92)90491-O.

33. Dilworth MJ, Eldridge ME, Eady RR. 1993. The molybdenum and vanadium nitrogenases of *Azotobacter chroococcum*: effect of elevated temperature on N_2 reduction. Biochem J 289:395– 400. https://doi.org/10.1042/bj2890395.

34. Eady RR. 1996. Structure-function relationships of alternative nitrogenases. Chem Rev 96:3013–3030. https://doi.org/10.1021/cr950057h.

35. Saeki K (ed). 2004. Electron transport to nitrogenase: diverse routes for a common destination. Kluwer, Dordrecht, the Netherlands.

36. Nieva-Gomez D, Roberts GP, Klevickis S, Brill WJ. 1980. Electron transport to nitrogenase in Klebsiella pneumoniae. Proc Natl Acad SciUSA 77:2555–2558. https://doi.org/10.1073/pnas.77.5.2555.

37. Wahl RC, Orme-Johnson WH. 1987. Clostridial pyruvate oxidoreductase and the pyruvate-oxidizing enzyme specific to nitrogen fixation in *Klebsiella pneumoniae* are similar enzymes. J Biol Chem 262:10489 –10496.

38. Yakunin AF, Hallenbeck PC. 1998. Purification and characterization of pyruvate oxidoreductase from the photosynthetic bacterium *Rhodobacter capsulatus*. Biochim Biophys Acta 1409:39 – 49. https://doi.org/10.1016/S0005-2728(98)00145-5.

39. Therien J, Artz JH, Poudel S, Hamilton TL, Liu Z, Noone S, Adams M, King PW, Bryant DA, Boyd ES, Peters JW. 2017. The physiological functions and structural determinants of catalytic bias in the [FeFe]-hydrogenases CpI and CpII of *Clostridium pasteurianum* strain W5. Front Microbiol 8:1305. https://doi.org/10.3389/fmicb.2017.01305.

40. Bothe H, Schmitz O, Yates MG, Newton WE. 2010. Nitrogen fixation and hydrogen metabolism in cyanobacteria. Microbiol Mol Biol Rev 74:529 –551. https://doi.org/10.1128/MMBR.00033-10.

41. Greening C, Biswas A, Carere CR, Jackson CJ, Taylor MC, Stott MB, Cook GM, Morales SE. 2016. Genomic and metagenomic surveys of hydrogenase distribution indicate H_2 is a widely utilised energy source for microbial growth and survival. ISME J 10:761. https://doi.org/10.1038/ismej.2015.153.

42. Gutekunst K, Chen X, Schreiber K, Kaspar U, Makam S, Appel J. 2014. The bidirectional NiFe-hydrogenase in *Synechocystis* sp PCC 6803 is reduced by flavodoxin and ferredoxin and is essential under mixotrophic, nitrate-limiting conditions. J Biol Chem 289:1930 –1937. https://doi.org/10.1074/jbc.M113.526376.

43. Khanna N, Lindblad P. 2015. Cyanobacterial hydrogenases and hydrogen metabolism revisited: recent progress and future prospects. Int J Mol Sci 16:10537–10561. https://doi.org/10.3390/ijms160510537.

44. Schmitz O, Boison G, Hilscher R, Hundeshagen B, Zimmer W, Lottspeich F, Bothe H. 1995. Molecular biological analysis of a bidirectional hydrogenase from

cyanobacteria. Eur J Biochem 233:266 –276. https://doi.org/10.1111/j.1432-1033.1995.266_1.x.

45. Rey FE, Oda Y, Harwood CS. 2006. Regulation of uptake hydrogenase and effects of hydrogen utilization on gene expression in *Rhodopseudomonas palustris*. J Bacteriol 188:6143– 6152. https://doi.org/10.1128/JB.00381-06.

46. Herrmann G, Jayamani E, Mai G, Buckel W. 2008. Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. J Bacteriol 190:784 –791. https://doi.org/10.1128/JB.01422-07.

47. Anraku Y. 1988. Bacterial electron-transport chains. Annu Rev Biochem 57:101–132. https://doi.org/10.1146/annurev.bi.57.070188.000533.

48. Watt G, Burns A, Lough S, Tennent D. 1980. Redox and spectroscopic properties of oxidized MoFe protein from *Azotobacter vinelandii*. Biochemistry 19:4926 – 4932. https://doi.org/10.1021/bi00562a035.

49. Lanzilotta WN, Ryle MJ, Seefeldt LC. 1995. Nucleotide hydrolysis and protein conformational changes in *Azotobacter vinelandii* nitrogenase iron protein: defining the function of aspartate 129. Biochemistry 34: 10713–10723. https://doi.org/10.1021/bi00034a003.

50. Braaksma A, Haaker H, Grande HJ, Veeger C. 1982. The effect of the redox potential on the activity of the nitrogenase and on the Fe protein of *Azotobacter vinelandii*. FEBS J 121:483– 491.

51. Dutton PL. 1986. Energy transduction in anoxygenic photosynthesis, p197–237. In Staehelin A, Arntzen CJ (ed), Photosynthesis III. Photosynthetic membranes and light harvesting systems. Encyclopedia of Plant Physiology (new series), vol 19. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-70936-4_5.

52. Jagannathan B, Golbeck J (ed). 2009. Photosynthesis: microbial, p325–341. In Schaechter M (ed), Encyclopedia of microbiology, 3rd ed. Academic press, Cambridge, MA. https://doi.org/10.1016/B978-012373944-5.00352-7.

53. Earl C, Ronson C, Ausubel F. 1987. Genetic and structural analysis of the *Rhizobium meliloti fixA*, *fixB*, *fixC*, and *fixX* genes. J Bacteriol 169:1127–1136. https://doi.org/10.1128/jb.169.3.1127-1136.1987.

54. Edgren T, Nordlund S. 2004. The *fixABCX* genes in *Rhodospirillum rubrum* encode a putative membrane complex participating in electron transfer to nitrogenase. J Bacteriol 186:2052–2060. https://doi.org/10.1128/JB.186.7.2052-2060.2004.

55. Biegel E, Schmidt S, Gonzalez JM, Muller V. 2011. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell Mol Life Sci 68:613– 634. https://doi.org/10.1007/s00018-010-0555-8.

56. Hess V, Schuchmann K, Muller V. 2013. The ferredoxin: NAD(+) oxidoreductase (Rnf) from the acetogen *Acetobacterium woodii* requires Na^+ and is reversibly coupled to the membrane potential. J Biol Chem 288:31496 –31502. https://doi.org/10.1074/jbc.M113.510255.

57. Jouanneau Y, Jeong HS, Hugo N, Meyer C, Willison JC. 1998. Overexpression in *Escherichia coli* of the rnf genes from *Rhodobacter capsulatus*: characterization of two membrane-bound iron-sulfur proteins. Eur J Biochem 251:54 – 64. https://doi.org/10.1046/j.1432-1327.1998.2510054.x.

58. Sarkar A, Köhler J, Hurek T, Reinhold-Hurek B. 2012. A novel regulatory role of the Rnf complex of *Azoarcus* sp. strain BH72. Mol Microbiol 83:408 – 422. https://doi.org/10.1111/j.1365-2958.2011.07940.x.

59. Schrautemeier B, Böhme H. 1985. A distinct ferredoxin for nitrogen fixation isolated from heterocysts of the cyanobacterium *Anabaena variabilis*. FEBS Lett 184:304 –308. https://doi.org/10.1016/0014-5793(85)80627-X.

60. Apte S, Rowell P, Stewart W. 1978. Electron donation to ferredoxin in heterocysts of the N_2-fixing alga *Anabaena cylindrica*. Proc R Soc B Biol Sci 200:1–25. https://doi.org/10.1098/rspb.1978.0001.

61. Aliverti A, Faber R, Finnerty CM, Ferioli C, Pandini V, Negri A, Karplus PA, Zanetti G. 2001. Biochemical and crystallographic characterization of ferredoxin-NADP(+) reductase from nonphotosynthetic tissues. Biochemistry 40:14501–14508. https://doi.org/10.1021/bi011224c.

62. Isas JM, Yannone SM, Burgess BK. 1995. *Azotobacter vinelandii* NADPH-ferredoxin reductase cloning, sequencing, and overexpression. J Biol Chem 270:21258 – 21263. https://doi.org/10.1074/jbc.270.36.21258.

63. Sattley WM, Madigan MT, Swingley WD, Cheung PC, Clocksin KM, Conrad AL, Dejesa LC, Honchak BM, Jung DO, Karbach LE, Kurdoglu A, Lahiri S, Mastrian SD, Page LE, Taylor HL, Wang ZT, Raymond J, Chen M, Blankenship RE, Touchman JW. 2008. The genome of *Heliobacterium modesticaldum*, a phototrophic representative of the Firmicutes containing the simplest photosynthetic apparatus. J Bacteriol 190: 4687–4696. https://doi.org/10.1128/JB.00299-08.

64. Seo D, Sakurai H. 2002. Purification and characterization of ferredoxinNAD(P)(+) reductase from the green sulfur bacterium *Chlorobium tepidum*. Biochim Biophys Acta 1597:123–132. https://doi.org/10.1016/S0167-4838(02)00269-8.

65. Muraki N, Seo D, Shiba T, Sakurai T, Kurisu G. 2010. Asymmetric dimeric structure of ferredoxin-NAD(P)$^+$ oxidoreductase from the green sulfur bacterium *Chlorobaculum tepidum*: implications for binding ferredoxin and NADP$^+$. J Mol Biol 401:403– 414. https://doi.org/10.1016/j.jmb.2010.06.024.

66. Petering D, Fee JA, Palmer G. 1971. The oxygen sensitivity of spinach ferredoxin and other iron-sulfur proteins. The formation of protein bound sulfur-zero. J Biol Chem 246:643– 653.

67. Knight E, Hardy R. 1966. Isolation and characteristics of flavodoxin from nitrogen-fixing *Clostridium pasteurianum*. J Biol Chem 241:2752–2756.

68. Lovenberg W, Buchanan BB, Rabinowitz JC. 1963. Studies on the chemical nature of clostridial ferredoxin. J Biol Chem 238:3899 –3913.

69. Imlay JA. 2006. Iron-sulphur clusters and the problem with oxygen. Mol Microbiol 59:1073–1082. https://doi.org/10.1111/j.1365-2958.2006.05028.x.

70. Fillat MF, Sandmann G, Gomez-Moreno C. 1988. Flavodoxin from the nitrogen-fixing cyanobacterium *Anabaena* PCC 7119. Arch Microbiol 150:160 –164. https://doi.org/10.1007/BF00425156.

71. Lodeyro AF, Ceccoli RD, Karlusich JJP, Carrillo N. 2012. The importance of flavodoxin for environmental stress tolerance in photosynthetic microorganisms and transgenic plants. Mechanism, evolution and biotechnological potential. FEBS Lett 586:2917–2924. https://doi.org/10.1016/j.febslet.2012.07.026.

72. Bennett LT, Jacobson MR, Dean DR. 1988. Isolation, sequencing, and mutagenesis of the nifF gene encoding flavodoxin from *Azotobacter vinelandii*. J Biol Chem 263:1364 –1369.

73. Hill S, Kavanagh EP. 1980. Roles of *nifF* and *nifJ* gene products in electron transport to nitrogenase in *Klebsiella pneumoniae*. J Bacteriol 141:470 – 475.

74. Shah VK, Stacey G, Brill WJ. 1983. Electron transport to nitrogenase: purification and characterization of pyruvate, flavodoxin oxidoreductase, the *nifJ*-gene product. J Biol Chem 258:2064 –2068.

75. Yoch DC, Arnon DI. 1972. Two biologically active ferredoxins from the aerobic nitrogen-fixing bacterium, *Azotobacter vinelandii*. J Biol Chem 247:4514 – 4520.

76. Jouanneau Y, Meyer C, Naud I, Klipp W. 1995. Characterization of an *fdxN* mutant of *Rhodobacter capsulatus* indicates that ferredoxin-I serves as electron donor to nitrogenase. Biochim Biophys Acta 1232:33– 42. https://doi.org/10.1016/0005-2728(95)00106-X.

77. McRose DL, Zhang XN, Kraepiel AML, Morel FMM. 2017. Diversity and activity of alternative nitrogenases in sequenced genomes and coastal environments. Front Microbiol 8:267. https://doi.org/10.3389/fmicb.2017.00267.

78. Dekas AE, Poretsky RS, Orphan VJ. 2009. Deep-sea Archaea fix and share nitrogen in methane-consuming microbial consortia. Science 326:422– 426. https://doi.org/10.1126/science.1178223.

79. Bolhuis H, Severin I, Confurius-Guns V, Wollenzien UIA, Stal LJ. 2010. Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. ISME J 4:121–130. https://doi.org/10.1038/ismej.2009.99.

80. Armengaud J, Meyer C, Jouanneau Y. 1994. Recombinant expression of the *fdxD* gene of *Rhodobacter capsulatus* and characterization of its product, a [2Fe-2S] ferredoxin. Biochem J 300:413– 418. https://doi.org/10.1042/bj3000413.

81. Schmehl M, Jahn A, Vilsendorf AMZ, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W. 1993. Identification of a new class of nitrogen-fixation genes in *Rhodobacter capsulatus*: a putative membrane complex involved in electron transport to nitrogenase. Mol Gen Genet 241:602– 615. https://doi.org/10.1007/BF00279903.

82. Hamilton TL, Ludwig M, Dixon R, Boyd ES, Dos Santos PC, Setubal JC, Bryant DA, Dean DR, Peters JW. 2011. Transcriptional profiling of nitrogen fixation in *Azotobacter vinelandii*. J Bacteriol 193:4477– 4486. https://doi.org/10.1128/JB.05099-11.

83. Jouanneau Y, Meyer C, Gaillard J, Forest E, Gagnon J. 1993. Purification and characterization of a novel dimeric ferredoxin (FdIII) from *Rhodobacter capsulatus*. J Biol Chem 268:10636 –10644.

84. Moreno-Vivian C, Hennecke S, Pühler A, Klipp W. 1989. Open reading frame 5 (ORF5), encoding a ferredoxin like protein, and *nifQ* are cotranscribed with *nifE*, *nifN*, *nifX*, and ORF4 in *Rhodobacter capsulatus*. J Bacteriol 171:2591–2598. https://doi.org/10.1128/jb.171.5.2591-2598.1989.

85. Masepohl B, Schölisch K, Görlitz K, Kutzki C, Böhme H. 1997. The heterocyst-specific *fdxH* gene product of the cyanobacterium *Anabaena* sp. PCC 7120 is important but not essential for nitrogen fixation. Mol Gen Genet 253:770 –776.

86. Souza ALF, Invitti AL, Rego FGM, Monteiro RA, Klassen G, Souza EM, Chubatsu LS, Pedrosa FO, Rigo LU. 2010. The involvement of the nif-associated ferredoxin-like genes *fdxA* and *fdxN* of *Herbaspirillum seropedicae* in nitrogen fixation. J Microbiol 48:77– 83. https://doi.org/10.1007/s12275-009-0077-y.

87. Hageman RV, Burris RH. 1978. Nitrogenase and nitrogenase reductase associate and dissociate with each catalytic cycle. Proc Natl Acad Sci USA 75:2699 –2702. https://doi.org/10.1073/pnas.75.6.2699.

88. Malkin R, Aparicio PJ, Arnon DI. 1974. The isolation and characterization of a new iron-sulfur protein from photosynthetic membranes. Proc Natl Acad Sci USA 71:2362–2366. https://doi.org/10.1073/pnas.71.6.2362.

89. Hallenbeck PC, Gennaro G. 1998. Stopped-flow kinetic studies of low potential electron carriers of the photosynthetic bacterium, *Rhodobacter capsulatus*: ferredoxin I and NifF. Biochim Biophys Acta 1365:435– 442. https://doi.org/10.1016/S0005-2728(98)00096-6.

90. Hallenbeck PC, Jouanneau Y, Vignais PM. 1982. Purification and molecular properties of a soluble ferredoxin from *Rhodopseudomonas capsulata*. Biochim Biophys Acta 681:168 –176. https://doi.org/10.1016/0005-2728(82)90020-2.

91. Edgren T, Nordlund S. 2005. Electron transport to nitrogenase in *Rhodospirillum rubrum*: identification of a new *fdxN* gene encoding the primary electron donor to nitrogenase. FEMS Microbiol Lett 245:345–351. https://doi.org/10.1016/j.femsle.2005.03.024.

92. Klipp W, Reilander H, Schluter A, Krey R, Puhler A. 1989. The *Rhizobium meliloti fdxN* gene encoding a ferredoxin-like protein is necessary for nitrogen fixation and is cotranscribed with *nifA* and *nifB*. Mol Gen Genet 216:293–302. https://doi.org/10.1007/BF00334368.

93. Armengaud J, Meyer C, Jouanneau Y. 1997. A [2Fe-2S] ferredoxin (FdVI) is essential for growth of the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 179:3304 –3309. https://doi.org/10.1128/jb.179.10.3304-3309.1997.

94. Schrautemeier B, Cassing A, Bohme H. 1994. Characterization of the genome region encoding an FdxH-type ferredoxin and a new 2[4fe-4s] ferredoxin from the nonheterocystous, nitrogen-fixing cyanobacterium *Plectonema boryanum* PCC-73110. J Bacteriol 176:1037–1046. https://doi.org/10.1128/jb.176.4.1037-1046.1994.

95. Ermakova M, Battchikova N, Richaud P, Leino H, Kosourov S, Isojarvi J, Peltier G, Flores E, Cournac L, Allahverdiyeva Y, Aro EM. 2014. Heterocyst-specific flavodiiron protein Flv3B enables oxic diazotrophic growth of the filamentous cyanobacterium

Anabaena sp. PCC 7120. Proc Natl Acad Sci USA 111:11205–11210. https://doi.org/10.1073/pnas.1407327111.

96. Peden EA, Boehm M, Mulder DW, Davis R, Old WM, King PW, Ghirardi ML, Dubini A. 2013. Identification of global ferredoxin interaction networks in *Chlamydomonas reinhardtii*. J Biol Chem 288:35192–35209.https://doi.org/10.1074/jbc.M113.483727.

97. Grabau C, Schatt E, Jouanneau Y, Vignais PM. 1991. A new [2Fe-2S] ferredoxin from *Rhodobacter capsulatus*. Coexpression with a 2 [4Fe-4S] ferredoxin in *Escherichia coli*. J Biol Chem 266:3294 –3299.

98. Deistung J, Cannon FC, Cannon MC, Hill S, Thorneley RNF. 1985. Electron transfer to nitrogenase in *Klebsiella pneumoniae*: *nifF* gene cloned and the gene product, a flavodoxin, purified. Biochem J 231: 743–753. https://doi.org/10.1042/bj2310743.

99. Drummond MH. 1985. The base sequence of the *nifF* gene of *Klebsiella pneumoniae* and homology of the predicted amino acid sequence of its protein product to other flavodoxins. Biochem J 232:891– 896. https://doi.org/10.1042/bj2320891.

100. Roberts GP, MacNeil T, MacNeil D, Brill WJ. 1978. Regulation and characterization of protein products coded by the nif (nitrogen fixation) genes of *Klebsiella pneumoniae*. J Bacteriol 136:267–279.

101. Arnold W, Rump A, Klipp W, Priefer UB, Puhler A. 1988. Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of *Klebsiella pneumoniae*. J Mol Biol 203:715–738. https://doi.org/10.1016/0022-2836(88)90205-7.

102. Peters JW, Schut GJ, Boyd ES, Mulder DW, Shepard EM, Broderick JB, King PW, Adams MWW. 2015. [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochim Biophys Acta 1853:1350 –1369. https://doi.org/10.1016/j.bbamcr.2014.11.021.

103. Poudel S, Tokmina-Lukaszewska M, Colman DR, Refai M, Schut GJ, King PW, Maness PC, Adams MWW, Peters JW, Bothner B, Boyd ES. 2016. Unification of [FeFe]-hydrogenases into three structural and functional groups. Biochim Biophys Acta 1860:1910 –1921. https://doi.org/10.1016/j.bbagen.2016.05.034.

104. Kaminski PA, Norel F, Desnoues N, Kush A, Salzano G, Elmerich C. 1988. Characterization of the *fixABC* region of *Azorhizobium caulinodans ors571* and identification of a new nitrogen fixation gene. Mol Gen Genet 214:496 –502. https://doi.org/10.1007/BF00330486.

105. Gubler M, Hennecke H. 1986. *fixA*, *B* and *C* genes are essential for symbiotic and free-living, microaerobic nitrogen fixation. FEBS Lett 200:186 –192. https://doi.org/10.1016/0014-5793(86)80536-1.

106. Garcia Costas AM, Poudel S, Miller A-F, Schut GJ, Ledbetter RN, Fixen KR, Seefeldt LC, Adams MW, Harwood CS, Boyd ES. 2017. Defining electron bifurcation in the electron transferring flavoprotein family. J Bacteriol 199:e00440-17. https://doi.org/10.1128/JB.00440-17.

107. Furdui C, Ragsdale SW. 2000. The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the Wood-Ljungdahl pathway. J Biol Chem 275:28494 –28499. https://doi.org/10.1074/jbc.M003291200.

108. Neuer G, Bothe H. 1985. Electron donation to nitrogenase in heterocysts of cyanobacteria. Arch Microbiol 143:185–191. https://doi.org/10.1007/BF00411045.

109. Lubitz W, Ogata H, Rudiger O, Reijerse E. 2014. Hydrogenases. Chem Rev 114:4081– 4148. https://doi.org/10.1021/cr4005814.

110. Volbeda A, Charon MH, Piras C, Hatchikian EC, Frey M, Fontecilla-Camps JC. 1995. Crystal structure of the nickel-iron hydrogenase from *Desulfovibrio gigas*. Nature 373:580 –587. https://doi.org/10.1038/373580a0.

111. Shock EL, Boyd ES. 2015. Principles of geobiochemistry. Elements 11:395– 401. https://doi.org/10.2113/gselements.11.6.395.

112. Dick JM, Shock EL. 2011. Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring. PLoS One 6:e22782. https://doi.org/10.1371/journal.pone.0022782.

113. Dick JM. 2014. Average oxidation state of carbon in proteins. J R Soc Interface 11:20131095. https://doi.org/10.1098/rsif.2013.1095.

114. Singh BB, Curdt I, Jakobs C, Schomburg D, Bisen PS, Bohme H. 1999. Identification of amino acids responsible for the oxygen sensitivity of ferredoxins from *Anabaena variabilis* using site-directed mutagenesis. Biochim Biophys Acta 1412:288 – 294. https://doi.org/10.1016/S0005-2728(99)00069-9.

115. Jannasch HW, Mottl MJ. 1985. Geomicrobiology of deep-sea hydrothermal vents. Science 229:717–725. https://doi.org/10.1126/science.229.4715.717.

116. Anantharaman K, Breier JA, Sheik CS, Dick GJ. 2013. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. Proc Natl Acad Sci USA 110:330 –335. https://doi.org/10.1073/pnas.1215340110.

117. Zheng Y, Harris DF, Yu Z, Fu Y, Poudel S, Ledbetter RN, Fixen KR, Yang Z-Y, Boyd ES, Lidstrom ME, Seefeldt LC, Harwood CS. 2018. A pathway for biological methane production using bacterial iron-only nitrogenase. Nat Microbiol 3:281–186. https://doi.org/10.1038/s41564-017-0091-5.

118. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.

119. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033.

120. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu SN, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang ZX, Yamashita RA, Zhang DC, Zheng CJ, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221.

121. Boyd ES, Schut GJ, Adams MWW, Peters JW. 2014. Hydrogen metabolism and the evolution of biological respiration. Microbe 9:361–367.

122. Schut GJ, Zadvornyy O, Wu C-H, Peters JW, Boyd ES, Adams MWW. 2016. The role of geochemistry and energetics in the evolution of modern respiratory complexes from a proton-reducing ancestor. Biochim Biophys Acta 1857:958 –970. https://doi.org/10.1016/j.bbabio.2016.01.010.

123. Hahsler M, Chelluboina S, Hornik K, Buchta C. 2011. The arules R-package ecosystem: analyzing interesting patterns from large transaction data sets. J Mach Learning Res 12:2021–2025.

124. Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules, p 487–499. In Bocca JB, Jarke M, Zaniolo C (ed), Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann, Burlington, MA.

125. R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

126. Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res 8:205–213. https://doi.org/10.1093/dnares/8.5.205.

127. Jacobson BL, Chae YK, Böhme H, Markley JL, Holden HM. 1992. Crystallization and preliminary analysis of oxidized, recombinant, heterocyst [2Fe-2S] ferredoxin from *Anabaena* 7120. Arch Biochem Biophys 294:279 –281. https://doi.org/10.1016/0003-9861(92)90169-W.

128. Rypniewski WR, Breiter DR, Benning MM, Wesenberg G, Oh BH, Markley JL, Rayment I, Holden HM. 1991. Crystallization and structure determination of 2.5 Å resolution of the oxidized [2Fe-2S] ferredoxin isolated from *Anabaena* 7120. Biochemistry 30:4126 – 4131. https://doi.org/10.1021/bi00231a003.

129. Jacobson M, Brigle K, Bennett L, Setterquist R, Wilson M, Cash V, Beynon J, Newton W, Dean D. 1989. Physical and genetic map of the major *nif* gene cluster from *Azotobacter vinelandii*. J Bacteriol 171:1017–1027. https://doi.org/10.1128/jb.171.2.1017-1027.1989.

130. Setubal JC, dos Santos P, Goldman BS, Ertesvag H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D, Cromes L, Curatti L, Du ZJ, Godsy E, Goodner B, Hellner-Burris K, Hernandez JA, Houmiel K, Imperial J, Kennedy C, Larson TJ, Latreille P, Ligon LS, Lu J, Maerk M, Miller NM, Norton S, O'Carroll IP, Paulsen I, Raulfs EC, Roemer R, Rosser J, Segura D, Slater S, Stricklin SL, Studholme DJ, Sun J, Viana CJ, Wallin E, Wang BM, Wheeler C, Zhu HJ, Dean DR, Dixon R, Wood D. 2009. Genome sequence of Azotobacter vinelandii, an obligate aerobe specialized to support diverse anaerobic metabolic processes. J Bacteriol 191:4534 – 4545. https://doi.org/10.1128/JB.00504-09.

131. Schatt E, Jouanneau Y, Vignais P. 1989. Molecular cloning and sequence analysis of the structural gene of ferredoxin I from the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 171:6218 –6226. https://doi.org/10.1128/jb.171.11.6218-6226.1989.

132. Reyntjens B, Jollie DR, Stephens PJ, Gao-Sheridan HS, Burgess BK. 1997. Purification and characterization of a *fixABCX*-linked 2[4Fe-4S] ferredoxin from *Azotobacter vinelandii*. J Biol Inorg Chem 2:595– 602. https://doi.org/10.1007/s007750050174.

133. Saeki K, Suetsugu Y, Tokuda K-I, Miyatake Y, Young D, Marrs B, Matsubara H. 1991. Genetic analysis of functional differences among distinct ferredoxins in *Rhodobacter capsulatus*. J Biol Chem 266:12889 –12895.

134. Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, Land ML, Pelletier DA, Beatty JT, Lang AS, Tabita FR, Gibson JL, Hanson TE, Bobst C, Torres JLTY, Peres C, Harrison FH, Gibson J, Harwood CS. 2004. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nat Biotechnol 22:55– 61. https://doi.org/10.1038/nbt923.

CHAPTER FIVE

CONCLUSIONS AND FUTURE RESEARCH DIRECTIONS

In this dissertation, I aimed to further the understanding of the distribution, evolution and function of FBEB using comparative genomics and structural bioinformatics approaches. Chapter 2 focused on the taxonomic and ecological distribution of FBEB enzymes and addresses the question of whether these enzymes were present in the LUCA of Archaea and Bacteria. The taxonomic distribution of FBEB in complete genomes revealed that homologs of FBEB enzymes are largely confined to obligate anaerobes belonging to Archaea and Bacteria but were absent from Eukarya. While homologs of FBEB enzymes were enriched in the genomes of anaerobes, more specifically in the genomes of anaerobic autotrophs (e.g., methanogens and acetogens) that have been widely speculated to have retained their primitive physiology, our phylogenetic analysis and the limited taxonomic distribution of homologs of eleven of the twelve (Hdr2 not evaluated) FBEB enzymes suggested that these enzymes emerged after the divergence of Archaea and Bacteria and therefore, were not a property of the LUCA. This indicates that FBEB is a recently evolved process and one that may have allowed microorganisms to diversify into new ecological niches.

To identify the type of the environment that likely promoted the emergence of FBEB enzymes, we scanned for the homologs of FBEB enzymes in a database comprising 3136 community metagenomes. Homologs of FBEB enzymes were detected in both surface and subsurface metagenomes but were highly enriched in subsurface

metagenomes, including those of hydrothermal vents/springs, groundwater, and subsurface sediments. These environments exhibit the characteristics of being highly reduced and contain low energy, a finding that is consistent with the results of our complete genome analyses where bifurcating enzymes were primarily found in strict anaerobes. Our phylogenetic analysis, conducted using the catalytic subunits of each bifurcating enzyme, revealed that homologs of most bifurcating enzymes (i.e., seven out of ten known unique bifurcating enzymes) emerged in deep subsurface or hydrothermal vents/springs. These results suggest that bifurcating enzymes likely emerged in highly reducing environments with similar physicochemical characteristics. Many of the homologs of FBEB enzymes recovered from these environmental genomes, including those that branch early in phylogenetic reconstructions, exhibit low sequence identities with homologs available in complete genome databases. This suggests the existence of other yet to be discovered FBEB homologs in natural environments that should perhaps be targeted in the future to reveal additional insights about the origin and function of FBEB enzymes.

Our observations suggest that FBEB is widely distributed among organisms inhabiting anoxic habitats, which are energy limited and where enzymatic mechanisms such as FBEB allow organisms to maximize energy yields from redox reactions. While FBEB is critical for strict anaerobes it is interesting that, with the exception of homologs involved in supplying reduced ferredoxin for nitrogen fixation (Fix), none of the other bifurcating enzymes are widely distributed in aerobes. This suggests that other enzymatic mechanisms for balancing pyridine and Fd pools were selected for over FBEB in

oxidized environments. This may be a result of widespread sensitivity of FBEB enzymes to oxygen, as suggested by the fact that the majority of bifurcating enzymes comprise multiple metal and iron sulfur (FeS) clusters that are themselves oftentimes sensitive to oxygen. On the other hand, Fix is primarily found in aerobes or facultative anaerobes. Fix belongs to a larger family of electron transfer flavoprotein (Etf) that have been found to be involved in a variety of metabolisms that include fatty acid and amino acids metabolisms (Chapter 3). To further our understanding of the functional role of Fix, I also examined the distribution of Etf in complete genome sequences (Chapter 3) and investigated the functional role of Fix (Chapter 4).

Both bifurcating and non-bifurcating Etf enzymes were distributed widely in both Archaea and Bacteria, consistent with the essential roles these enzymes have in the metabolism of organisms operating a wide variety of metabolisms. Since, these enzymes were involved in various metabolisms we sought to group these novel unknown homologs into unique functional bins to understand their potential functional role. We therefore carried out phylogenetic analysis to identify these unique groups. Based on phylogeny, Etf homologs (or paralogs, if they are not functionally redundant) were divided into five distinct groups: group 1 (G1) to group 5 (G5). Of these five groups of Etfs, all bifurcating Etfs belonged to group G2. Comparisons of structural models of non-bifurcating (i.e., G1 Etfs) and bifurcating Etfs (i.e., G2 Etfs) using homology modeling approaches revealed that bifurcating flavin replaces the oxidized form of ATP [i.e., Adenosine monophosphate (AMP)] in bifurcating Etfs and this occurred comparatively recently in evolutionary time. In addition, unique residues were identified that bind the

flavin and NADH in bifurcating Etfs compared to non-bifurcating Etfs, suggesting that bifurcating Etfs that included Fix have a unique structural signature.

Fix or Fix-like enzymes were only found in the subgroup D2 of G2 Etfs. While there exist no predicted functional roles for Fix-like enzymes that were detected in non-nitrogen fixing Archaea, Fix identified in bacterial genomes were shown to be associated with genes involved in nitrogen fixation (Ledbetter et al., 2017). Biological nitrogen fixation is carried out by nitrogenases but interestingly, they are oxygen sensitive. However, they are found in organisms living in wide variety of environments, including oxic conditions. Recently, it was shown that Fix supplies reduced Fd to nitrogenase in the strict aerobe *A. vinelandii* (Ledbetter et al., 2017). Fix couples the reduction of Fd or flavodoxin (Fld) and quinone to the oxidation of two NADH (Ledbetter et al., 2017). We therefore scanned for Fix in our complete genome database to document the distribution of the enzyme in nitrogen fixers (diazotrophs) to investigate if Fix is indeed the primary enzyme among aerobic diazotrophs (Chapter 4).

We first scanned for homologs of nitrogenases (Nif) in available complete genomes. Nif homologs were identified in both archaeal and bacterial genomes but were not detected among eukaryal genomes. We identified a total of four isoforms of Nif: NifA, NifB, NifC and NifD. Of the four isoforms, aerobes and facultative anaerobes primarily encoded NifA nitrogenases while the other three isoforms were primarily found in the genomes of facultative to obligate anaerobes. While anaerobes that inhabit highly reduced environments encoded enzyme homologs that can generate reduced Fd other than Fix [i.e., pyruvate-flavodoxin oxidoreductase (Wahl et al., 1987), *Rhodobacter*

nitrogen fixation protein (Schmehl et al., 1993), [FeFe]-hydrogenase (Therien et al., 2017) and [NiFe]-hydrogenase (Gutekunst et al., 2014)], aerobes primarily encoded Fix suggesting that it is likely the primary enzyme involved in supplying reduced Fd or Fld for nitrogen fixation. Statistical analyses indicate that the reduced Fd or Fld generated by Fix is the primary electron donor to nitrogenase in aerobes. One potential reason for the near obligate requirement of Fd/Fld reduction by Fix in aerobes and not anaerobes is due to the oxidation-reduction potential of the environments where these respective metabolic guilds live. Whereas anaerobes have little difficulty in generating low potential Fd or Fld as electron carriers during metabolism, aerobes typically rely on NADH or quinone for electron transport but these are not of enough low potential to drive nitrogen fixation (electrons need to be ~ 470 mV) (Braaksma et al, 1982; Lanzilotta et al, 1995; Watt et al, 1980). Thus, aerobic diazotrophs that diversified into oxidized environments (aerobes or facultative anaerobes) needed to evolve a mechanism to generate electron carriers with low enough potential (i.e., Fd/Fld) to donate to nitrogenase. We speculate that this is the primary trigger for the emergence and maintenance of Fix in aerobic diazotrophs.

The presence of bifurcating Fix that can reduce Fd/Fld amongst obligate aerobes is a strong indication that FBEB likely emerged or evolved specifically to control the balance of oxidized and reduced Fd/Fld pools. Since anaerobes primarily rely on Fd/Fld as their electron carriers, it is only logical that FBEB is primarily found in anaerobes rather than in aerobes. That being said, bifurcating enzymes can be found in aerobes (e.g., Fix), but likely only if they also carry out anaerobic processes such as nitrogen fixation. It is important to note that not all obligate anaerobes encode bifurcating enzymes, raising

the question of what dictated the acquisition of these enzymes in early evolving lineages. Further work is therefore needed to understand the distribution and acquisition of FBEB in early evolving cells.

Overall, FBEB is an important process with potential industrial applications including in improving the efficiency of production of H_2, ethanol, and methanol in anaerobic cells. The work presented here documents the overall taxonomic and ecological distribution of FBEB enzyme homologs, identified unique homologs of FBEB enzymes in genomic and metagenomic datasets, and examined its importance for key metabolic processes including biological nitrogen fixation. Our work strongly indicates that FBEB emerged early to maintain the balance of oxidized and reduced Fd in anaerobes and this seems to be its main function in contemporary cells. However, FBEB enzymes have diversified over time to function within other metabolic roles as well, including in carbon metabolism in anaerobes. Fd is considered to be the simplest metal-sulfide containing proteins and is widely believed to be the progenitor of many complex metalloenzymes including nitrogenases (Kim *et al*, 2012). For this reason, it is believed that Fd was a property of the LUCA. Since FBEB is involved in balancing the oxidation-reduction state of Fd in modern organisms, it has been hypothesized that it played this role on early earth as well and was thus a property of the LUCA. However, our work clearly indicates that FBEB was not a property of LUCA but rather emerged after the divergence of Archaea and Bacteria from LUCA. This leaves unsolved the puzzle of how early evolving microorganisms reduced Fd. Through additional studies of the distribution of the

environment types and microorganisms that inhabit these environments and that are dependent on FBEB, new insights will be obtained that will help unravel this mystery.

REFERENCES

Braaksma A, Haaker H, Grande HJ, Veeger C. 1982. The effect of the redox potential on the activity of the nitrogenase and on the Fe protein of *Azotobacter vinelandii*. FEBS J 121:483– 491.

Gutekunst K, Chen X, Schreiber K, Kaspar U, Makam S, Appel J. 2014. The bidirectional NiFe-hydrogenase in *Synechocystis* sp PCC 6803 is reduced by flavodoxin and ferredoxin and is essential under mixotrophic, nitrate-limiting conditions. J Biol Chem 289:1930-1937.

Lanzilotta WN, Ryle MJ, Seefeldt LC. 1995. Nucleotide hydrolysis and protein conformational changes in *Azotobacter vinelandii* nitrogenase iron protein: defining the function of aspartate 129. Biochemistry 34: 10713–10723. https://doi.org/10.1021/bi00034a003.

Ledbetter R, Garcia-Costas AM, Lubner CE, Mulder DW, Tokmina-Lukaszewska M, Artz JH et al (2017). The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: Generation of low-potential reducing equivalents for nitrogenase catalysis. Biochemistry 56: 4177-4190.

Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG (2012). Energetic selection of topology in ferredoxins. PLoS computational biology 8(4):e1002463.

Schmehl M, Jahn A, Vilsendorf AMZ, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W. 1993. Identification of a new class of nitrogen-fixation genes in *Rhodobacter capsulatus* - A putative membrane complex involved in electron transport to nitrogenase. Mol Gen Genet 241:602-615

Therien J, Artz JH, Poudel S, Hamilton TL, Liu Z, Noone S, Adams M, King PW, Bryant DA, Boyd ES, Peters JW. 2017. The physiological functions and structural determinants of catalytic bias in the [FeFe]-hydrogenases CpI and CpII of *Clostridium pasteurianum* strain W5. Front Microbiol 8:1305.

Wahl RC, Orme-Johnson WH. 1987. Clostridial pyruvate oxidoreductase and the pyruvate-oxidizing enzyme specific to nitrogen fixation in *Klebsiella pneumoniae* are similar enzymes. J Biol Chem 262:10489-10496.

Watt G, Burns A, Lough S, Tennent D. 1980. Redox and spectroscopic properties of oxidized MoFe protein from *Azotobacter vinelandii*. Biochemistry 19:4926 – 4932. https://doi.org/10.1021/bi00562a035.

CUMULATIVE REFERENCES

Agrawal R, Srikant R. 1994. Fast algorithms for mining association rules, p 487– 499. In Bocca JB, Jarke M, Zaniolo C (ed), Proceedings of the 20th International Conference on Very Large Data Bases. Morgan Kaufmann, Burlington, MA.

Aliverti A, Faber R, Finnerty CM, Ferioli C, Pandini V, Negri A, Karplus PA, Zanetti G. 2001. Biochemical and crystallographic characterization of ferredoxin-NADP(+) reductase from nonphotosynthetic tissues. Biochemistry 40:14501–14508. https://doi.org/10.1021/bi011224c.

Amenabar, M.J., Urschel, M.R., and Boyd, E.S. 2015. "Metabolic and taxonomic diversification in continental magmatic hydrothermal systems," in Microbial Evolution Under Extreme Conditions. Berlin: De Gruyter.), 57-96.

Amend, J.P., and Shock, E.L. 2001. Energetics of overall metabolic reactions of thermophilic and hyperthermophilic Archaea and Bacteria. FEMS Microbiol. Rev. 25, 175-243.

Anantharaman K, Breier JA, Sheik CS, Dick GJ. 2013. Evidence for hydrogen oxidation and metabolic plasticity in widespread deep-sea sulfur-oxidizing bacteria. Proc Natl Acad Sci USA 110:330 –335. https://doi.org/10.1073/pnas.1215340110.

Anraku Y. 1988. Bacterial electron-transport chains. Annu Rev Biochem 57:101–132. https://doi.org/10.1146/annurev.bi.57.070188.000533.

Apte S, Rowell P, Stewart W. 1978. Electron donation to ferredoxin in heterocysts of the N_2-fixing alga *Anabaena cylindrica*. Proc R Soc B Biol Sci 200:1–25. https://doi.org/10.1098/rspb.1978.0001.

Armengaud J, Meyer C, Jouanneau Y. 1994. Recombinant expression of the *fdxD* gene of *Rhodobacter capsulatus* and characterization of its product, a [2Fe-2S] ferredoxin. Biochem J 300:413– 418. https://doi.org/10.1042/bj3000413.

Armengaud J, Meyer C, Jouanneau Y. 1997. A [2Fe-2S] ferredoxin (FdVI) is essential for growth of the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 179:3304 –3309. https://doi.org/10.1128/jb.179.10.3304-3309.1997.

Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195–201. https://doi.org/10.1093/bioinformatics/bti770.

Arnold W, Rump A, Klipp W, Priefer UB, Puhler A. 1988. Nucleotide sequence of a 24,206-base-pair DNA fragment carrying the entire nitrogen fixation gene cluster of *Klebsiella pneumoniae*. J Mol Biol 203:715–738. https://doi.org/10.1016/0022-2836(88)90205-7.

Barkay, T., Kritee, K., Boyd, E., and Geesey, G. 2010. A thermophilic bacterial origin and subsequent constraints by redox, light and salinity on the evolution of the microbial mercuric reductase. Environ. Microbiol. 12, 2904-2917.

Baymann, F., Schoepp-Cothenet, B., Duval, S., Guiral, M., Brugna, M., Baffert, C., Russell, M.J., and Nitschke, W. 2018. On the natural history of flavin-based electron bifurcation. Frontiers in Microbiology.

Benkert P, Biasini M, Schwede T. 2011. Toward the estimation of the absolute quality of individual protein structure models. Bioinformatics 27:343–350. https://doi.org/10.1093/bioinformatics/btq662.

Benkert P, Tosatto SC, Schomburg D. 2008. QMEAN: a comprehensive scoring function for model quality assessment. Proteins 71:261–277. https://doi.org/10.1002/prot.21715.

Bennett LT, Jacobson MR, Dean DR. 1988. Isolation, sequencing, and mutagenesis of the nifF gene encoding flavodoxin from *Azotobacter vinelandii*. J Biol Chem 263:1364 – 1369.

Bergström J, Eady RR, Thorneley RN. 1988. The vanadium- and molybdenum-containing nitrogenases of *Azotobacter chroococcum*. Comparison of mid-point potentials and kinetics of reduction by sodium dithionite of the iron proteins with bound magnesium adenosine 5'-diphosphate. Biochem J 251:165–169.

Berry L, Poudel S, Tokmina-Lukaszewska M, Colman DR, Nguyen DM, Schut GJ et al. 2017. H/D exchange mass spectrometry and statistical coupling analysis reveal a role for allostery in a ferredoxin-dependent bifurcating transhydrogenase catalytic cycle. Biochimica et Biophysica Acta (BBA)-General Subjects 1862: 9-17.

Bertsch J, Parthasarathy A, Buckel W, Muller V. 2013. An electron bifurcating caffeyl-CoA reductase. J Biol Chem 288: 11304-11311.

Bertsch, J., Oppinger, C., Hess, V., Langer, J.D., and Muller, V. 2015. Heterotrimeric NADH-oxidizing methylenetetrahydrofolate reductase from the acetogenic bacterium *Acetobacterium woodii*. J. Bacteriol. 197, 1681-1689.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, Schwede T. 2014. SWISSMODEL:modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258. https://doi.org/ 10.1093/nar/gku340.

Biegel E, Schmidt S, Gonzalez JM, Muller V. 2011. Biochemistry, evolution and physiological function of the Rnf complex, a novel ion-motive electron transport complex in prokaryotes. Cell Mol Life Sci 68:613– 634. https://doi.org/10.1007/s00018-010-0555-8.

Bogaardt, C. 2014. FigTree [Online]. Available: http://tree.bio.ed.ac.uk/software/figtree/ [Accessed].

Bolhuis H, Severin I, Confurius-Guns V, Wollenzien UIA, Stal LJ. 2010. Horizontal transfer of the nitrogen fixation gene cluster in the cyanobacterium *Microcoleus chthonoplastes*. ISME J 4:121–130. https://doi.org/10.1038/ismej.2009.99.

Bothe H, Schmitz O, Yates MG, Newton WE. 2010. Nitrogen fixation and hydrogen metabolism in cyanobacteria. Microbiol Mol Biol Rev 74:529 –551. https://doi.org/10.1128/MMBR.00033-10.

Boughanemi, S., Lyonnet, J., Infossi, P., Bauzan, M., Kosta, A., Lignon, S., Giudici-Orticoni, M.T., and Guiral, M. 2016. Microbial oxidative sulfur metabolism: biochemical evidence of the membrane-bound heterodisulfide reductase-like complex of the bacterium Aquifex aeolicus. FEMS Microbiol. Lett. 363.

Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW. 2011. A late methanogen origin for molybdenum-dependent nitrogenase. Geobiology 9:221–232. https://doi.org/10.1111/j.1472-4669.2011.00278.x.

Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW. 2015. Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. J Bacteriol 197:1690 –1699. https://doi.org/10.1128/JB.02611-14.

Boyd ES, Hamilton TL, Peters JW. 2011. An alternative path for the evolution of biological nitrogen fixation. Front Microbiol 2:205. https://doi.org/10.3389/fmicb.2011.00205.

Boyd ES, Peters JW. 2013. New insights into the evolutionary history of biological nitrogen fixation. Front Microbiol 4:201. https://doi.org/10.3389/fmicb.2013.00201.

Boyd ES, Schut GJ, Adams MWW, Peters JW. 2014. Hydrogen metabolism and the evolution of biological respiration. Microbe 9:361–367.

Braaksma A, Haaker H, Grande HJ, Veeger C. 1982. The effect of the redox potential on the activity of the nitrogenase and on the Fe protein of *Azotobacter vinelandii*. FEBS J 121:483– 491.

Buan, N.R., and Metcalf, W.W. 2010. Methanogenesis by *Methanosarcina acetivorans* involves two structurally and functionally distinct classes of heterodisulfide reductase. Mol. Microbiol. 75, 843-853.

Buckel W, Thauer RK. 2013. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. Bba-Bioenergetics 1827: 94-113.

Buckel W, Thauer RK. 2018a. Flavin-based electron bifurcation, ferredoxin, flavodoxin, and anaerobic respiration with protons (Ech) or NAD+ (Rnf) as electron acceptors: A historical review. Front Microbiol 9: 401.

Buckel W, Thauer RK. 2018b. Flavin-Based Electron Bifurcation, A New Mechanism of Biological Energy Coupling. Chem Rev.

Bulen W, LeComte J. 1966. The nitrogenase system from *Azotobacter*: two-enzyme requirement for N_2 reduction, ATP-dependent H_2 evolution, and ATP hydrolysis. Proc Natl Acad SciUSA 56:979 –986. https://doi.org/10.1073/pnas.56.3.979.

Burk D, Lineweaver H, Horner CK. 1934. The specific influence of acidity on the mechanism of nitrogen fixation by *Azotobacter*. J Bacteriol 27:325–340.

Calusinska, M., Happe, T., Joris, B., and Wilmotte, A. 2010. The surprising diversity of clostridial hydrogenases: a comparative genomic perspective. Microbiol. 156, 1575-1588.

Chen D, Swenson RP. 1994. Cloning, sequence analysis, and expression of the genes encoding the two subunits of the methylotrophic bacterium W3A1 electron transfer flavoprotein. J Biol Chem 269:32120 –32130.

Chowdhury NP, Kahnt J, Buckel W. 2015. Reduction of ferredoxin or oxygen by flavin-based electron bifurcation in *Megasphaera elsdenii*. FEBS J 282:3149 –3160. https://doi.org/10.1111/febs.13308.

Chowdhury, N.P., Mowafy, A.M., Demmer, J.K., Upadhyay, V., Koelzer, S., Jayamani, E., Kahnt, J., Hornung, M., Demmer, U., Ermler, U., and Buckel, W. 2014. Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J. Biol. Chem .289, 5145-5157.

Colman, D.R., Poudel, S., Stamps, B.W., Boyd, E.S., and Spear, J.R. 2017. The deep, hot biosphere: Twenty-five years of retrospection. Proc. Natl. Acad. Sci. USA 114, 6895-6903.

Costa, K.C., Lie, T.J., Xia, Q., and Leigh, J.A. 2013. VhuD facilitates electron flow from H2 or formate to heterodisulfide reductase in *Methanococcus maripaludis*. J. Bacteriol. 195, 5160-5165.

Costa, K.C., Wong, P.M., Wang, T.S., Lie, T.J., Dodsworth, J.A., Swanson, I., Burn, J.A., Hackett, M., and Leigh, J.A. 2010. Protein complexing in a methanogen suggests electron bifurcation and electron delivery from formate to heterodisulfide reductase. Proc. Natl. Acad. Sci. USA 107, 11050-11055.

Crane F, Beinert H. 1956. On the mechanism of dehydrogenation of fatty acyl derivatives of coenzyme A. J Biol Chem 218: 701-716.

Davidson VL, Husain M, Neher JW. 1986. Electron transfer flavoprotein from *Methylophilus methylotrophus*: properties, comparison with other electron transfer flavoproteins, and regulation of expression by carbon source. J Bacteriol 166:812– 817. https://doi.org/10.1128/jb.166.3.812-817.1986.

Deistung J, Cannon FC, Cannon MC, Hill S, Thorneley RNF. 1985. Electron transfer to nitrogenase in *Klebsiella pneumoniae*: *nifF* gene cloned and the gene product, a flavodoxin, purified. Biochem J 231: 743–753. https://doi.org/10.1042/bj2310743.

Deistung J, Thorneley RNF. 1986. Electron transfer to nitrogenase. Characterization of flavodoxin from Azotobacter chroococcum and comparison of its redoc potentials with those of flavodoxins from *Azotobacter vinelandii* and *Klebsiella pneumoniae* (*nifF*-gene product). Biochemical Journal 239: 69-75.

Dekas AE, Poretsky RS, Orphan VJ. 2009. Deep-sea Archaea fix and share nitrogen in methane-consuming microbial consortia. Science 326:422– 426. https://doi.org/10.1126/science.1178223.

Demmer JK, Chowdhury NP, Selmer T, Ermler U, Buckel W. 2017. The semiquinone swing in the bifurcating electron transferring flavoprotein/butyryl-CoA dehydrogenase complex from *Clostridium difficile*. Nat Commun 8.

Demmer JK, Huang H, Wang SN, Demmer U, Thauer RK, Ermler U. 2015. Insights into Flavin-based Electron Bifurcation via the NADH-dependent reduced Ferredoxin:NADP oxidoreductase structure. J Biol Chem 290: 21985-21995.

D'hondt, S., Inagaki, F., Zarikian, C.A., Abrams, L.J., Dubois, N., Engelhardt, T., Evans, H., Ferdelman, T., Gribsholt, B., Harris, R.N., Hoppie, B.W., Hyun, J.H., Kallmeyer, J., Kim, J., Lynch, J.E., Mckinley, C.C., Mitsunobu, S., Morono, Y., Murray, R.W., Pockalny, R., Sauvage, J., Shimono, T., Shiraishi, F., Smith, D.C., Smith-Duque, C.E., Spivack, A.J., Steinsbu, B.O., Suzuki, Y., Szpak, M., Toffin, L., Uramoto, G., Yamaguchi, Y.T., Zhang, G.L., Zhang, X.H., and Ziebis, W. 2015. Presence of oxygen and aerobic communities from sea floor to basement in deep-sea sediments. Nat. Geosci. 8, 299-304.

Dick JM, Shock EL. 2011. Calculation of the relative chemical stabilities of proteins as a function of temperature and redox chemistry in a hot spring. PLoS One 6:e22782. https://doi.org/10.1371/journal.pone.0022782.

Dick JM. 2014. Average oxidation state of carbon in proteins. J R Soc Interface 11:20131095. https://doi.org/10.1098/rsif.2013.1095.

Dilworth MJ, Eldridge ME, Eady RR. 1992. Correction for creatine interference with the direct indophenol measurement of NH_3 in steadystate nitrogenase assays. Anal Biochem 207:6 –10. https://doi.org/10.1016/0003-2697(92)90491-O.

Dilworth MJ, Eldridge ME, Eady RR. 1993. The molybdenum and vanadium nitrogenases of *Azotobacter chroococcum*: effect of elevated temperature on N_2 reduction. Biochem J 289:395– 400. https://doi.org/10.1042/bj2890395.

Djordjevic, S., Pace, C.P., Stankovich, M.T., and Kim, J.-J.P. 1995. Three-dimensional structure of butyryl-CoA dehydrogenase from *Megasphaera elsdenii*. Biochem. 34, 2163-2171.

Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. 2012. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13:162. https://doi.org/10.1186/1471-2164-13-162.

Drummond MH. 1985. The base sequence of the *nifF* gene of *Klebsiella pneumoniae* and homology of the predicted amino acid sequence of its protein product to other flavodoxins. Biochem J 232:891– 896. https://doi.org/10.1042/bj2320891.

DuPlessis ER, Rohlfs RJ, Hille R, Thorpe C. 1994. Electron-transferring flavoprotein from pig and the methylotrophic bacterium W3A1 contains AMP as well as FAD. Biochem Mol Biol Int 32:195–199.

Dutton PL. 1986. Energy transduction in anoxygenic photosynthesis, p197–237. In Staehelin A, Arntzen CJ (ed), Photosynthesis III. Photosynthetic membranes and light harvesting systems. Encyclopedia of Plant Physiology (new series), vol 19. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-70936-4_5.

E, Kahnt J, Hornung M, Demmer U, Ermler U, Buckel W. 2014. Studies on the mechanism of electron bifurcation catalyzed by electron transferring flavoprotein (Etf) and butyryl-CoA dehydrogenase (Bcd) of *Acidaminococcus fermentans*. J Biol Chem 289:5145–5157. https://doi.org/10.1074/jbc.M113.521013.

Eady RR. 1996. Structure-function relationships of alternative nitrogenases. Chem Rev 96:3013–3030. https://doi.org/10.1021/cr950057h.

Earl C, Ronson C, Ausubel F. 1987. Genetic and structural analysis of the *Rhizobium meliloti fixA*, *fixB*, *fixC*, and *fixX* genes. J Bacteriol 169:1127–1136. https://doi.org/10.1128/jb.169.3.1127-1136.1987.

Eddy, S. 2015. HMMER: biosequence analysis using profile hidden Markov models [Online]. Available: hmmer.org [Accessed].

Edgren T, Nordlund S. 2005. Electron transport to nitrogenase in *Rhodospirillum rubrum*: identification of a new *fdxN* gene encoding the primary electron donor to nitrogenase. FEMS Microbiol Lett 245:345–351. https://doi.org/10.1016/j.femsle.2005.03.024.

Edgren, T., and Nordlund, S. 2004. The *fixABCX* genes in Rhodospirillum rubrum encode a putative membrane complex participating in electron transfer to nitrogenase. J. Bacteriol. 186, 2052-2060.

Edwards, K.J., Becker, K., and Colwell, F. 2012. The deep, dark energy biosphere: Intraterrestrial life on Earth. Annu. Rev. Earth. Plan. Sci. 40, 551-568.

Egener T, Martin DE, Sarkar A, Reinhold-Hurek B. 2001. Role of a ferredoxin gene cotranscribed with the *nifHDK* operon in N_2 fixation and nitrogenase "switch-off" of *Azoarcus* sp. strain BH72. J Bacteriol 183:3752–3760. https://doi.org/10.1128/JB.183.12.3752-3760.2001.

Eichler K, Buchet A, Bourgis F, Kleber HP, Mandrand-Berthelot MA. 1995.The *fix Escherichia coli* region contains four genes related to carnitine metabolism. J Basic Microbiol 35:217–227. https://doi.org/10.1002/jobm.3620350404.

Ermakova M, Battchikova N, Richaud P, Leino H, Kosourov S, Isojarvi J, Peltier G, Flores E, Cournac L, Allahverdiyeva Y, Aro EM. 2014. Heterocyst-specific flavodiiron protein Flv3B enables oxic diazotrophic growth of the filamentous cyanobacterium *Anabaena* sp. PCC 7120. Proc Natl Acad Sci USA 111:11205–11210. https://doi.org/10.1073/pnas.1407327111.

Estelmann S, Boll M. 2014. Glutaryl-coenzyme A dehydrogenase from *Geobacter metallireducens*: interaction with electron transferring flavoprotein and kinetic basis of unidirectional catalysis. FEBS J 281: 5120 –5131. https://doi.org/10.1111/febs.13051.

Evans, P.N., Parks, D.H., Chadwick, G.L., Robbins, S.J., Orphan, V.J., Golding, S.D., and Tyson, G.W. 2015. Methane metabolism in the archaeal phylum Bathyarchaeota revealed by genome-centric metagenomics. Science 350, 434-438.

Falkowski PG. 1997. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO_2 in the ocean. Nature 387:272–275. https://doi.org/10.1038/387272a0.

Fay P. 1992. Oxygen relations of nitrogen fixation in cyanobacteria. Microbiol Rev 56:340 –373.

Ferry, J.G. 1990. Formate dehydrogenase. FEMS Microbiol. Lett. 87, 377-382.

Fillat MF, Sandmann G, Gomez-Moreno C. 1988. Flavodoxin from the nitrogen-fixing cyanobacterium *Anabaena* PCC 7119. Arch Microbiol 150:160 –164. https://doi.org/10.1007/BF00425156.

Forterre, P. 2015. The universal tree of life: an update. Front Microbiol 6. 717.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150-3152.

Furdui C, Ragsdale SW. 2000. The role of pyruvate ferredoxin oxidoreductase in pyruvate synthesis during autotrophic growth by the Wood-Ljungdahl pathway. J Biol Chem 275:28494 –28499. https://doi.org/10.1074/jbc.M003291200.

Garcia Costas, A.M., Poudel, S., Miller, A.-F., Schut, G.J., Ledbetter, R.N., Fixen, K.R., Seefeldt, L.C., Adams, M.W., Harwood, C.S., and Boyd, E.S. 2017. Defining electron bifurcation in the electron transferring flavoprotein family. J. Bacteriol., e00440-17.

Garvie, E.I. 1980. Bacterial lactate dehydrogenases. Microbiol. Rev. 44, 106-139.

Gennaro G, Hubner P, Sandmeier U, Yakunin AF, Hallenbeck PC. 1996. Cloning, characterization, and regulation of *nifF* from *Rhodobacter capsulatus*. J Bacteriol 178:3949 –3952. https://doi.org/10.1128/jb.178.13.3949-3952.1996.

Ghisla S, Thorpe C. 2004. Acyl-CoA dehydrogenases: a mechanistic overview. Eur J Biochem 271:494 –508. https://doi.org/10.1046/j.1432-1033.2003.03946.x.

Grabau C, Schatt E, Jouanneau Y, Vignais PM. 1991. A new [2Fe-2S] ferredoxin from *Rhodobacter capsulatus*. Coexpression with a 2 [4Fe-4S] ferredoxin in *Escherichia coli*. J Biol Chem 266:3294 –3299.

Gray HB, Winkler JR. 2015. Hole hopping through tyrosine/tryptophan chains protects proteins from oxidative damage. Proc Natl Acad SciUSA 112:10920 –10925. https://doi.org/10.1073/pnas.1512704112.

Greening, C., Biswas, A., Carere, C.R., Jackson, C.J., Taylor, M.C., Stott, M.B., Cook, G.M., and Morales, S.E. 2016. Genomic and metagenomic surveys of hydrogenase distribution indicate H2 is a widely utilised energy source for microbial growth and survival. ISME J. 10, 761.

Griffith, D.M., Veech, J.A., and Marsh, C.J. 2016. cooccur: Probabilistic species co-occurrence analysis in R. J. Stat. Software. 69, 1-17.

Grönger P, Manian SS, Reiländer H, O'Connell M, Priefer UB, Pühler A. 1987. Organization and partial sequence of a DNA region of the *Rhizobium leguminosarum* symbiotic plasmid pRL6JI containing the genes *fixABC*, *nifA*, *nifB* and a novel open reading frame. Nucleic Acids Res15:31– 49. https://doi.org/10.1093/nar/15.1.31.

Gubler M, Hennecke H. 1986. *fixA*, *B* and *C* genes are essential for symbiotic and free-living, microaerobic nitrogen fixation. FEBS Lett 200:186 –192. https://doi.org/10.1016/0014-5793(86)80536-1.

Gubler M, Zürcher T, Hennecke H. 1989. The *Bradyrhizobium japonicum fixBCX* operon: identification of *fixX* and of a 5'mRNA region affecting the level of the *fixSCX* transcript. Mol Microbiol 3:141–148. https://doi.org/10.1111/j.1365-958.1989.tb01803.x.

Guenther, B.D., Sheppard, C.A., Tran, P., Rozen, R., Matthews, R.G., and Ludwig, M.L. 1999. The structure and properties of methylenetetrahydrofolate reductase from *Escherichia coli* suggest how folate ameliorates human hyperhomocysteinemia. Nat. Struct. Biol. 6, 359-365.

Gutekunst K, Chen X, Schreiber K, Kaspar U, Makam S, Appel J. 2014. The bidirectional NiFe-hydrogenase in *Synechocystis* sp PCC 6803 is reduced by flavodoxin and ferredoxin and is essential under mixotrophic, nitrate-limiting conditions. J Biol Chem 289:1930 –1937. https://doi.org/10.1074/jbc.M113.526376.

Hageman RV, Burris RH. 1978. Nitrogenase and nitrogenase reductase associate and dissociate with each catalytic cycle. Proc Natl Acad Sci USA 75:2699 –2702. https://doi.org/10.1073/pnas.75.6.2699.

Hahsler M, Chelluboina S, Hornik K, Buchta C. 2011. The arules R-package ecosystem: analyzing interesting patterns from large transaction data sets. J Mach Learning Res 12:2021–2025.

Hallenbeck PC, Gennaro G. 1998. Stopped-flow kinetic studies of low potential electron carriers of the photosynthetic bacterium, *Rhodobacter capsulatus*: ferredoxin I and NifF. Biochim Biophys Acta 1365:435– 442. https://doi.org/10.1016/S0005-2728(98)00096-6.

Hallenbeck PC, Jouanneau Y, Vignais PM. 1982. Purification and molecular properties of a soluble ferredoxin from *Rhodopseudomonas capsulata*. Biochim Biophys Acta 681:168 –176. https://doi.org/10.1016/0005-2728(82)90020-2.

Hamilton TL, Ludwig M, Dixon R, Boyd ES, Dos Santos PC, Setubal JC, Bryant DA, Dean DR, Peters JW. 2011. Transcriptional profiling of nitrogen fixation in *Azotobacter vinelandii*. J Bacteriol 193:4477– 4486. https://doi.org/10.1128/JB.05099-11.

Hedderich, R., Koch, J., Linder, D., and Thauer, R.K. 1994. The heterodisulfide reductase from *Methanobacterium hermoautotrophicum* contains sequence motifs characteristic of pyridine-nucleotide-dependent thioredoxin reductases. FEBS J. 225, 253-261.

Herrmann G, Jayamani E, Mai G, Buckel W. 2008. Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. J Bacteriol 190:784 –791. https://doi.org/10.1128/JB.01422-07.

Hess, V., Schuchmann, K., and Müller, V. 2013. The ferredoxin:NAD+ oxidoreductase (Rnf) from the acetogen *Acetobacterium woodii* requires Na+ and is reversibly coupled to the membrane potential. J. Biol. Chem. 288, 31496-31502.

Hill S, Kavanagh EP. 1980. Roles of *nifF* and *nifJ* gene products in electron transport to nitrogenase in *Klebsiella pneumoniae*. J Bacteriol 141:470 – 475.

Hochstrasser M. 2009. Origin and function of ubiquitin-like protein conjugation. Nature 458:422– 429. https://doi.org/10.1038/nature07958.

Huang JJ, Heiniger EK, McKinlay JB, Harwood CS. 2010. Production of hydrogen gas from light and the inorganic electron donor thiosulfate by *Rhodopseudomonas palustris*. Appl Environ Microbiol 76:7717–7722. https://doi.org/10.1128/AEM.01143-10.

Huang L, Rohlfs RJ, Hille R. 1995. The reaction of trimethylamine dehydrogenase with electron transferring flavoprotein. J Biol Chem 270:23958 –23965. https://doi.org/10.1074/jbc.270.41.23958.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D.A., Finstad, K.M., Amundson, R., Thomas, B.C., and Banfield, J.F. 2016. A new view of the tree of life. Nat. Microbiol 1. 16048.

Imlay JA. 2006. Iron-sulphur clusters and the problem with oxygen. Mol Microbiol 59:1073–1082. https://doi.org/10.1111/j.1365-2958.2006.05028.x.

Isas JM, Yannone SM, Burgess BK. 1995. *Azotobacter vinelandii* NADPH-ferredoxin reductase cloning, sequencing, and overexpression. J Biol Chem 270:21258 –21263. https://doi.org/10.1074/jbc.270.36.21258.

Jacobson BL, Chae YK, Böhme H, Markley JL, Holden HM. 1992. Crystallization and preliminary analysis of oxidized, recombinant, heterocyst [2Fe-2S] ferredoxin from *Anabaena* 7120. Arch Biochem Biophys 294:279 –281. https://doi.org/10.1016/0003-9861(92)90169-W.

Jacobson M, Brigle K, Bennett L, Setterquist R, Wilson M, Cash V, Beynon J, Newton W, Dean D. 1989. Physical and genetic map of the major *nif* gene cluster from *Azotobacter vinelandii*. J Bacteriol 171:1017–1027. https://doi.org/10.1128/jb.171.2.1017-1027.1989.

Jagannathan B, Golbeck J (ed). 2009. Photosynthesis: microbial, p325–341. In Schaechter M (ed), Encyclopedia of microbiology, 3rd ed. Academic press, Cambridge, MA. https://doi.org/10.1016/B978-012373944-5.00352-7.

Jang M-H, Scrutton NS, Hille R. 2000. Formation of W3A1 electron transferring flavoprotein (ETF) hydroquinone in the trimethylamine dehydrogenase·ETF protein complex. J Biol Chem 275:12546 –12552.https://doi.org/10.1074/jbc.275.17.12546.

Jannasch HW, Mottl MJ. 1985. Geomicrobiology of deep-sea hydrothermal vents. Science 229:717–725. https://doi.org/10.1126/science.229.4715.717.

Joerger RD, Bishop PE, Bishop PE. 1988. Bacterial alternative nitrogenfixation systems. Crit Rev Microbiol 16:1–14. https://doi.org/10.3109/10408418809104465.

Jouanneau Y, Jeong HS, Hugo N, Meyer C, Willison JC. 1998. Overexpression in *Escherichia coli* of the rnf genes from *Rhodobacter capsulatus*: characterization of two membrane-bound iron-sulfur proteins. Eur J Biochem 251:54 – 64. https://doi.org/10.1046/j.1432-1327.1998.2510054.x.

Jouanneau Y, Meyer C, Gaillard J, Forest E, Gagnon J. 1993. Purification and characterization of a novel dimeric ferredoxin (FdIII) from *Rhodobacter capsulatus*. J Biol Chem 268:10636 –10644.

Jouanneau Y, Meyer C, Naud I, Klipp W. 1995. Characterization of an *fdxN* mutant of *Rhodobacter capsulatus* indicates that ferredoxin-I serves as electron donor to nitrogenase. Biochim Biophys Acta 1232:33– 42. https://doi.org/10.1016/0005-2728(95)00106-X.

Jungermann, K., Thauer, R.K., Leimenstoll, G., and Decker, K. 1973. Function of reduced pyridine nucleotide-ferredoxin oxidoreductases in saccharolytic Clostridia. Biochim. Biophys. Acta 305, 268-280.

Kaminski PA, Norel F, Desnoues N, Kush A, Salzano G, Elmerich C. 1988. Characterization of the *fixABC* region of *Azorhizobium caulinodans* ors571 and identification of a new nitrogen fixation gene. Mol Gen Genet 214: 496-502.

Kaminski PA, Norel F, Desnoues N, Kush A, Salzano G, Elmerich C. 1988. Characterization of the *fixABC* region of *Azorhizobium caulinodans ors571* and identification of a new nitrogen fixation gene. Mol Gen Genet 214:496 –502. https://doi.org/10.1007/BF00330486.

Kaneko T, Nakamura Y, Wolk CP, Kuritz T, Sasamoto S, Watanabe A, Iriguchi M, Ishikawa A, Kawashima K, Kimura T, Kishida Y, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakazaki N, Shimpo S, Sugimoto M, Takazawa M, Yamada M, Yasuda M, Tabata S. 2001. Complete genomic sequence of the filamentous nitrogen-fixing cyanobacterium *Anabaena* sp. strain PCC 7120. DNA Res 8:205–213. https://doi.org/10.1093/dnares/8.5.205.

Kaster, A.K., Moll, J., Parey, K., and Thauer, R.K. 2011. Coupling of ferredoxin and heterodisulfide reduction via electron bifurcation in hydrogenotrophic methanogenic archaea. Proc. Natl. Acad. Sci. USA 108, 2981-2986.

Keltch AK, Strittmatter CF, Walters CP, Clowes GHA. 1950. Oxidative phosphorylation by a cell-free particulate system from unfertilized Arbacia eggs. J Gen Physiol 33: 547-553.

Khanna N, Lindblad P. 2015. Cyanobacterial hydrogenases and hydrogen metabolism revisited: recent progress and future prospects. Int J Mol Sci 16:10537–10561. https://doi.org/10.3390/ijms160510537.

Kieft, T.L. 2016. Microbiology of the deep continental biosphere. Ad.v Env. Microbiol. 1, 225-249.

Kim JD, Rodriguez-Granillo A, Case DA, Nanda V, Falkowski PG. 2012. Energetic selection of topology in ferredoxins. PLoS computational biology 8(4):e1002463.

Klipp W, Reilander H, Schluter A, Krey R, Puhler A. 1989. The *Rhizobium meliloti fdxN* gene encoding a ferredoxin-like protein is necessary for nitrogen fixation and is cotranscribed with *nifA* and *nifB*. Mol Gen Genet 216:293–302. https://doi.org/10.1007/BF00334368.

Knight E, Hardy R. 1966. Isolation and characteristics of flavodoxin from nitrogen-fixing *Clostridium pasteurianum*. J Biol Chem 241:2752–2756.

Kumar, S., Stecher, G., and Tamura, K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870-1874.

Lanzilotta WN, Ryle MJ, Seefeldt LC. 1995. Nucleotide hydrolysis and protein conformational changes in *Azotobacter vinelandii* nitrogenase iron protein: defining the function of aspartate 129. Biochemistry 34: 10713–10723. https://doi.org/10.1021/bi00034a003.

Larimer FW, Chain P, Hauser L, Lamerdin J, Malfatti S, Do L, Land ML, Pelletier DA, Beatty JT, Lang AS, Tabita FR, Gibson JL, Hanson TE, Bobst C, Torres JLTY, Peres C, Harrison FH, Gibson J, Harwood CS. 2004. Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodopseudomonas palustris*. Nat Biotechnol 22:55– 61. https://doi.org/10.1038/nbt923.

Laso-Perez, R., Wegener, G., Knittel, K., Widdel, F., Harding, K.J., Krukenberg, V., Meier, D.V., Richter, M., Tegetmeyer, H.E., Riedel, D., Richnow, H.H., Adrian, L., Reemtsma, T., Lechtenfeld, O.J., and Musat, F. 2016. Thermophilic archaea activate butane via alkyl-coenzyme M formation. Nature 539, 396-401.

Ledbetter R, Garcia-Costas AM, Lubner CE, Mulder DW, Tokmina-Lukaszewska M, Artz JH et al. 2017. The electron bifurcating FixABCX protein complex from *Azotobacter vinelandii*: Generation of low-potential reducing equivalents for nitrogenase catalysis. Biochemistry 56: 4177-4190.

Lee PT, Hsu AY, Ha HT, Clarke CF. 1997. A C-methyltransferase involved in both ubiquinone and menaquinone biosynthesis: isolation and identification of the *Escherichia coli ubiE* gene. J Bacteriol 179:1748 –1754. https://doi.org/10.1128/jb.179.5.1748-1754.1997.

Leigh JA. 2000. Nitrogen fixation in methanogens: the archaeal perspective. Curr Issues Mol Biol 2:125–131.

Letunic, I., and Bork, P. 2016. Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 44, W242-W245.

Li F, Hinderberger J, Seedorf H, Zhang J, Buckel W, Thauer RK. 2008. Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. J Bacteriol 190: 843-850.

Li, W., and Godzik, A. 2006. Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22, 1658-1659.

Lindsay, M.R., Anderson, C., Fox, N., Scofield, G., Allen, J., Anderson, E., Bueter, L., Poudel, S., Sutherland, K., Munson-Mcgee, J.H., Van Nostrand, J.D., Zhou, J., Spear, J.R., Baxter, B.K., Lageson, D.R., and Boyd, E.S. 2017. Microbialite response to an anthropogenic salinity gradient in Great Salt Lake, Utah. Geobiology 15, 131-145.

Lipmann F. 1941. Metabolic generation and utilization of phosphate bond energy. Adv Enzymol Rel S Bi 1: 99-162.

Lodeyro AF, Ceccoli RD, Karlusich JJP, Carrillo N. 2012. The importance of flavodoxin for environmental stress tolerance in photosynthetic microorganisms and transgenic plants. Mechanism, evolution and biotechnological potential. FEBS Lett 586:2917–2924. https://doi.org/10.1016/j.febslet.2012.07.026.

Lovenberg W, Buchanan BB, Rabinowitz JC. 1963. Studies on the chemical nature of clostridial ferredoxin. J Biol Chem 238:3899 –3913.

Lubitz W, Ogata H, Rudiger O, Reijerse E. 2014. Hydrogenases. Chem Rev 114:4081–4148. https://doi.org/10.1021/cr4005814.

Lubner CE, Jennings DP, Mulder DW, Schut GJ, Zadvornyy OA, Hoben JP, Tokmina-Lukaszewska M, Berry L, Nguyen DM, Lipscomb GL, Bothner B, Jones AK, Miller AF, King PW, Adams MWW, Peters JW. 2017. Mechanistic insights into energy conservation by flavin-based electron bifurcation. Nat Chem Biol 13:655– 659. https://doi.org/10.1038/nchembio.2348.

Luo H, Lin Y, Gao F, Zhang C-T, Zhang R. 2014. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements. Nucleic Acids Res 42:D574 –D580. https://doi.org/10.1093/nar/gkt1131.

Lyons, T.W., Reinhard, C.T., and Planavsky, N.J. 2014. The rise of oxygen in Earth's early ocean and atmosphere. Nature 506, 307-315.

Makarova KS, Koonin EV. 2010. Archaeal ubiquitin-like proteins: functional versatility and putative ancestral involvement in tRNA modification revealed by comparative genomic analysis. Archaea 2010:710303. https://doi.org/10.1155/2010/710303.

Malkin R, Aparicio PJ, Arnon DI. 1974. The isolation and characterization of a new iron-sulfur protein from photosynthetic membranes. Proc Natl Acad Sci USA 71:2362–2366. https://doi.org/10.1073/pnas.71.6.2362.

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu SN, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang ZX, Yamashita RA, Zhang DC, Zheng CJ, Bryant SH. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res 43:D222–D226. https://doi.org/10.1093/nar/gku1221.

Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the Integrated Microbial Genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122. https://doi.org/10.1093/nar/gkr1044.

Martin AE, Burgess BK, Iismaa SE, Smartt CT, Jacobson MR, Dean DR. 1989. Construction and characterization of an *Azotobacter vinelandii* strain with mutations in the genes encoding flavodoxin and ferredoxin I. J Bacteriol 171:3162–3167. https://doi.org/10.1128/jb.171.6.3162-3167.1989.

Martin WF. 2012. Hydrogen, metals, bifurcating electrons, and proton gradients: The early evolution of biological energy conservation. Febs Lett 586: 485-493.

Martin, W., Baross, J., Kelley, D., and Russell, M.J. 2008. Hydrothermal vents and the origin of life. Nat. Rev. Microbiol. 6, 805-814.

Martin, W.F., Bryant, D.A., and Beatty, J.T. 2017. A physiological perspective on the origin and evolution of photosynthesis. FEMS Microbiol. Rev.

Masepohl B, Schölisch K, Görlitz K, Kutzki C, Böhme H. 1997. The heterocyst-specific *fdxH* gene product of the cyanobacterium *Anabaena* sp. PCC 7120 is important but not essential for nitrogen fixation. Mol Gen Genet 253:770 –776.

McGlynn SE, Boyd ES, Peters JW, Orphan VJ. 2013. Classifying the metal dependence of uncharacterized nitrogenases. Front Microbiol 3:419. https://doi.org/10.3389/fmicb.2012.00419.

McRose DL, Zhang XN, Kraepiel AML, Morel FMM. 2017. Diversity and activity of alternative nitrogenases in sequenced genomes and coastal environments. Front Microbiol 8:267. https://doi.org/10.3389/fmicb.2017.00267.

Mehta MP, Baross JA. 2006. Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. Science 314:1783–1786. https://doi.org/10.1126/science.1134772.

Meyer, J. 2007. [FeFe] hydrogenases and their evolution: a genomic perspective. Cell Mol. Life Sci. 64, 1063-1084.

Mitchell P 1961. Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. Nature 191: 144-148.

Mitchell P 1975. The protonmotive Q cycle: a general formulation. Febs Lett 59: 137-139.

Mock, J., Wang, S.N., Huang, H.Y., Kahnt, J., and Thauer, R.K. 2014. Evidence for a gexaheteromeric methylenetetrahydrofolate reductase in *Moorella thermoacetica*. J. Bacteriol. 196, 3303-3314.

Moreno-Vivian C, Hennecke S, Pühler A, Klipp W. 1989. Open reading frame 5 (ORF5), encoding a ferredoxin like protein, and *nifQ* are cotranscribed with *nifE*, *nifN*, *nifX*, and ORF4 in *Rhodobacter capsulatus*. J Bacteriol 171:2591–2598. https://doi.org/10.1128/jb.171.5.2591-2598.1989.

Mulder, D.W., Boyd, E.S., Sarma, R., Lange, R.K., Endrizzi, J.A., Broderick, J.B., and Peters, J.W. 2010. Stepwise [FeFe]-hydrogenase H-cluster assembly revealed in the structure of HydAΔEFG. Nature 465, 248-U143.

Müller, V., Chowdhury, N.P., and Basen, M. 2018. Electron bifurcation: A long-hidden energy-coupling mechanism. Annual Review of Microbiology 72:null.

Muraki N, Seo D, Shiba T, Sakurai T, Kurisu G. 2010. Asymmetric dimeric structure of ferredoxin-NAD(P)$^+$ oxidoreductase from the green sulfur bacterium *Chlorobaculum tepidum*: implications for binding ferredoxin and NADP$^+$. J Mol Biol 401:403– 414. https://doi.org/10.1016/j.jmb.2010.06.024.

Nelson DL, Michael M. Cox. 2008. Lehninger principles of biochemistry Fifth.

Neuer G, Bothe H. 1985. Electron donation to nitrogenase in heterocysts of cyanobacteria. Arch Microbiol 143:185–191. https://doi.org/10.1007/BF00411045.

Nguyen DMN, Schut GJ, Zadvornyy OA, Tokmina-Lukaszewska M, Poudel S, Lipscomb GL et al. 2017. Two functionally distinct NADP(+)-dependent ferredoxin oxidoreductases maintain the primary redox balance of *Pyrococcus furiosus*. J Biol Chem 292: 14603-14616.

Nieva-Gomez D, Roberts GP, Klevickis S, Brill WJ. 1980. Electron transport to nitrogenase in *Klebsiella pneumoniae*. Proc Natl Acad SciUSA 77:2555–2558. https://doi.org/10.1073/pnas.77.5.2555.

Nishizawa M, Miyazaki J, Makabe A, Koba K, Takai K. 2014. Physiological and isotopic characteristics of nitrogen fixation by hyperthermophilic methanogens: key insights into nitrogen anabolism of the microbial communities in archean hydrothermal systems. Geochim Cosmochim Acta 138:117–135. https://doi.org/10.1016/j.gca.2014.04.021.

Nitschke, W., and Russell, M.J. 2012. Redox bifurcations: Mechanisms and importance to life now, and at its origin. Bioessays 34, 106-109.

Paradis, E., Claude, J., and Strimmer, K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20, 289-290.

Paul, K., Nonoh, J.O., Mikulski, L., and Brune, A. 2012. "Methanoplasmatales," Thermoplasmatales-related archaea in termite guts and other environments, are the seventh order of methanogens. Appl. Environ. Microbiol. 78, 8245-8253.

Pechter KB, Gallagher L, Pyles H, Manoil CS, Harwood CS. 2015. Essential genome of the metabolically versatile alphaproteobacterium *Rhodopseudomonas palustris*. J Bacteriol 198:867– 876. https://doi.org/10.1128/JB.00771-15.

Peden EA, Boehm M, Mulder DW, Davis R, Old WM, King PW, Ghirardi ML, Dubini A. 2013. Identification of global ferredoxin interaction networks in *Chlamydomonas reinhardtii*. J Biol Chem 288:35192–35209.https://doi.org/10.1074/jbc.M113.483727.

Petering D, Fee JA, Palmer G. 1971. The oxygen sensitivity of spinach ferredoxin and other iron-sulfur proteins. The formation of protein bound sulfur-zero. J Biol Chem 246:643– 653.

Peters, JW, Beratan, DN, Schut, GJ, & Adams, MW. 2018. On the nature of organic and inorganic centers that bifurcate electrons, coupling exergonic and endergonic oxidation–reduction reactions. Chemical Communications, 54(33), 4091-4099.

Peters JW, Boyd ES, Hamilton T, Rubio LM (ed). 2011. Biochemistry of Mo-nitrogenase. Caister Academic Press, Poole, United Kingdom.

Peters JW, Schut GJ, Boyd ES, Mulder DW, Shepard EM, Broderick JB, King PW, Adams MWW. 2015. [FeFe]- and [NiFe]-hydrogenase diversity, mechanism, and maturation. Biochim Biophys Acta 1853:1350 –1369. https://doi.org/10.1016/j.bbamcr.2014.11.021.

Peters, J.W., Lanzilotta, W.N., Lemon, B.J., and Seefeldt, L.C. 1998. X-ray crystal structure of the Fe-only hydrogenase (Cpl) from *Clostridium pasteurianum* to 1.8 angstrom resolution. Science 282, 1853-1858.

Peters, J.W., Miller, A.F., Jones, A.K., King, P.W., and Adams, M.W.W. 2016. Electron bifurcation. Curr. Opin. Chem. Biol. 31, 146-152.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera: a visualization system for exploratory research and analysis. J Comput Chem 25:1605–1612. https://doi.org/10.1002/jcc.20084.

Poole RK, Hill S. 1997. Respiratory protection of nitrogenase activity in *Azotobacter vinelandii*: roles of the terminal oxidases. Biosci Rep 17:303–317. https://doi.org/10.1023/A:1027336712748.

Posewitz MC, King PW, Smolinski SL, Zhang L, Seibert M, Ghirardi ML. 2004. Discovery of two novel radical S-adenosylmethionine proteins required for the assembly of an active [Fe] hydrogenase. J Biol Chem 279:25711–25720. https://doi.org/10.1074/jbc.M403206200.

Poudel, S., Colman, D.R., Fixen, K.R., Ledbetter, R., Zheng, Y., Pence, N., Seefeldt, L.C., Peters, J.W., Harwood, C.S., and Boyd, E.S. 2018. Electron transfer to nitrogenase in different genomic and metabolic backgrounds. J. Bacteriol. 200, e00757-17.

Poudel, S., Tokmina-Lukaszewska, M., Colman, D.R., Refai, M., Schut, G.J., King, P.W., Maness, P.C., Adams, M.W.W., Peters, J.W., Bothner, B., and Boyd, E.S. 2016. Unification of [FeFe]-hydrogenases into three structural and functional groups. Biochim. Biophys. Acta, Gen. Subj. 1860, 1910-1921.

Pradeep L, Shin HC, Scheraga HA. 2006. Correlation of folding kinetics with the number and isomerization states of prolines in three homologous proteins of the RNase family. FEBS Lett 580:5029 –5032. https://doi.org/10.1016/j.febslet.2006.08.024.

Pühler A, Aguilar MO, Hynes M, Müller P, Klipp W, Priefer U, Simon R, Weber G. 1984. Advances in the genetics of free-living and symbiotic nitrogen fixing bacteria, p 609 – 619. In Veeger C, Newton WE (ed), Advances in nitrogen fixation research: proceedings of the 5th International Symposium on Nitrogen Fixation, Noordwijkerhout, The Netherlands, August 28 –September 3, 1983. Springer Netherlands, Dordrecht, Netherlands.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F.O. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 41, D590-D596.

R Development Core Team. 2010. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The natural history of nitrogen fixation. Mol Biol Evol 21:541–554. https://doi.org/10.1093/molbev/msh047.

Rey FE, Oda Y, Harwood CS. 2006. Regulation of uptake hydrogenase and effects of hydrogen utilization on gene expression in *Rhodopseudomonas palustris*. J Bacteriol 188:6143– 6152. https://doi.org/10.1128/JB.00381-06.

Reyntjens B, Jollie DR, Stephens PJ, GaoSheridan HS, Burgess BK. 1997. Purification and characterization of a *fixABCX*-linked 2[4Fe-4S] ferredoxin from *Azotobacter vinelandii*. J Biol Inorg Chem 2: 595-602.

Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.F., Darling, A., Malfatti, S., Swan, B.K., Gies, E.A., Dodsworth, J.A., Hedlund, B.P., Tsiamis, G., Sievert, S.M., Liu, W.T., Eisen, J.A., Hallam, S.J., Kyrpides, N.C., Stepanauskas, R., Rubin, E.M., Hugenholtz, P., and Woyke, T. 2013. Insights into the phylogeny and coding potential of microbial dark matter. Nature 499, 431-437.

Roberts DL, Frerman FE, Kim JJ. 1996. Three-dimensional structure of human electron transfer flavoprotein to 2.1-Å resolution. Proc Natl Acad Sci U S A 93:14355–14360. https://doi.org/10.1073/pnas.93.25.14355.

Roberts DL, Salazar D, Fulmer JP, Frerman FE, Kim JJ. 1999. Crystal structure of *Paracoccus denitrificans* electron transfer flavoprotein: structural and electrostatic analysis of a conserved flavin binding domain. Biochemistry 38:1977–1989. https://doi.org/10.1021/bi9820917.

Roberts GP, MacNeil T, MacNeil D, Brill WJ. 1978. Regulation and characterization of protein products coded by the *nif* (nitrogen fixation) genes of *Klebsiella pneumoniae*. J Bacteriol 136:267–279.

Rubio LM, Ludden PW. 2008. Biosynthesis of the iron-molybdenumcofactor of nitrogenase. Annu Rev Microbiol 62:93–111. https://doi.org/10.1146/annurev.micro.62.081307.162737.

Ruvkun GB, Sundaresan V, Ausubel FM. 1982. Directed transposon Tn5 mutagenesis and complementation analysis of *Rhizobium meliloti* symbiotic nitrogen fixation genes. Cell 29:551–559. https://doi.org/10.1016/0092-8674(82)90171-4.

Rypniewski WR, Breiter DR, Benning MM, Wesenberg G, Oh BH, Markley JL, Rayment I, Holden HM. 1991. Crystallization and structure determination of 2.5 Å resolution of the oxidized [2Fe-2S] ferredoxin isolated from *Anabaena* 7120. Biochemistry 30:4126 – 4131. https://doi.org/10.1021/bi00231a003.

Saeki K (ed). 2004. Electron transport to nitrogenase: diverse routes for a common destination. Kluwer, Dordrecht, the Netherlands.

Saeki K, Suetsugu Y, Tokuda K-I, Miyatake Y, Young D, Marrs B, Matsubara H. 1991. Genetic analysis of functional differences among distinct ferredoxins in *Rhodobacter capsulatus*. J Biol Chem 266:12889 –12895.

Sait, M., Hugenholtz, P., and Janssen, P.H. 2002. Cultivation of globally distributed soil bacteria from phylogenetic lineages previously only detected in cultivation-independent surveys. Environ. Microbiol. 4, 654-666.

Sandmann G, Peleato ML, Fillat MF, Lazaro MC, Gomezmoreno C. 1990. Consequences of the iron-dependent formation of ferredoxin and flavodoxin on photosynthesis and nitrogen fixation on *Anabaena* strains. Photosynth Res 26:119 –125. https://doi.org/10.1007/BF00047083.

Sapra R, Bagramyan K, Adams MWW. 2003. A simple energy-conserving system: Proton reduction coupled to proton translocation. P Natl Acad Sci USA 100: 7545-7550.

Sarkar, A., Köhler, J., Hurek, T., and Reinhold-Hurek, B. 2012. A novel regulatory role of the Rnf complex of Azoarcus sp. strain BH72. Mol. Microbiol. 83, 408-422.

Sato K, Nishina Y, Shiga K. 2003. Purification of electron-transferring flavoprotein from *Megasphaera elsdenii* and binding of additional FAD with an unusual absorption spectrum. J Biochem 134:719 –729. https://doi.org/10.1093/jb/mvg199.

Sattley WM, Madigan MT, Swingley WD, Cheung PC, Clocksin KM, Conrad AL, Dejesa LC, Honchak BM, Jung DO, Karbach LE, Kurdoglu A, Lahiri S, Mastrian SD, Page LE, Taylor HL, Wang ZT, Raymond J, Chen M, Blankenship RE, Touchman JW. 2008. The genome of *Heliobacterium modesticaldum*, a phototrophic representative of the Firmicutes containing the simplest photosynthetic apparatus. J Bacteriol 190: 4687–4696. https://doi.org/10.1128/JB.00299-08.

Schatt E, Jouanneau Y, Vignais P. 1989. Molecular cloning and sequence analysis of the structural gene of ferredoxin I from the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 171:6218 –6226. https://doi.org/10.1128/jb.171.11.6218-6226.1989.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., and Robinson, C.J. 2009. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75, 7537-7541.

Schmehl M, Jahn A, Vilsendorf AMZ, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W. 1993. Identification of a new class of nitrogen-fixation genes in *Rhodobacter capsulatus*: a putative membrane complex involved in electron transport to nitrogenase. Mol Gen Genet 241:602– 615. https://doi.org/10.1007/BF00279903.

Schmitz O, Boison G, Hilscher R, Hundeshagen B, Zimmer W, Lottspeich F, Bothe H. 1995. Molecular biological analysis of a bidirectional hydrogenase from cyanobacteria. Eur J Biochem 233:266 –276. https://doi.org/10.1111/j.1432-1033.1995.266_1.x.

Schrautemeier B, Böhme H. 1985. A distinct ferredoxin for nitrogen fixation isolated from heterocysts of the cyanobacterium *Anabaena variabilis*. FEBS Lett 184:304 –308. https://doi.org/10.1016/0014-5793(85)80627-X.

Schrautemeier B, Cassing A, Bohme H. 1994. Characterization of the genome region encoding an FdxH-type ferredoxin and a new 2[4fe-4s] ferredoxin from the nonheterocystous, nitrogen-fixing cyanobacterium *Plectonema boryanum* PCC-73110. J Bacteriol 176:1037–1046. https://doi.org/10.1128/jb.176.4.1037-1046.1994.

Schuchmann K, Müller V. 2012. A bacterial electron-bifurcating hydrogenase. J Biol Chem 287: 31165-31171.

Schut GJ, Adams MWW. 2009. The iron-hydrogenase of *Thermotoga maritima* utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production. J Bacteriol 191: 4451-4457.

Schut GJ, Zadvornyy O, Wu C-H, Peters JW, Boyd ES, Adams MWW. 2016. The role of geochemistry and energetics in the evolution of modern respiratory complexes from a proton-reducing ancestor. Biochim Biophys Acta 1857:958 –970. https://doi.org/10.1016/j.bbabio.2016.01.010.

Schut, G.J., Boyd, E.S., Peters, J.W., and Adams, M.W.W. 2013. The modular respiratory complexes involved in hydrogen and sulfur metabolism by heterotrophic hyperthermophilic archaea and their evolutionary implications. FEMS Microbiol. Rev. 37, 182-203.

Schut, G.J., Zadvornyy, O., Wu, C.H., Peters, J.W., Boyd, E.S., and Adams, M.W.W. 2016. The role of geochemistry and energetics in the evolution of modern respiratory complexes from a proton-reducing ancestor. Biochim. Biophys. Acta, Bioenerg. 1857, 958-970.

Scott JD, Ludwig RA. 2004. *Azorhizobium caulinodans* electron transferring flavoprotein N electrochemically couples pyruvate dehydrogenase complex activity to N_2 fixation. Microbiology 150:117–126. https://doi.org/10.1099/mic.0.26603-0.

Seefeldt LC, Hoffman BM, Dean DR. 2009. Mechanism of Mo-dependent nitrogenase. Annu Rev Biochem 78:701–722. https://doi.org/10.1146/annurev.biochem.78.070907.103812.

Seo D, Sakurai H. 2002. Purification and characterization of ferredoxinNAD(P)(+) reductase from the green sulfur bacterium *Chlorobium tepidum*. Biochim Biophys Acta 1597:123–132. https://doi.org/10.1016/S0167-4838(02)00269-8.

Setubal JC, dos Santos P, Goldman BS, Ertesvag H, Espin G, Rubio LM, Valla S, Almeida NF, Balasubramanian D, Cromes L, Curatti L, Du ZJ, Godsy E, Goodner B, Hellner-Burris K, Hernandez JA, Houmiel K, Imperial J, Kennedy C, Larson TJ, Latreille P, Ligon LS, Lu J, Maerk M, Miller NM, Norton S, O'Carroll IP, Paulsen I, Raulfs EC, Roemer R, Rosser J, Segura D, Slater S, Stricklin SL, Studholme DJ, Sun J, Viana CJ, Wallin E, Wang BM, Wheeler C, Zhu HJ, Dean DR, Dixon R, Wood D. 2009. Genome sequence of Azotobacter vinelandii, an obligate aerobe specialized to support diverse

anaerobic metabolic processes. J Bacteriol 191:4534 – 4545. https://doi.org/10.1128/JB.00504-09.

Shah VK, Stacey G, Brill WJ. 1983. Electron transport to nitrogenase: purification and characterization of pyruvate, flavodoxin oxidoreductase, the *nifJ*-gene product. J Biol Chem 258:2064 –2068.

Shen, L.D., Wu, H.S., Gao, Z.Q., Cheng, H.X., Li, J., Liu, X., and Ren, Q.Q. 2016. Distribution and activity of anaerobic ammonium-oxidising bacteria in natural freshwater wetland soils. Appl. Microbiol. Biotech. 100, 3291-3300.

Shock EL, Boyd ES. 2015. Principles of geobiochemistry. Elements 11:395– 401. https://doi.org/10.2113/gselements.11.6.395.

Shock, E.L., Holland, M., Meyer-Dombard, D., Amend, J.P., Osburn, G.R., and Fischer, T.P. 2010. Quantifying inorganic sources of geochemical energy in hydrothermal ecosystems, Yellowstone National Park, USA. Geochim. Cosmochim. Acta 74, 4005-4043.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539. https://doi.org/10.1038/msb.2011.75.

Silvestro D, Michalak I. 2012. raxmlGUI: a graphical front-end for RAxML. Org Divers Evol 12:335–337. https://doi.org/10.1007/s13127-011-0056-0.

Simpson FB, Burris RH. 1984. A nitrogen pressure of 50 atmospheres does not prevent evolution of hydrogen by nitrogenase. Science 224: 1095–1097. https://doi.org/10.1126/science.6585956.

Singh BB, Curdt I, Jakobs C, Schomburg D, Bisen PS, Bohme H. 1999. Identification of amino acids responsible for the oxygen sensitivity of ferredoxins from *Anabaena variabilis* using site-directed mutagenesis. Biochim Biophys Acta 1412:288 –294. https://doi.org/10.1016/S0005-2728(99)00069-9.

Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431– 432.

Sorokin, D.Y., Makarova, K.S., Abbas, B., Ferrer, M., Golyshin, P.N., Galinski, E.A., Ciordia, S., Mena, M.C., Merkel, A.Y., Wolf, Y.I., Van Loosdrecht, M.C.M., and Koonin, E.V. 2017. Discovery of extremely halophilic, methyl-reducing euryarchaea provides insights into the evolutionary origin of methanogenesis. Nat. Microbiol. 2.

Sousa FL, Nelson-Sathi S, Martin WF. 2016. One step beyond a ribosome: The ancient anaerobic core. Biochimica et Biophysica Acta (BBA)-Bioenergetics 1857: 1027-1038.

Sousa FL, Preiner M, Martin WF. 2018. Native metals, electron bifurcation, and CO2 reduction in early biochemical evolution. Curr Opin Microbiol 43: 77-83.

Souza ALF, Invitti AL, Rego FGM, Monteiro RA, Klassen G, Souza EM, Chubatsu LS, Pedrosa FO, Rigo LU. 2010. The involvement of the nif-associated ferredoxin-like genes *fdxA* and *fdxN* of *Herbaspirillum seropedicae* in nitrogen fixation. J Microbiol 48:77– 83. https://doi.org/10.1007/s12275-009-0077-y.

Stal LJ, Krumbein WE. 1985. Nitrogenase activity in the nonheterocystous cyanobacterium *Oscillatoria* sp. grown under alternating light-dark cycles. Arch Microbiol 143:67–71. https://doi.org/10.1007/BF00414770.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics (Oxford, England) 30, 1312-1313.

Stein, L.Y., La Duc, M.T., Grundl, T.J., and Nealson, K.H. 2001. Bacterial and archaeal populations associated with freshwater ferromanganous micronodules and sediments. Environ. Microbiol. 3, 10-18.

Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596-1599.

Tan X, Volbeda A, Fontecilla-Camps JC, Lindahl PA. 2006. Function of the tunnel in acetylcoenzyme A synthase/carbon monoxide dehydrogenase. J Biol Inorg Chem 11:371–378. https://doi.org/10.1007/s00775-006-0086-9.

Team R, R.C. 2014. "R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2014".).

Thauer RK, Kaster AK, Seedorf H, Buckel W, Hedderich R. 2008. Methanogenic archaea: ecologically relevant differences in energy conservation. Nat Rev Microbiol 6: 579-591.

Thauer, R.K., Kaster, A.-K., Goenrich, M., Schick, M., Hiromoto, T., and Shima, S. 2010. Hydrogenases from methanogenic archaea, nickel, a novel cofactor, and H2 storage. Ann. Rev. Biochem. 79, 507-536.

Thauer, R.K., Jungermann, K. and Decker, K., 1977. Energy conservation in chemotrophic anaerobic bacteria. Bacteriological reviews, 41(1), p.100.

Therien J, Artz JH, Poudel S, Hamilton TL, Liu Z, Noone S, Adams M, King PW, Bryant DA, Boyd ES, Peters JW. 2017. The physiological functions and structural determinants of catalytic bias in the [FeFe]-hydrogenases CpI and CpII of *Clostridium pasteurianum* strain W5. Front Microbiol 8:1305. https://doi.org/10.3389/fmicb.2017.01305.

Toogood HS, Leys D, Scrutton NS. 2007. Dynamics driving function: new insights from electron transferring flavoproteins and partner complexes. FEBS J 274:5481–5504. https://doi.org/10.1111/j.1742-4658.2007.06107.x.

Tsai MH, Saier MH, Jr. 1995. Phylogenetic characterization of the ubiquitous electron transfer flavoprotein families ETF-α and ETF-β. Res Microbiol 146:397– 404. https://doi.org/10.1016/0923-2508(96)80285-3.

Tso M-YW, Ljones T, Burris R. 1972. Purification of the nitrogenase proteins from *Clostridium pasteurianum*. Biochim Biophys Acta 267:600 – 604. https://doi.org/10.1016/0005-2728(72)90193-4.

Vanwonterghem, I., Evans, P.N., Parks, D.H., Jensen, P.D., Woodcroft, B.J., Hugenholtz, P., and Tyson, G.W. 2016. Methylotrophic methanogenesis discovered in the archaeal phylum Verstraetearchaeota. Nat. Microbiol. 1. 16170.

Varma SJ, Muchowska KB, Chatelain P, Moran J. 2018. Native iron reduces CO2 to intermediates and end-products of the acetyl-CoA pathway. Nature ecology & evolution: 1.

Vignais, P.M., Billoud, B., and Meyer, J. 2001. Classification and phylogeny of hydrogenases. FEMS Microbiol. Rev. 25, 455-501.

Vitousek PM, Menge DNL, Reed SC, Cleveland CC. 2013. Biological nitrogen fixation: rates, patterns and ecological controls in terrestrial ecosystems. Philos Trans R Soc B Biol Sci 368:20130119. https://doi.org/10.1098/rstb.2013.0119.

Volbeda A, Charon MH, Piras C, Hatchikian EC, Frey M, Fontecilla-Camps JC. 1995. Crystal structure of the nickel-iron hydrogenase from *Desulfovibrio gigas*. Nature 373:580 –587. https://doi.org/10.1038/373580a0.

Wahl RC, Orme-Johnson WH. 1987. Clostridial pyruvate oxidoreductase and the pyruvate-oxidizing enzyme specific to nitrogen fixation in *Klebsiella pneumoniae* are similar enzymes. J Biol Chem 262:10489 –10496.

Walt A, Kahn ML. 2002. The *fixA* and *fixB* genes are necessary for anaerobic carnitine reduction in Escherichia coli. J Bacteriol 184:4044 – 4047. https://doi.org/10.1128/JB.184.14.4044-4047.2002.

Wang J, Wang W, Kollman PA, Case DA. 2006. Automatic atom type and bond type perception in molecular mechanical calculations. J Mol Graph Model 25:247–260. https://doi.org/10.1016/j.jmgm.2005.12.005.

Wang S, Huang H, Kahnt J, Thauer RK. 2013. *Clostridium acidurici* electron-bifurcating formate dehydrogenase. Appl Environ Microb 79: 6176-6179.

Wang, S., Huang, H., Kahnt, J., Mueller, A.P., Köpke, M., and Thauer, R.K. 2013a. NADP-specific electron-bifurcating [FeFe]-hydrogenase in a functional complex with formate dehydrogenase in *Clostridium autoethanogenum* grown on CO. J. Bacteriol. 195, 4373-4386.

Wang, S.N., Huang, H.Y., Moll, J., and Thauer, R.K. 2010. NADP+ reduction with reduced ferredoxin and NADP+ reduction with NADH are coupled via an electron-bifurcating enzyme complex in *Clostridium kluyveri*. J. Bacteriol. 192, 5115-5123.

Watmough NJ, Kiss J, Frerman FE. 1992. Structural and redox relationships between *Paracoccus denitrificans*, porcine and human electrontransferring flavoproteins. Eur J Biochem 205:1089 –1097. https://doi.org/10.1111/j.1432-1033.1992.tb16877.x.

Watt G, Burns A, Lough S, Tennent D. 1980. Redox and spectroscopic properties of oxidized MoFe protein from *Azotobacter vinelandii*. Biochemistry 19:4926 – 4932. https://doi.org/10.1021/bi00562a035.

Weghoff MC, Bertsch J, Muller V 2015. A novel mode of lactate metabolism in strictly anaerobic bacteria. Environmental Microbiology 17: 670-677.

Weiss, M.C., Sousa, F.L., Mrnjavac, N., Neukirchen, S., Roettger, M., Nelson-Sathi, S., and Martin, W.F. 2016. The physiology and habitat of the last universal common ancestor. Nature Microbiology 1:16116.

Wimley WC, Creamer TP, White SH. 1996. Solvation energies of amino acid side chains and backbone in a family of host-guest pentapeptides. Biochemistry 35:5109 –5124. https://doi.org/10.1021/bi9600153.

Yakunin AF, Gennaro G, Hallenbeck PC. 1993. Purification and properties of a nif-specific flavodoxin from the photosynthetic bacterium *Rhodobacter capsulatus*. J Bacteriol 175:6775–6780. https://doi.org/10.1128/jb.175.21.6775-6780.1993.

Yakunin AF, Hallenbeck PC. 1998. Purification and characterization of pyruvate oxidoreductase from the photosynthetic bacterium *Rhodobacter capsulatus*. Biochim Biophys Acta 1409:39 – 49. https://doi.org/10.1016/S0005-2728(98)00145-5.

Yan Z, Wang MY, Ferry JG. 2017. A ferredoxin- and F420H2-dependent, electron bifurcating, heterodisulfide reductase with homologs in the domains Bacteria and Archaea. Mbio 8.

Yang Z-Y, Ledbetter R, Shaw S, Pence N, Tokmina-Lukaszewska M, Eilers B, Guo Q, Pokhrel N, Cash VL, Dean DR. 2016. Evidence that the Pi release event is the rate-limiting step in the nitrogenase catalytic cycle. Biochemistry 55:3625–3635. https://doi.org/10.1021/acs.biochem.6b00421.

Yoch DC, Arnon DI. 1972. Two biologically active ferredoxins from the aerobic nitrogen-fixing bacterium, *Azotobacter vinelandii*. J Biol Chem 247:4514 – 4520.

Zhang HB, Chobot SE, Osyczka A, Wraight CA, Dutton PL, Moser CC. 2008. Quinone and non-quinone redox couples in Complex III. Journal of Bioenergetics and Biomembranes 40: 493-499.

Zhang J, Frerman FE, Kim JJ. 2006. Structure of electron transfer flavoprotein-ubiquinone oxidoreductase and electron transfer to the mitochondrial ubiquinone pool. Proc Natl Acad SciUSA 103:16212–16217. https://doi.org/10.1073/pnas.0604567103.

Zhang P, Yuly JL, Lubner CE, Mulder DW, King PW, Peters JW et al. 2017. Electron Bifurcation: Thermodynamics and kinetics of two-electron brokering in biological redox chemistry. Accounts Chem Res 50: 2410-2417.

Zheng Y, Harris DF, Yu Z, Fu Y, Poudel S, Ledbetter RN, Fixen KR, Yang Z-Y, Boyd ES, Lidstrom ME, Seefeldt LC, Harwood CS. 2018. A pathway for biological methane production using bacterial iron-only nitrogenase. Nat Microbiol 3:281–186. https://doi.org/10.1038/s41564-017-0091-5.

APPENDICES

APPENDIX A

UNIFICATION OF [FEFE]-HYDROGENASES INTO THREE STRUCTURAL AND FUNCTIONAL GROUPS

Contribution of Authors and Co-Authors

Manuscript in Appendix A

Author: Saroj Poudel

Contributions: Designed and conducted the experiment. Contributed to the writing of the manuscript.
Co-Author: Monika Tokmina-Lukaszewska

Contributions: Conducted mass spectrometry analysis. Contributed to the writing of the manuscript.

Co-Author: Daniel R. Colman

Contributions: Assisted with bioinformatics analysis. Contributed to the writing of the manuscript.

Co-Author: Mohammed Refai

Contributions: Assisted in mass spectrometry analysis.

Co-Author: Gerrit J. Schut
Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Paul W. King

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Pin-Ching Maness

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Michael W.W. Adams

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: John W. Peters

Contributions: Provided field expertise and feedback on drafts of the manuscripts.

Co-Author: Brian Bothner

Contributions: Designed and conducted the mass spectrometry experiment. Contributed to the writing of the manuscript.

Co-Author: Eric S. Boyd

Contributions: Provided feedback on the study design. Provided statistical advice and contributed to the writing of the manuscript.

Manuscript Information Page

Saroj Poudel, Monika Tokmina-Lukaszewska, Daniel R. Colman, Mohammed Refai, Gerrit J. Schut, Paul W. King, Pin-Ching Maness, Michael W.W. Adams, John W. Peters, Brian Bothner, Eric S. Boyd

Biochimica et Biophysica Acta – General Subjects

Status of Manuscript:
____ Prepared for submission to a peer-reviewed journal
____ Officially submitted to a peer-review journal
____ Accepted by a peer-reviewed journal
__x__ Published in a peer-reviewed journal

Published online September 2016, https://doi.org/10.1016/j.bbagen.2016.05.034.

Unification of [FeFe]-hydrogenases into Three Structural and Functional Groups

Saroj Poudel[1], Monika Tokmina-Lukaszewska[2], Daniel R. Colman[1], Mohammed Refai[2], Gerrit J. Schut[3], Paul W. King[4], Pin-Ching Maness[4], Michael W.W. Adams[3], John W. Peters[2], Brian Bothner[2] and Eric S. Boyd[1*]

[1]*Department of Microbiology and Immunology, Montana State University, Bozeman, MT 59717*
[2]*Department of Chemistry and Biochemistry, Montana State University, Bozeman, MT 59717*
[3]*Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA 30602*
[4]*Biosciences Center, National Renewable Energy Laboratory, Golden, CO 80401*

Correspondence* Eric S. Boyd (*eboyd@montana.edu*)
Department of Microbiology & Immunology
Montana State University
PO Box 173520
Bozeman, MT 59717
Phone: (406) 994-7046
Fax: (406) 994-4926

ABSTRACT

Background. [FeFe]-hydrogenases (Hyd) are structurally diverse enzymes that catalyze the reversible oxidation of hydrogen (H_2). Recent biochemical data demonstrate new functional roles for these enzymes, including those that function in electron bifurcation where an exergonic reaction is coupled with an endergonic reaction to drive the reversible oxidation/production of H_2.

Methods. To identify the structural determinants that underpin differences in enzyme functionality, a total of 714 homologous sequences of the catalytic subunit, HydA, were compiled. Bioinformatics approaches informed by biochemical data were then used to characterize differences in inferred quaternary structure, HydA active site protein environment, accessory iron-sulfur clusters in HydA, and regulatory proteins encoded in HydA gene neighborhoods.

Results. HydA homologs were clustered into one of three classification groups, Group 1 (G1), Group 2 (G2), and Group 3 (G3). G1 enzymes were predicted to be monomeric while those in G2 and G3 were predicted to be multimeric and include HydB, HydC (G2/G3) and HydD (G3) subunits. Variation in the HydA active site and accessory iron-sulfur clusters did not vary by group type. Group-specific regulatory genes were identified in the gene neighborhoods of both G2 and G3 Hyd. Analyses of purified G2 and G3 enzymes by mass spectrometry strongly suggests that they are post-translationally modified by phosphorylation.

Conclusions. These results suggest that bifurcation capability is dictated primarily by the presence of both HydB and HydC in Hyd complexes, rather than by variation in HydA.

General Significance. This classification scheme provides a framework for future biochemical and mutagenesis studies to elucidate the functional role of Hyd enzymes.

1. INTRODUCTION

Hydrogenases catalyze the reversible reduction of protons to hydrogen (H_2) gas using a variety of electron donors and acceptors. The three primary variants of hydrogenase, termed [NiFe]-, [FeFe]-, and [Fe]-hydrogenases, are evolutionarily distinct and differ in the composition of their active site metal clusters [1–10]. Of the three variants, [FeFe]-hydrogenases (Hyd) are of substantial interest for biotechnology and bioenergy purposes due to their high rate of proton reduction [11–13]. Hyds are widely distributed in Bacteria and also have been detected in a number of Eukarya but have yet to be identified in Archaea [14–17]. Hyds vary in their cellular localization which corresponds with differing functionalities. For instance, Hyds can be localized in the periplasm or the cytoplasm and can be either membrane-bound or soluble [18–24]. Periplasmic Hyds are primarily involved in H_2 uptake while membrane-bound or soluble cytoplasmic Hyds are involved in reversible oxidation of H_2 [18–22,24].

At a minimum Hyds comprise a catalytic subunit, HydA, which contains a conserved ~350 residue H-cluster domain that harbors the ligands for the novel active site iron-sulfur cluster. The active site iron-sulfur cluster consists of a [4Fe-4S] cluster bridged to a 2Fe subcluster via a cysteine thiol [1,11,17,25]. In addition to the H-cluster

domain, HydA often contains N-terminal (F-cluster) and C-terminal (C-cluster) domains that add flexibility to the enzyme and that modulate electron flow to and/or from the active site H-cluster [11,14,16,17,25,26]. The simplest form of HydA includes those identified in green algae, such as *Chlamydomonas reinhardtii*, which generally only contains the H-cluster domain [27–32]. More complex forms of HydA are more common and possess several additional iron-sulfur clusters, including those with both the F- and C-clusters [2,11,33]. In addition to F- and C-cluster mediated functional flexibility, HydA can exist in multiple isoforms with different quaternary structures, including those that are monomeric and those that are multimeric [14,16,17,26,34]. Multiple HydAs of variable quaternary structure can be present in a single genome [22,35–37] where they form complexes with paralogs of the large and small subunits of NADH dehydrogenase (referred to as HydB and HydC, respectively) [21,35–39].

Recent biochemical studies indicate that several soluble, cytoplasmic and, multimeric Hyds are capable of electron bifurcation, a novel mechanism of energy conservation that involves the simultaneous reduction or oxidation of two electron acceptors or donors in an enzyme complex, whereby a thermodynamically favorable exergonic reaction drives a thermodynamically unfavorable endergonic reaction [21,35,37–42]. Soluble cytoplasmic Hyds identified from *Acetobacterium woodii* [21], *Clostridium autoethanogenum* [40], *Moorella thermoacetica* [36] and *Thermotoga maritima* [35] are all capable of electron bifurcation [38,39]. The trimeric Hyd from *T. maritima* (i.e. HydABC complex) couples the simultaneous oxidation of ferredoxin (Fd) and NADH to the production of H_2 [35]. In contrast, the tetrameric Hyd from *A. woodii*

(i.e. HydABCD complex) couples the oxidation of H_2 to the simultaneous reduction of Fd and NAD^+ [21]. Interestingly, a trimeric Hyd from the acetogen *M. thermoacetica* functions in vivo to either oxidize or produce H_2, contingent on cultivation conditions, by coupling Fd and NAD^+/NADH reduction/oxidation [43]. Autotrophically-grown *M. thermoacetica* cells oxidize H_2 by simultaneously reducing both Fd and NAD^+ while glucose-grown cells used the same enzyme complex to reduce protons to produce H_2 by coupling with the oxidation of Fd and NADH. Conversely, the Hyd from *Clostridium pasteurianum* (cytoplasmic and soluble) has been shown to not bifurcate electrons, which is consistent with its lack of a full length HydB [2,11,36]. Likewise, the 'dimeric' Hyd from *Desulfovibrio desulfuricans*, which has a truncated HydB that functions in part to target the enzyme to the periplasm, is unlikely to bifurcate [23,44].

The genomes of many organisms encode for monomeric and multimeric Hyds, for example, *A. woodii* [21], *Ruminococcus albus* [37], and *T. maritima* [35], among others [14,16]. The presence of multiple copies of HydA including those that bifurcate and those that do not suggests a need to regulate one or more of the enzymes in order to maximize energy conservation. Indeed, a recent study identified several genes that encode for proteins putatively involved in sensing H_2 and in post-translational modification (PTM) in the gene neighborhood of seven different hydAs (including both monomeric and trimeric proteins) [37]. One of those proteins, HydS, was proposed as a H_2 sensor based on the presence of a Per-ARNT-Sim (PAS) domain, which is widely used to promote protein-protein interactions or signal transduction [37,45]. Moreover, proteins hypothesized to be involved in signal transduction (e.g. serine-threonine

kinases/phosphatases and histidine kinases) have been identified in the flanking gene environment of hydA in *Clostridium thermocellum* and *T. maritima* [37]. Whether these enzymes are involved in PTM of Hyd proteins and the extent to which they are distributed in the gene neighborhoods of hydA among other microorganisms is not yet known.

In order to develop a robust classification scheme for use in predicting bifurcating and non-bifurcating HydA, we compiled all the available HydAs in completed genomes. Our goal was to build upon and at the same time reconcile past classification schemes [14,16]. A comprehensive characterization of the proteins encoded in the neighboring genomic regions of hydA was also conducted using bioinformatics approaches. This information was then used to develop classifications of Hyd that unify patterns in enzyme diversity at the level of primary and predicted quaternary structure. These groups were then analyzed in the context of biochemically characterized bifurcating and non-bifurcating enzymes to demarcate their diverse functionalities. Finally, we used this classification scheme to identify proteins that might be involved in the PTM of Hyds and to characterize their distribution in the Hyd classification groupings. The regulatory proteins identified include a histidine kinase-like (Hkl) protein (sometimes annotated as HydE [21]), serine/threonine kinase (Stk), and serine phosphatase (Sp) that may be involved in PTM of putatively multimeric bifurcating Hyds. Herein we report a new classification scheme that assigns HydA into one of these three groups, based on the composition of proteins encoded in adjacent gene neighborhoods and structural features

associated with these groups. Finally, we provide evidence from biochemical and mass spectrometry assays for PTM in multimeric Hyd complexes.

2. MATERIALS AND METHODS

2.1 Identification and compilation of HydA homologs

The amino acid sequence of *C. pasterianum* [FeFe]-hydrogenase CpI (WP_004455619.1) was used in a BLASTp query to identify homologs present in publicly available genomes in the Department of Energy (DOE) Integrated Microbial Genomes (IMG) database in June 2014 using default settings and an e-value of 10^{-5}. The extracted homologs were aligned using Clustal Omega [46] specifying default alignment parameters. Aligned homologs were screened manually for the presence of three H-cluster binding motifs, previously defined as L1 (TSCCPxW), L2 (MPCxxKxxE), and L3 (ExMACxxGCxxGGGxP) [34]. A total of 714 sequences were identified that contained the signature L1, L2, and L3 motifs. Despite previously reported conservation of the three motifs [34], our analyses indicated slight variations in each of them. Therefore, we analyzed the extent of variation in these three H-cluster binding motifs using the weblogo server [47]. The taxonomy associated with each HydA sequence was compiled using the DOE-IMG database.

2.2 Construction of an extended sequence dataset

A custom python (ver. 2.7.6) script was used to extract gene sequences (10 upstream and 10 downstream) that flanked each hydA. The twenty inferred protein

sequences were clustered by pairwise BLASTp using the NCBI blastclust program [48] specifying a threshold of 30% sequence identities and a coverage of 60% for each pair of sequences. Clusters that contained a minimum of 12 sequences were considered for further analysis, resulting in a total of 123 clusters. The protein sequences within each of the 123 clusters were blasted against the Conserved Domain Database (CDD): a public database that contains all known protein motifs and domains [49] using an e-value of 0.01 to confirm bin assignments that were made by cluster analysis. Protein sequences were pruned from the clusters based on the absence of key domains (as identified by CDD blast) associated with each protein cluster. In rare cases the proteins were manually reassigned to other clusters. The filtered protein clusters were then annotated using the UniProt database [50]. Proteins encoded in the gene neighborhood of hydA are reported in **Supplementary Table 1**.

2.3 Frequent pattern mining using the extended dataset

The protein clusters (n = 123) encoded in the gene neighborhoods of the 714 HydA homologs were converted into binary matrix for use in classifying the HydA homologs based on the presence/absence of proteins. To objectively identify the most frequent patterns in the composition of proteins encoded by genes flanking hydA, the Apriori algorithm was used as implemented with the arules package in R [51]. Apriori is an algorithm that identifies frequent items (i.e., protein clusters) within a database and characterizes associations among the distribution of those items to objectively identify patterns. We applied the Apriori algorithm to our protein cluster database while

specifying the following parameters: a confidence threshold of 0.95 and a support threshold of 0.1 [51,52]. The confidence threshold is the measure of the statistical significance of an identified pattern within a dataset while the support threshold is indicative of the relative abundance of a given pattern in a dataset. The Apriori algorithm works in two phases with the first phase consisting of a scan of the entire database of proteins encoded in flanking regions of hydA in order to objectively identify and extract frequently observed proteins. In the second phase, the algorithm uses those frequently occurring proteins to identify the dominant patterns in the distribution of these proteins or their combinations. In doing so, the algorithm identifies correlations (within the user defined confidence and support threshold) between and among proteins in the database [52]. The 32 most frequent patterns and their confidence intervals are reported in **Supplementary Table 2**.

2.4 Statistical analysis

To further evaluate the extent to which Hyd group designations correspond to differences in the composition of proteins encoded in the flanking regions of hydA, the binary presence/absence table of the protein encoding genes flanking hydA (+/−10 genes) were subjected to Principal Component Analysis (PCA). Only proteins that were present in flanking genes that were represented in N2% of all HydA encoding genomes were considered in the generation of the dissimilarity matrix. The dissimilarity matrix was subjected to PCA (using a variance covariance matrix) to visualize variation in the composition of genes that flank hydA using PAST (version 3) [53].

2.5 Network analysis

The abundant patterns that were (i.e. ≥95% confidence value and ≥10% support value) identified by the Apriori algorithm were used to create a pattern correlating network with Cytoscape specifying the force directed organic layout [54]. Each unique pattern was denoted as a node and the edges between nodes represent the degree of correlation between the nodes. The network was further analyzed using the Network Analyzer plugin in Cytoscape [55] to calculate the relatedness (i.e., betweenness centrality) of each node in the network where the edges represent the degree of correlation between the nodes. Proteins with the highest betweenness centrality were overlaid on the network.

Proteins encoded in the flanking regions of hydA that had a relative frequency of ≥10% were extracted and used to create a binary matrix to identify the presence/absence of the protein in the gene context of each hydA. This matrix was then used to generate a network map to identify the protein clusters most commonly identified in each Hyd group. Cytoscape was used to visualize the network using a force directed organic layout and the network was further analyzed using the Network Analyzer plugin.

2.6 Determining the F- and C-cluster composition of HydA homologs

The N- and C-terminal regions that flank the H-cluster (i.e., F- and C-cluster domains, respectively) of HydA were trimmed from alignment blocks and were subjected to BLASTp against CDD using the CDSEARCH/cdd v3.13 algorithm [49] (version 3.13)

using an e-value of 0.01 as previously described [14]. The identified F- and C-cluster domains were compiled for each sequence using a custom python (ver.2.7.6) script and were used to classify the putative domains of HydA and to determine if these structures varied according to group designation.

2.7 Analysis of Post-Translational Modification

The Hyd (i.e. HydABC) from *T. maritima* was purified as described previously [22] while purification and properties of the HydABCD enzyme from *C. bescii* will be reported elsewhere. *Pyrococcus furiosus* transhydrogenase (also known as NfnI or sulfide dehydrogenase) was purified as previously described [56].

PTM identification (phosphorylation) was conducted on intact protein complexes. Samples were generated by spiking 4 μg of protein complexes with Calf Intestinal Alkaline Phosphatase (CIP, New England BioLabs) (1 CIP unit/1 μg protein) and incubation for 3 h. at 37 °C in 50 mM ammonium bicarbonate, pH 8.5. Casein (Affymetrix Life Science) was chosen as a positive control for phosphorylation. After treatment samples were immediately mixed with gel-loading buffer, boiled for 5 min, centrifuged, and loaded onto 4–20% mini-protean TGX gel (Bio-Rad). Controls were prepared using the same protocol, however no CIP was added. Proteins were visualized with Pro-Q Diamond Phosphoprotein Stain (Invitrogen/Molecular Probes) and GelCode Blue Stain (ThermoFisher) according to manufacturer instructions. PTM identification (phosphorylation) was conducted on intact protein complexes. Samples were generated by spiking 4 μg of protein complexes with Calf Intestinal Alkaline Phosphatase (CIP,

New England BioLabs) (1 CIP unit/1 µg protein) and incubation for 3 h. at 37 °C in 50 mM ammonium bicarbonate, pH 8.5. Casein (Affymetrix Life Science) was chosen as a positive control for phosphorylation. After treatment samples were immediately mixed with gel-loading buffer, boiled for 5 min., centrifuged, and loaded onto 4-20% mini-protean TGX gel (Bio-Rad). Controls were prepared using the same protocol, however no CIP was added. Proteins were visualized with Pro-Q Diamond Phosphoprotein Stain (Invitrogen/Molecular Probes) and GelCode Blue Stain (ThermoFisher) according to manufacturer instructions.

Concurrently, intact protein samples were analyzed using a 1290 ultrahigh pressure (UPLC) series chromatography stack (Agilent Technologies) coupled directly to an electrospray-time of flight (ESI-TOF) mass spectrometer (Bruker Micro-TOF). Before infusion to ESI source, samples were separated on a reverse-phase (RP) column PLRP-S column (50 x 2.1 mm, 3 µm, 100 Å, Agilent Technologies) or a size exclusion (SEC) Polyhydroxyethyl A column (100 x 4.6 mm, 5 µm, 500 Å, PolyLC Inc.) with flow rates of 600 µL/min. (RP column) and 300 µL/min. (SEC column). Fast gradient conditions (1 min., 10% B; 1.0-6.0 min., 10-70% B; 6.0-7.0 min., 10% B) were applied to elute proteins from the RP column (temperature of 50°C), while isocratic conditions (25% B) were used to elute proteins for 4 min. from the SEC column (temperature of 25°C). Solvent A comprised 0.1% formic acid (FA, Sigma) in water (ThermoFisher) and solvent B comprised 0.1% FA in acetonitrile (ThermoFisher).

Electrospray conditions for RP and SEC chromatography were as follow: nebulizer 3.0 or 3.5 bar (RP and SEC columns, respectively), drying gas at a flow rate of

7.0 or 6.0 L/min. (RP and SEC columns, respectively), drying temperature at 200°C, capillary voltage at 4.5 kV, and capillary exit voltage at 100 V. Data was collected in positive mode only at 2 Hz rate over the 200-3000 m/z scan range. Data processing and analysis were performed using the Bruker Data Analysis package 4.0. Charge deconvolution was performed using a maximum entropy algorithm for H^+ adducts only and 0.1 m/z data point spacing. The low mass end was defined by the mass of lightest component while the high mass end was defined as 3.3x of the heaviest component within the complex. Measured m/z errors for all proteins were less than 0.8 m/z from calculated values.

3. RESULTS AND DISCUSSION

3.1 Taxonomic distribution and classification of HydA homologs.

The taxonomic distribution of *hydA* was examined in 30 bacterial, 5 archaeal, and 9 eukaryotic phyla with available and complete genome sequences, as determined in June 2014 (**Table 1**). None of the archaeal phyla (167 total genomes) contained organisms whose genomes encoded for HydA while 17 of the 30 bacterial phyla (287 total genomes) contained organisms whose genomes encoded for HydA. Three of the 9 eukaryotic phyla (5 total genomes) contained organisms whose genomes encoded for HydA. Ninety percent of the 714 *hydA* homologs identified were distributed among the bacterial phyla Firmicutes, Proteobacteria, Spirochaetes, Thermotogae and which encoded on average ~ 2 HydA isoforms per HydA encoding genome.

The genes flanking (+/- 10 genes) the 714 *hydA* homologs were compiled and the encoded proteins clustered into homologous bins. The Apriori algorithm was then applied to this dataset to identify genes present at a high frequency in genomic regions flanking *hydA* and to calculate their co-occurrence patterns for use in predicting potential interactions. HydB and HydC were the most prevalent proteins encoded in genomic regions flanking *hydA* and these were identified in ~40% of the 714 *hydA* flanking gene neighborhoods (**Fig. 1**). An ABC transporter protein, HydD, and Hkl were the next most frequently identified proteins which were encoded in ~28%, 15%, and 12% of the *hydA* flanking gene neighborhoods, respectively **(Fig. 1)**.

The categorized proteins from the first phase of the Apriori algorithm were then used to generate association rules for classifying HydA based on the distribution of proteins or combinations therein encoded in the flanking regions of *hydA*. A total of 32 different flanking protein combinations were identified that were within the confidence value of 0.95 or higher (**Supplementary Table 1**). Of the 32 statistically supported flanking protein combinations, 98% of them contained either HydB, HydC, HydD, or combinations of these proteins (**Supplementary Table 1)**. The identified flanking protein combinations (n=32) were subjected to network analysis to identify the most prevalent co-occurrence patterns.

HydB, HydC, HydD, Hkl, a protein containing a polymerase and histidinol phosphatase (PHP) domain, and a protein with a DRTGG domain also had high betweenness centrality (e.g., connectivity among nodes) (**Fig. 2**). This indicates that these proteins were shared with a number of the 32 patterns and therefore behave as ‘bridges’

that connect the entire protein cluster network. Of these proteins, HydB, HydC, and HydD had the highest betweenness centrality. This indicates that these three proteins are highly represented in flanking regions of genomes that encode *hydA* and presumably, given the available biochemical evidence discussed below, interact with HydA.

To further characterize the proteins in the flanking regions of hydA, a network map was generated using the proteins that were identified within 10 genes upstream and downstream of >2% of the 714 hydA homologs (**Fig. 3**). This network, like that presented in **Fig. 2**, again indicates that the HydB and HydC are highly prevalent in hydA gene neighborhoods and their distribution is strongly co-correlated. Likewise, the distribution of HydD and the Hkl protein also strongly covary. Rex (redox sensing repressor protein which is involved in transcriptional redox regulation [57–59]) was commonly found in the gene neighborhood of all three groups of Hyd (thus giving it a high 'betweenness centrality' value) and may represent a general transcriptional level mechanism to regulate Hyd as a function of cellular redox status. Based on these results we categorized each HydA into 3 groups: Group 1 (G1), Group 2 (G2), and Group 3 (G3), which are differentiated primarily on the basis of the presence of HydBCD (**Fig. 4** and **Supplementary Table 2**). The Hyds identified in each group were then compared to those of biochemically characterized Hyds [21,35–37,40,60] in order to putatively assign functionality and quaternary structure.

On the basis of sequence alignments and BLASTp analyses against the CDD, HydB is predicted to bind a flavin and ligate several [4Fe-4S] clusters while HydC is predicted to ligate a [2Fe-2S] cluster. Electron paramagnetic resonance (EPR) on HydB

from *T. maritima* also indicated the presence of multiple iron-sulfur clusters and sequence analysis of HydB predicted the presence of a FMN binding site and a variable number of iron-sulfur clusters [22]. Similarly, bioinformatics analysis of HydB from *A. woodii*, *C. autoethanogenum*, *M. thermoacetica*, and *R. albus* also indicated the presence of variable number of iron sulfur clusters and a FMN binding site [21,35–37,40]. Sequence analysis of HydC from *T. maritima* predicted the presence of one [2Fe-2S] cluster [22,35]. In addition, based on our sequence analysis, HydD exhibits homology to a thioredoxin-like Fd and, based on the presence of two vicinal Cys residues, is predicted to bind a [2Fe-2S] cluster.

Hyd gene clusters that did not encode for or encoded incomplete HydB (e.g., truncated HydB lacking a flavin binding site) and HydC proteins in the hydA gene neighborhood (i.e. +/−10 genes) were classified as G1 Hyds (n = 483 HydA). Representatives of these monomeric enzymes are from the bacterium *C. pasteurianum* [2,11,14,16,34,61] and the alga *C. reinhardtii* [14,16,27,29,30]. These have been subjected to extensive biochemical analysis and are prototypical examples of this group. HydA from *C. pasteurianum* [2,11,36] and *C. reinhardtii* [27,29,30] have shown to produce H_2 using ferredoxin as the only electron donor and are thus unlikely to bifurcatei. These findings suggest that the other G1-group Hyds also likely do not bifurcate, including the previously characterized ‘dimeric’ Hyd from *D. desulfuricans* (note the presence of truncated HydB which lacks a flavin binding site [23,44]). Of the 483 G1 HydAs, the gene neighborhood of seven homologs encoded for HydB, HydC, or HydD or a combination of these, but they did not meet the classification scheme for complete

HydBC in their gene neighborhoods. For example, of these seven Hyds, one encoded for only HydB, three encoded for only HydC, one encoded for only HydCD, and two encoded for only HydBD. Based on available biochemical data, it is not clear how or if these proteins interact with HydA. At this point, it is not known whether these Hyds can bifurcate since all biochemically characterized bifurcating enzymes encode for either HydABC or HydABCD complexes [21,35–37,40,41].

Hyds that encode HydBC in the flanking regions of hydA were categorized as G2 enzymes (n = 155 HydA). Of these 155 Hyds, 92% shared a common pattern of HydABC encoded in the ABC orientation (**Fig. 4**). Representatives of this Hyd group have been purified from *M. thermoacetica* and *T. maritima*, which were biochemically characterized as trimeric complexes (i.e. HydABC) [35,36]. In these examples, HydB is the subunit that has been predicted to contain a unique flavin binding site along with variable numbers of iron-sulfur clusters while HydC contains a single [2Fe-2S] cluster [35,36]. Both of these trimeric Hyds have been shown to bifurcate electrons, indicating that other G2 Hyds are likely capable of bifurcation [35,36]. Both Fd and NADH are necessary for bifurcating Hyd activity and it has been proposed that HydC, which contains a Fd-like [2Fe-2S] cluster, accepts electrons from Fd while HydB accepts electrons from NADH [35]. If verified, this suggests the importance of both subunits HydBC for Hyd and bifurcating capability. Furthermore, the biochemical study conducted on the trimeric G2 Hyd from the thermophile *T. maritima* indicated that when HydBC were separated from HydA, the HydA enzyme lost its stability at elevated temperature and also lost the ability to reduce

anthraquinone-2,6-disulfonate with H_2 [22]. This indicates that these two HydBC subunits not only play a role in electron transfer but also function to stabilize HydA.

In addition to HydBC, genes flanking certain G2 hydA were found to encode electron transfer proteins (etp) (n = 2), formate dehydrogenase D (fdhD) (n = 2), a homolog of HydA termed HydS (n = 17), and Sp (n = 28). HydS contains an additional sensory domain (i.e. PAS [45]) at its C-terminus that is not present in canonical HydA. These results are consistent with previous informatics studies which revealed the genes flanking *hydA* included *hydS* and *sp* in *C. thermocellum* DSM 1313 [37,62] and *T. maritima* [35,37]. In addition, ETP and FdhD, the latter of which has a $NADP^+$ binding site, have been shown to form a complex with G2 Hyds in *C. autoethanogenum* [40]. This multimeric complex is involved in the reversible reduction of CO_2 in the presence of H_2 to formate and also in the reversible reduction of Fd and $NADP^+$using formate. Interestingly, six of the 155 G2 Hyds were found to have fused HydB and HydC subunits and one G2 Hyd (*Nyctotherus ovalis*; AAU14235) was found to have fused HydABC subunits. These observations provide additional evidence that these three proteins function synergistically, as has been observed in numerous other cooccurring enzyme systems, including Hyd maturases [1,63–67].

Lastly, hydA-containing gene clusters that encode for HydBCD were classified into G3 enzymes (n = 76). Interestingly, 71% of the G3 Hyds (58 of 76) also encoded an Hkl protein that was present within the hyd gene cluster. The identity of Hkl proteins was confirmed based on the presence of a HATPase_c motif (PF02518) which is the ATP binding site of the ATPase domain [68,69]. This motif composition differs from

canonical histidine kinase proteins. Canonical histidine kinase typically has an additional two motifs including the HAMP motif [histidine kinase, adenylyl cyclase, methyl-accepting protein and phosphatase (PF00672)] which is a phosphoacceptor domain that acts as a cytoplasmic intermediate linker connecting the phosphoryl transfer (HisKA) motif [68,70–74] with the sensor domains outside of cytoplasm [70,75,76]. Biochemical characterization of the G3 Hyd in *A. woodii* indicates that it forms a tetrameric complex that includes HydBC and a Fd-like HydD protein that contains a [2Fe-2S] cluster [21].

Like representatives of G2 Hyds, the tetrameric HydABCD enzyme of *A. woodii* has been shown to bifurcate and also to require two electron carriers, Fd and NADH, for activity [21]. However, unlike the trimeric G2 Hyd identified in *T. maritima* [35], which is involved in fermentative H_2 production, the tetrameric G3 Hyd identified in *A. woodii* is involved in oxidation of H_2 [35]. It has been proposed that electrons are transferred from HydA to HydC and then to HydB where the electrons are bifurcated simultaneously to reduce Fd and NAD^+ [21]. This mechanism, like that proposed for trimeric G2 Hyd [35], implies an important role for HydBC in electron transfer. The function of HydD in G3 Hyd and its role in bifurcation and in reaction directionality is still unknown. Importantly, in several instances, HydBCD were not present in this arrangement, but rather, were interspersed among other genes in the 20 protein encoding genes flanking G3 hydA. Examples include Hyd from *Desulfarculus baarsii* DSM 2075 (Deba_1523), *Desulfobacterium autotrophicum* HRM2 (HRM2_16550), and *Kosmotoga olearia* TBF 19.5.1 (Kole_0172). Thus, it is unclear if these Hyds have the capability to bifurcate. In addition to HydBCD and Hkl, genes flanking a number of G3 Hyds encoded various

regulatory genes including Stk proteins (n = 39 HydA), and/or transpeptidase proteins (n = 20 HydA).

In support of the group designations identified by Apriori algorithm classification, PCA ordination of a matrix describing the dissimilarity in the composition of proteins encoded by genes in flanking regions of hydA formed clusters that largely corresponded to the same group designations (**Supplementary Fig. 1**). A network map was also used to confirm the extent to which proteins encoded by genes in flanking regions of hydA correlated with group designation (**Fig. 5**). Based on the Apriori analysis, regulatory genes including Sp and HydS were enriched in G2 Hyds while Hkl, DRTGG, a PHP domain-containing protein and Stk were enriched in flanking regions of G3 Hyds. The difference in regulatory gene association between the two groups indicates that G2 and G3 Hyds may be subject to different modes of PTM and/or that these PTM enzymes exhibit specificity for target Hyd proteins allowing for their differential regulation if both types of protein are present in the same organism.

The HydA copy number per genome varied as a function of taxonomy (**Table 1**). Moreover, the HydA copy number per genome, when broken out by group designation, also varied (**Supplementary Fig. 2A**). Multiple copies (isoforms) of G1 Hyd in the same genome was the most common pattern identified (n = 117 of 287 genomes). Several genomes contained three or more copies of HydA, with a maximum copy number/genome of seven HydAs (*Desulfotomaculum carboxydivorans* CO-1-SRB, DSM 14880). Fifty-two percent of the HydA-encoding genomes encoded for at least two different types of Hyd (**Supplementary Fig. 2B**). For example, 77 genomes encoded for

G1 and G2 Hyds, 44 encoded for G2 and G3 Hyds, and 17 genomes encoded for all three groups of Hyds. The co-existence of G2 and G3 Hyds with G1 Hyds potentially suggests they have different functional roles within the organisms and again is consistent with data presented above indicating that they may be regulated at the post-translational level by different PTM enzymes.

Hyd variants were distributed unequally among bacterial phylum level lineages. G1 Hyds were prevalent in the Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes and Thermotogae phyla while G2 Hyds were enriched in Choloroflexi and Thermotoga phyla. G3 Hyds were enriched in Firmicutes and Spirochaetes phyla (**Table 1**). Two of the three eukaryotic phyla that contained homologs of Hyd encoded G1 Hyds: heterokonts with one genome (i.e. *Thalassiosira pseudonana*) and chlorophytes with 3 genomes. One of the three eukaryotic phyla encoded a G2 Hyd: Ciliophora with one genome (i.e. *Nyctotherus ovalis*). Multiple copies of HydA were often encoded in chlorophyte genomes and these were all classified as G1 Hyd. Interestingly, HydA identified in the ciliate *Nyctotherus ovalis* [77–79] forms a fused complex with HydB-like and HydC-like subunits, which begs the question as to whether this enzyme can also bifurcate electrons as it meets the minimal requirement of gene complements (i.e., encodes HydABC) hypothesized to afford bifurcation activity.

3.2 Structural composition of HydA (H-, F- and C- cluster).

Various biochemical [2,11,29,30] and bioinformatics studies [14,16,32,34,61] indicate substantial variation in the iron-sulfur cluster content of the F- and C-clusters of

HydA, which are hypothesized to act as electron transfer conduits for the various Hyds. Thus, the composition of iron-sulfur cluster binding domains in the F- and C-clusters of HydA was characterized in order to determine if variations in these clusters correlated with Hyd group designations. Based on the presence or absence of cluster binding motifs and other identifiable motifs defined in the CDD, the 714 HydA homologs were categorized into modular structures according to the scheme of Calusinska et al. [14]. A total of 12 modular structures were identified and these were nearly equally represented across all groups of Hyd (**Fig. 6**). Out of 12 modular structures, 10 were predicted to ligate two [4Fe-4S] clusters in the F-cluster domain. More than 50% of HydA in G2 and G3 exhibited the M2i modular structure which were predicted to ligate two [4Fe-4S] clusters or the M3a modular structure which comprised multiple iron-sulfur clusters (i.e. [2Fe-2S], histidine substituted [4Fe-4S] and 2[4Fe-4S]) in the F-cluster domain. Likewise, more than 50% of the HydA lacked C-clusters and of those that did have C-clusters, most were classified as the non-bifurcating G1 Hyds.

We also examined the variation in the three amino acid sequence motifs (i.e. L1, L2 and L3 [14,16]) that are involved in ligating the H-cluster in relation to group designations (Fig. 7). The majority of the alignment positions were comprised of residues that did not vary substantially in amino acid usage for any single group. However, two positions in each of the three motifs contained residues that differed among groups. For example, in the L1 motif region, the 4th residue, cysteine, is more common in G1 and G2 Hyds while serine is more common in G3 Hyds. Similarly, glycine is commonly observed in the 6th position of the L1 motif in G2 and G3 Hyds, whereas alanine is more

commonly observed at this position in G1 Hyds. Within the L2 motif region, methionine is more common in the 1st position in G2 and G3 Hyds, whereas glycine is more common in G1 Hyds. Within the L3 motif region, glycine and valine in the 2nd position is more common in G1 Hyds while valine and isoleucine are more common in G2 and G3 Hyds. Finally, proline and glycine are more common in the 13th position of the L3 motif in G1 Hyds whereas glycine is more common in G2 and G3 Hyds.

We also examined the variation in the three amino acid sequence motifs (i.e. L1, L2 and L3 [14, 16]) that are involved in ligating the H-cluster in relation to group designations (**Fig. 7**). The majority of the alignment positions were comprised of residues that did not vary substantially in amino acid usage for any single group. However, two positions in each of the three motifs contained residues that differed among groups. For example, in the L1 motif region, the 4th residue, cysteine, is more common in G1 and G2 Hyds while serine is more common in G3 Hyds. Similarly, glycine is commonly observed in the 6th position of the L1 motif in G2 and G3 Hyds, whereas alanine is more commonly observed at this position in G1 Hyds. Within the L2 motif region, methionine is more common in the 1st position in G2 and G3 Hyds, whereas glycine is more common in G1 Hyds. Within the L3 motif region, glycine and valine in the 2nd position are more common in G1 Hyds while valine and isoleucine are more common in G2 and G3 Hyds. Finally, proline and glycine are more common in the 13th position of the L3 motif in G1 Hyds whereas glycine is more common in G2 and G3 Hyds.

The lack of consistency between F-, C-, and H-cluster and Hyd group designations indicates that bifurcation capability is more likely to be predicated by the

presence/absence of accessory subunits, in particular the flavin binding subunit HydB. Moreover, these results may suggest that variations in F-, C-, and H-cluster domains are refinements to enzymes that are primarily under selection for bifurcating versus non-bifurcating abilities. Biochemical characterization of Hyds from *A. woodii, C. autoethanogenum, M. thermoacetica* and *T. maritima* all showed the ability to bifurcate and all the Hyds identified in these species were either trimeric or tetrameric [21,35,36,40]. Conversely, the monomeric Hyd from *C. pasteurianum* [2, 11, 36] and *C. reinhardtii* [27, 29, 30] indicate that they are unable to bifurcate electrons. These findings suggest that the other G1-group Hyds also likely do not bifurcate, including the 'dimeric' Hyd from *D. desulfuricans* which forms a complex with a truncated HydB that lacks a flavin binding site (Malki *et al*, 1995; Nicolet *et al*, 1999). A few additional putative dimeric forms of Hyd were identified that contained full length HydB with a flavin binding motif, but due to lack of biochemical characterization of an enzyme with this configuration, they were classified as G1 enzymes. For example, the HydA encoded in the *Coprococcus catus* GD/7 genome (locus tag: CC1_10640) also encoded for HydB in its flanking region but lacked genes encoding for HydC. Hence, based on available biochemical data, we hypothesize that the enzymes that lack either of the subunits i.e. HydB or HydC do not have the capability to bifurcate.

3.3 Post translational modification of bifurcating G2 and G3 Hyds

Among the more intriguing and novel findings from our bioinformatics work was the identification of genes encoding proteins that could be involved in PTM in the gene

neighborhoods of G2 and G3 Hyds. Identified proteins included Hkl and Stk in G3 Hyds while Sp was commonly observed in G2 Hyds. Despite both having an ATPase domain and exhibiting ~25% homology with ~40% pairwise sequence coverage, an alignment of our identified Hkl and Stk proteins indicated they are distinct (**Supplementary Figure 3**). Protein kinases have been a subject of extensive study which have demonstrated their involvement in various regulatory processes. Among those most widely studied are Hkl and Stk which have been biochemically shown to target histidine (H) and serine/threonine (S/T) for phosphorylation, respectively [70, 80-83]. The presence of Hkl and Stk in the gene neighborhood of G3 *hydA* indicate that these kinases and phosphatases may be involved in reversible phosphorylation of H, S and T residues. By extension, this may suggest conservation of these residues in Hyd subunits.

Multiple sequence alignments (MSA) of all the sequences for the Hyd subunits, i.e. HydA, HydB, and HydC from G1, G2, and G3, revealed few or no conserved (i.e. ≥ 90% conservation) H, S, and T residues. We therefore screened alignments of only HydA and its subunits in gene clusters that also encode Sp (G2), Stk (G3), and Hkl (G3) in the flanking region, which resulted in a total of 18 and 46 sequences, respectively. When HydABC that were not flanked by Hkl or Stk were removed from MSA, several conserved H, S, and T were identified (**Supplementary Figure 4**, **5**, and **6**). HydB subunits also harbored conserved (i.e. ≥ 90% conservation) H and S/T residues regardless of whether they were classified as G2 or G3 Hyds (**Supplementary Figure 5**). HydC in G2 harbored multiple conserved (i.e. ≥ 90% conservation) H and T residues while G3 harbored multiple conserved T residues only at the same position (**Supplementary**

Figure 6). Lastly, HydD in G3 also harbored conserved (i.e. ≥ 90%) H and T residues (**Supplemental Figures 7**). However, the extent of conservation of H, S, and T residues in HydD did not vary substantially in enzymes that were not flanked by Hkl or Stk versus those that were flanked by these proteins. While the presence of conserved H and S/T residues does not, by itself, indicate that those conserved sites are post-translationally modified, they do provide potential targets for *in vitro* studies. To provide further evidence for PTM of Hyd, we undertook a biochemical assay to probe PTM using the G2 Hyd from *T. maritima* and the G3 Hyd from *Caldicellulosiruptor bescii*.

Sequence analysis of the G2 HydA from *T. maritima* indicates that it comprises multiple iron-sulfur cluster (i.e. 2[Fe-2S], histidine substituted [4Fe-4S] and 2[4Fe-4S]) binding motifs in the F-cluster and a [2Fe-2S] cluster binding motif in the C-cluster and thus is classified within the 3c modular structure (**Fig. 6**). The structural proteins identified in *T. maritima* G2 Hyd (i.e. HydABC) and a homolog of Sp (TM1423) have been identified in the flanking region of the trimeric Hyd complex (TM1424-TM1426). Similarly, sequence analysis of HydA from *C. bescii* indicates that it also contains multiple iron-sulfur cluster (i.e. 2[Fe-2S], histidine substituted [4Fe-4S] and 2[4Fe-4S]) binding motifs in its F-cluster and is thus classified as having the M3a modular structure, which is a common structure associated with G3 Hyds (**Fig. 6**). The structural proteins identified in *C. bescii* G3 Hyd (i.e. HydABCD) are encoded in an apparent operon (YP_002573168 to YP_002573172). Homologs of Hkl (YP_002573169) and Stk (YP_002573164) were also identified in the gene neighborhood of *hydA* in *C. bescii*. The presence of Hkl and Stk in the flanking gene neighborhood of *hydA* in *C. bescii* and Sp in

the flanking region of *hydA* in *T. maritima* suggests that phosphorylation of these Hyd proteins may occur.

To test the possibility for PTM, purified Hyd protein complexes from *T. maritima* and *C. bescii* were analyzed by SDS-PAGE followed by phospho-specific protein staining. Image analysis showed strong signals for HydA/B proteins (68/72 kDa for *T. maritima* and 63/64 kDa for *C. bescii*) as well as for HydD (14 kDa) from *C. bescii* (**Fig. 8**, left panel). Positive (casein, Cas from *Bos taurus*) and negative (transhydrogenase, Nfn1 from *Pyrococcus furiosus*) controls for phospho-staining were included on the gel for reference. To further establish the presence of phosphorylation, protein samples were treated with calf intestinal phosphatase (CIP). A reduction in phosphorylation intensity was observed for high molecular weight Hyd subunits from *T. maritima* and casein. Incubation of *C. bescii* Hyd complex with CIP did not lead to a significant change in fluorescence. The same gel was then stained with Coomassie Blue to verify protein loading (**Fig. 8**, right panel). *T. maritima* HydC (18 kDa) was poorly visualized, as observed by the less than stoichiometric band intensity. Coomassie stain showed that HydC (17 kDa) and HydD (14 kDa) from *C. bescii* were present at an equal ratio, as expected, in contrast to their signal when visualized for phosphorylation.

Mass spectrometry was used to further investigate Hyd subunit modification. High-resolution mass analysis of intact proteins is a sensitive method for observing heterogeneity. Purified Hyd complexes were subjected to size-exclusion and reverse-phase chromatography to remove salts and buffers prior to electrospray time of flight mass analysis. All Hyd subunits, except for HydC from *C. bescii*, were observed as a

mixture of several predominant forms. HydB from *T. maritima* showed significant heterogeneity, with all of the forms being 300-400 Da greater than the expected molecular weight of 68,551 Da, based on the amino acid sequence (**Fig. 9,** top panel). HydC from *T. maritima* had three distinct forms separated by 98 Da (**Fig. 9,** middle panel). Analysis of *C. bescii* proteins showed that HydD was present as two dominant forms (1:3 ratio), separated by 80 Da (**Fig. 9,** bottom panel).

Phosphorylation is often difficult to confirm by mass spectrometry due to sub-stoichiometric modification and loss of the attached phosphate during sample processing and electrospray ionization. Attachment of a phosphate group (HPO_3) adds 80 Da to the protein for each phosphorylated amino acid [84-87]) as we observed for HydD from *C. bescii*. In contrast, a difference 98 Da (H_3PO_4) is indicative of an ionization induced loss of phosphorylation specific to serine and threonine residues [86, 88, 89]. This result suggests modification of serine and threonine residues of *T. maritima* HydC, although phosphorylation at the peptide level will be necessary to confirm the specific sites of phosphorylation. Together, the SDS-PAGE and LCMS analysis strongly suggest that HydAB and HydC of *T. maritima* G2 Hyd are phosphorylated. Along with our bioinformatics-based analysis, this provides strong evidence that Hyd proteins are likely to be post-translationally modified. This is an attractive mechanism for regulating the activity of bifurcating G2 Hyd complexes. Similarly, SDS-PAGE and LCMS analysis strongly suggest that the HydAB and HydD subunits of *C. bescii* G3 Hyd are phosphorylated. Interestingly, HydC in *C. bescii* G3 Hyd was not modified. Since G3 HydC lacked conserved H residues but G3 HydABD all contained a number of conserved

H residues, this result suggests that Hkl is potentially responsible for phosphorylating the H residues in G3 HydA. Future biochemical studies are needed to identify the specific sites that are putatively involved in phosphorylation activity, their effects on modulating hydrogenase activities, and the specific enzymes involved in PTM activities.

4. CONCLUSIONS

Our bioinformatics-based analysis of the gene neighborhood of hydA revealed three major groups of Hyd (G1, G2, and G3) that are differentiated on the basis of the presence of HydBC (G2) or HydBCD (G3). The G1 enzymes appear to be monomeric and based on biochemical data are predicted to not bifurcate while G2 and G3 Hyds are in trimeric and tetrameric forms, respectively. Based on bioinformatics analysis and albeit limited biochemical data, G2 and G3 enzymes are predicted to bifurcate. Variation in the F- and C-clusters and the residues that ligate the H-cluster of HydA do not vary in a systematic manner with respect to Hyd group designations, indicating that they are under different selection and unlikely to delineate the ability of enzymes to bifurcate electrons.

Our analysis also identified regulatory genes in the flanking regions of *hydA*, including Sp (in G2 Hyds), Stk (in G3 Hyds), and an Hkl (in G3 Hyds), all of which we hypothesize to play a role in PTM regulation of Hyds. Indeed, our MS and biochemical data indicate that examples of G2 and G3 Hyds are phosphorylated alluding to the role of the aforementioned proteins in potentially regulating Hyd. That PTM proteins associate with G2/G3 Hyd but not G1 Hyd is consistent with the prevalence of G1 Hyd in genomes

that encode for G2/G3 Hyd, necessitating mechanisms to differentially regulate the activities of these enzymes. Moreover, the presence of different PTM proteins in the gene neighborhoods of G2 and G3 Hyds and differences in the protein subunits that were apparently phosphorylated is consistent with the co-occurrence of G2 and G3 Hyd in the same organisms and their need to differentially regulate these enzyme complexes. Collectively, these results show that the vast inferred structural diversity of Hyds can be classified into only three fundamental groups, and variations within them allude to their various functions in microbial metabolism. The classification scheme reported herein provides the first framework for prioritizing biochemical and mutagenesis studies of Hyds aimed at furthering our understanding of the functional role of this surprisingly cohesive group of enzymes.

5. ACKNOWLEDGEMENTS

This work was supported as part of the Biological Electron Transfer and Catalysis Energy Frontier Research Center funded by the U.S. Department of Energy, Office of Science, and Basic Energy Sciences under Award # DE-SC0012518.

Supplementary Table 1. The 32 unique combinations of genes that flank *hydA* identified using frequent pattern mining algorithm, as conducted with the Apriori algorithm. Only those patterns that had a confidence value of 0.95 or higher are presented.

Supplementary Table 2. Complete database including proteins encoded in the +/- 10 flanking gene regions of *hydA* (n=714), F- and C- cluster composition of HydA, and HydA H-cluster (L1, L2, and L3) residue composition.

Table 1. Taxonomic distribution of HydA homologs in 2921 complete archaeal, bacterial, and eukaryal genomes. The total number of genomes for each phylum is indicated along with the number of genomes within that phylum that encode for HydA. The average number of HydA in genomes that encode HydA is indicated along with the number of HydA in each of the 3 defined Hyd groups (see results and discussion section for a description of the group designations).

Domain	Phylum	Total number of genomes	Total number of genomes with HydA	Average # of HydA in the genomes	Total number of Hyd in each group		
					Group 1	Group 2	Group 3
Archaea	Crenarchaeota	50	0	0	0	0	0
	Euryarchaeota	110	0	0	0	0	0
	Korarchaeota	1	0	0	0	0	0
	Nanoarchaeota	1	0	0	0	0	0
	Thaumarchaeota	5	0	0	0	0	0
Bacteria	Acidobacteria	7	0	0	0	0	0
	Actinobacteria	301	6	1.5	9	0	0
	Aquificae	14	1	1	1	0	0
	Armatimonadetes	2	0	0	0	0	0
	Bacteroidetes	94	10	1.5	11	2	2
	Caldiserica	1	0	0	0	0	0
	Chlamydia	109	0	0	0	0	0
	Chlorobi	11	0	0	0	0	0
	Chloroflexi	24	8	1	0	8	0
	Chrysiogenetes	1	0	0	0	0	0
	Cloacimonetes	1	1	1	0	0	1
	Cyanobacteria	79	0	0	0	0	0
	Deferribacteres	4	0	0	0	0	0
	Deinococcus-Thermus	21	0	0	0	0	0
	Dictyoglomi	2	2	2	2	2	0
	Elusimicrobia	1	1	1	0	1	0
	Fibrobacteres	2	0	0	0	0	0
	Firmicutes	622	165	2.9	347	92	52
	Fusobacteria	11	2	4.5	4	5	0
	Gemmatimonadetes	1	0	0	0	0	0
	Ignavibacteriae	2	2	3.5	2	5	0
	Nitrospirae	4	1	2	2	0	0
	Planctomycetes	8	0	0	0	0	0
	Proteobacteria	1240	39	1.5	44	7	6
	Spirochaetes	65	22	2.5	31	14	11
	Synergistetes	5	2	1	0	1	1
	Tenericutes	89	0	0	0	0	0
	Thermodesulfobacteria	2	1	4	4	0	0
	Thermotogae	18	17	2.3	20	16	3
	Verrucomicrobia	5	1	1	0	1	0
Eukaryota	Heterokonts	1	1	1	1	0	0
	Chlorophyta	3	3	1.7	5	0	0
	Ciliophora	1	1	1	0	1	0
	Ascomycota	14	0	0	0	0	0
	Microsporidia	1	0	0	0	0	0
	Euglenozoa	1	0	0	0	0	0
	Mognoliophyta	1	0	0	0	0	0

	Chordata Nematoda	2 1	0 0	0 0	0 0	0 0	0 0

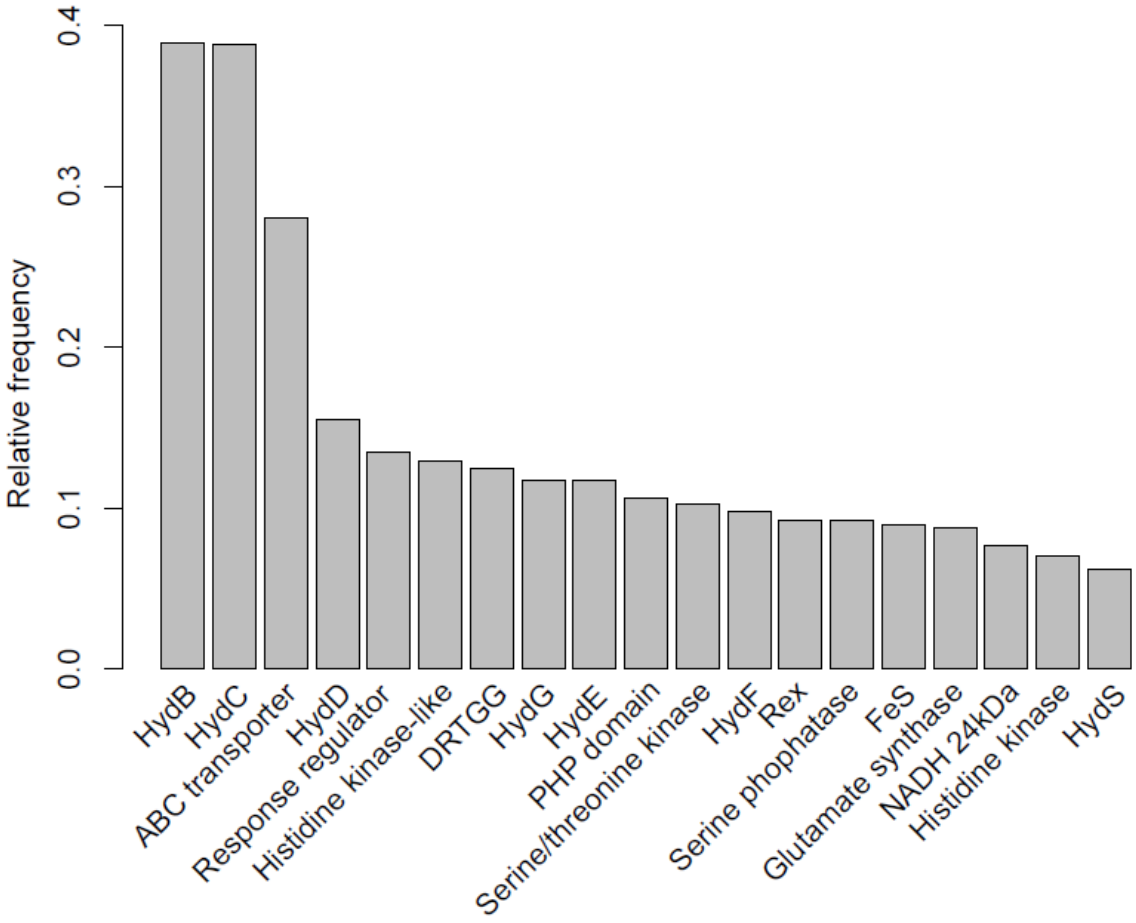


Fig. 1. Relative frequency of encoded proteins that were identified among genes flanking 714 hydA homologs (+/−10 genes). For simplicity, only those proteins that were encoded in ≥6% of the 714 gene neighborhoods are shown. The identity of all proteins that flank each individual HydA can be found in **Supplementary Table 2**.

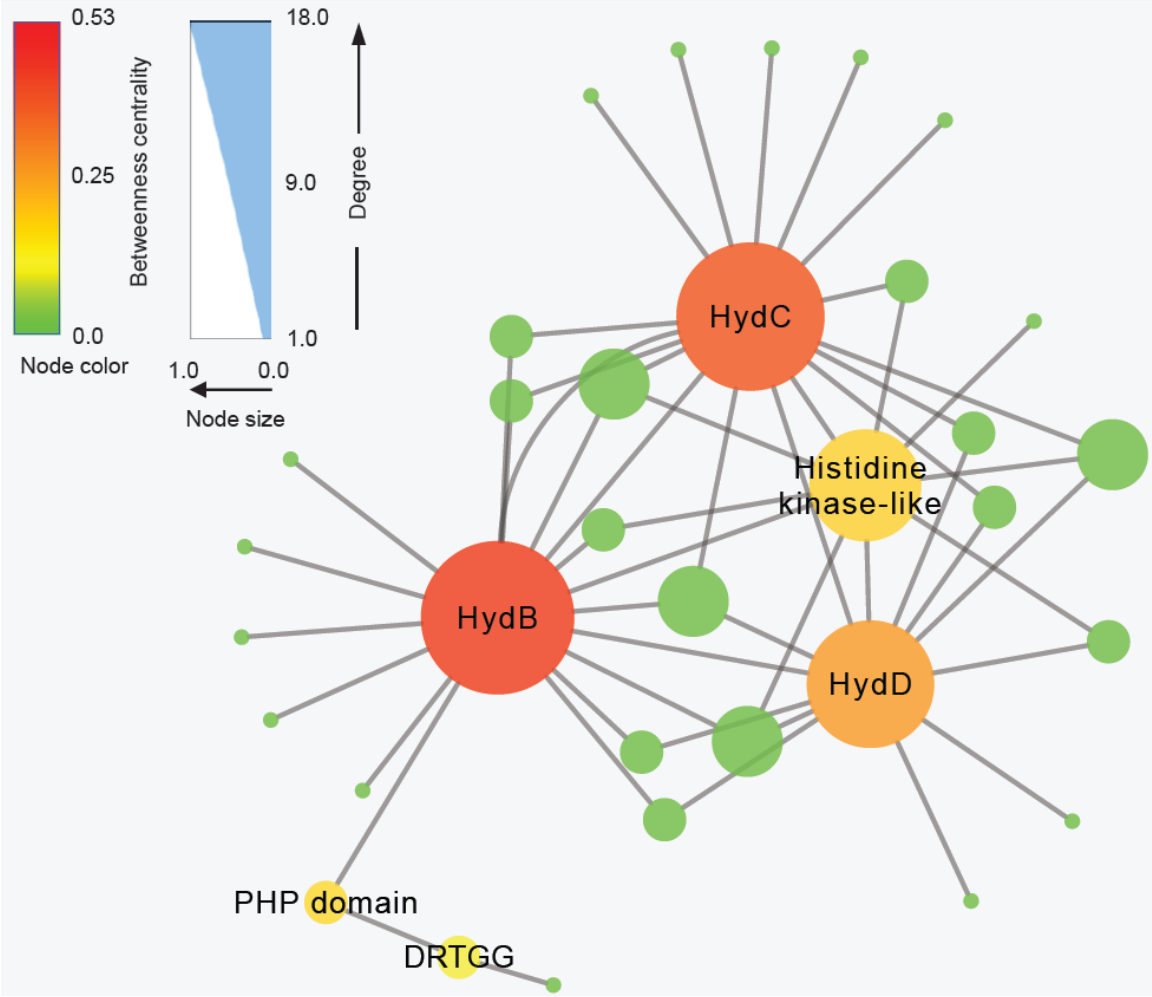


Fig. 2. Network map depicting the dominant patterns in proteins encoded in the genes that flank hydA (+/−10 genes), as identified by the Apriori algorithm (18). Each unique pattern is given as a node and the edges between nodes represent significant correlations between those nodes (confidence value of ≥0.95). Node color represents the betweenness centrality (a measure of the 'connectedness' of each gene to the total network) that connects multiple groups (i.e., are shared in multiple groups). Node size represents the number of patterns/genes significantly associated with other patterns/genes. The force directed organic layout within Cytoscape was used to visualize the correlation in the network. Patterns in the distribution of genes flanking hydA, as revealed by this network, were used to classify the 714 HydA homologs into 3 categories (described in text). For simplicity, only nodes with betweenness centrality of 0.06 or greater are labeled.

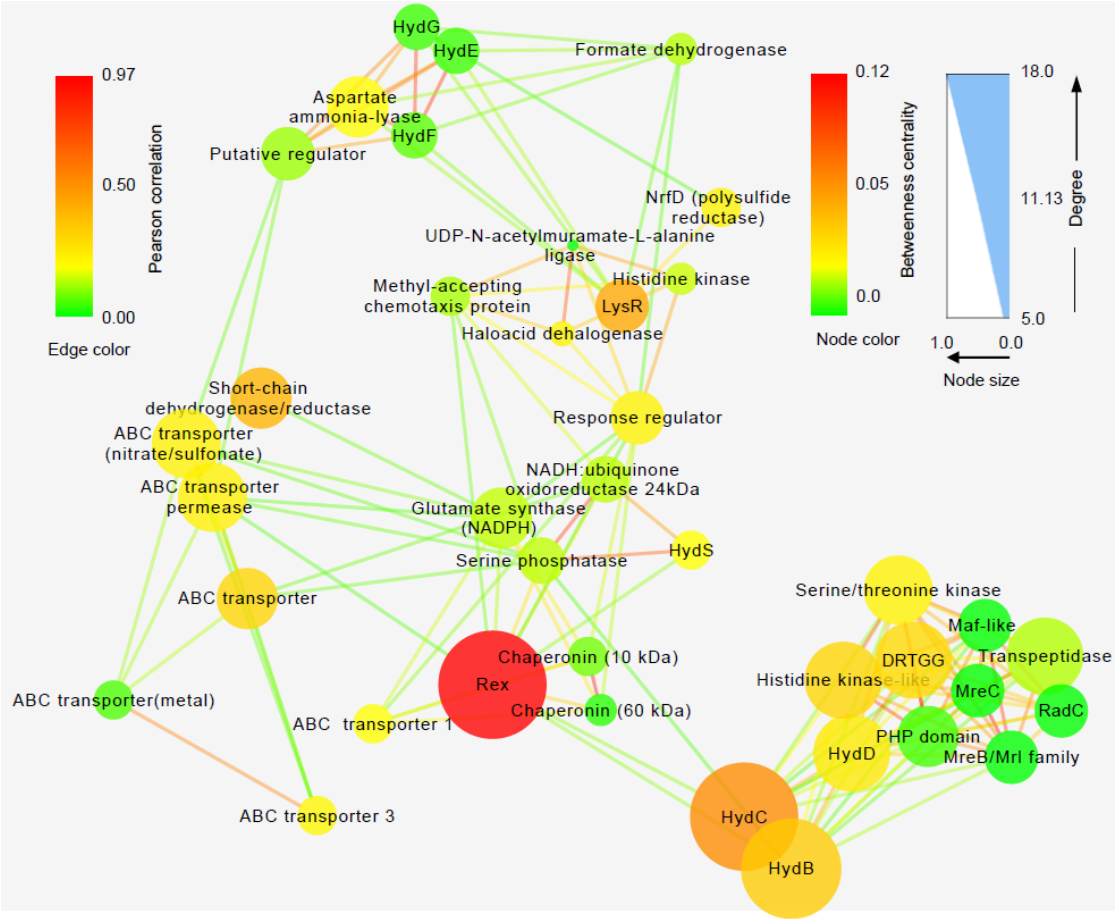


Fig. 3. Network map depicting the correlation of proteins encoded by genes flanking (+/−10 genes) hydA. Each unique protein is given as a node and the edges between nodes represent significantly correlated proteins ($p < 0.05$). Node color represents the betweenness centrality that connects proteins (a measure of the 'connectedness' of each gene to the total network) while edge color represents the Pearson correlation describing the co-distribution of proteins encoded in flanking regions of hydA. Node size represents the number of proteins encoded by genes flanking hydA that exhibit significant associations with other proteins in these flanking regions.

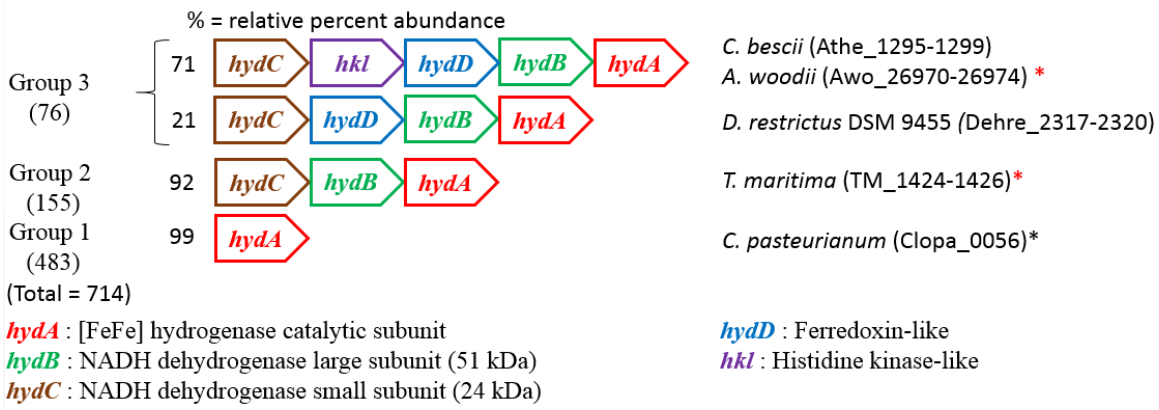


Fig. 4. Classification of HydA (n = 714) homologs based on the composition of proteins encoded in flanking gene neighborhoods (+/−10 genes). The number of HydA homologs that fall within each group is indicated in parentheses. The percent of HydA homologs within a specified group that exhibit the exact gene neighborhood as depicted is given as a percentage. An organism and the locus tag for a representative gene cluster for each specified group are indicated. Asterisks denote Hyd homologs that have been biochemically characterized (see Results and discussion section for details). Those with a red asterisk have been biochemically shown to bifurcate while those with a black asterisk have been biochemically shown to not bifurcate.

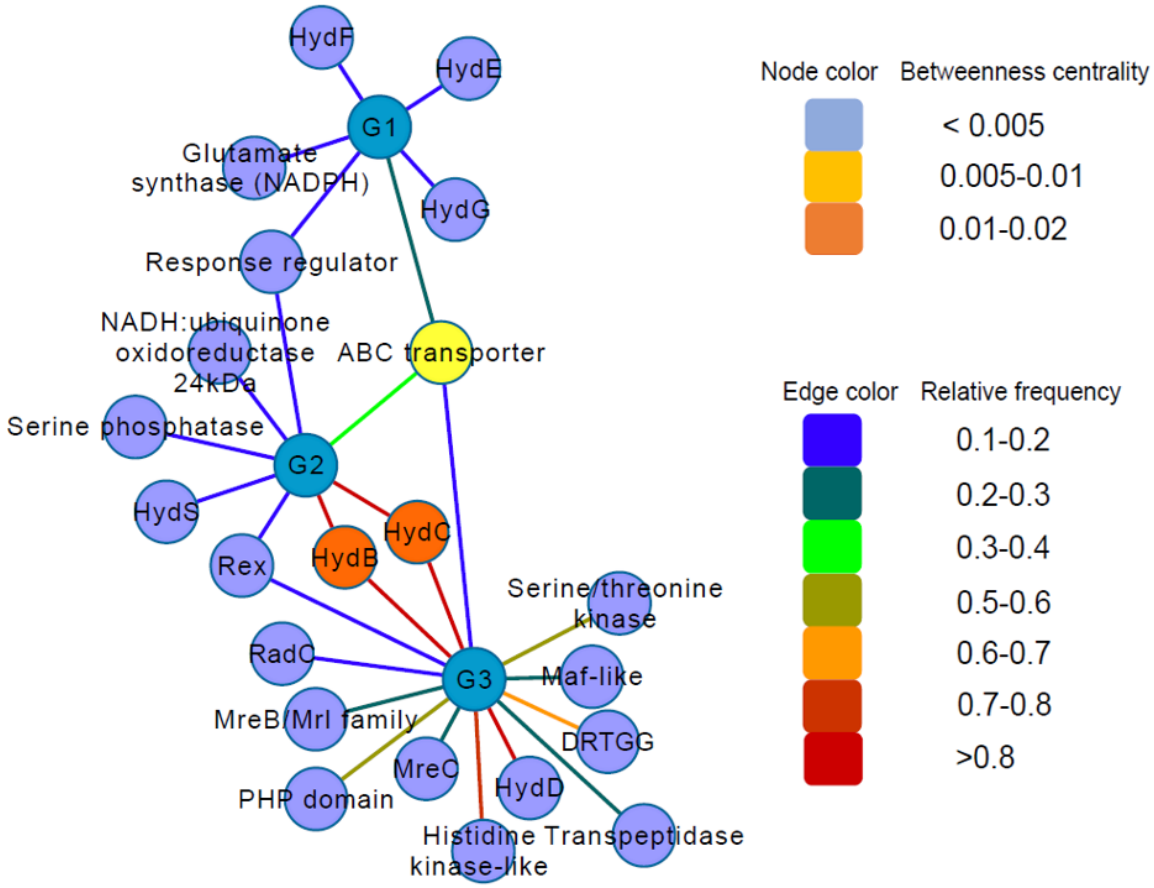


Fig. 5. Network analysis of proteins encoded by genes flanking (+/−10 genes) hydA. Only proteins (n = 22) encoded in the flanking regions of >10% of the each HydA group (i.e. relative frequency of >10%) were considered in this analysis. Here, node color represents the betweenness centrality (a measure of the 'connectedness' of each gene to the total network) while edge color represents the relative frequency of each protein in each group. The force directed organic layout was used to visualize the correlation in the network. The average pattern associated with each of the 3 defined Hyd groups was included in this analysis [indicated by Group 1 (G1), Group 2 (G2), and Group 3 (G3)].

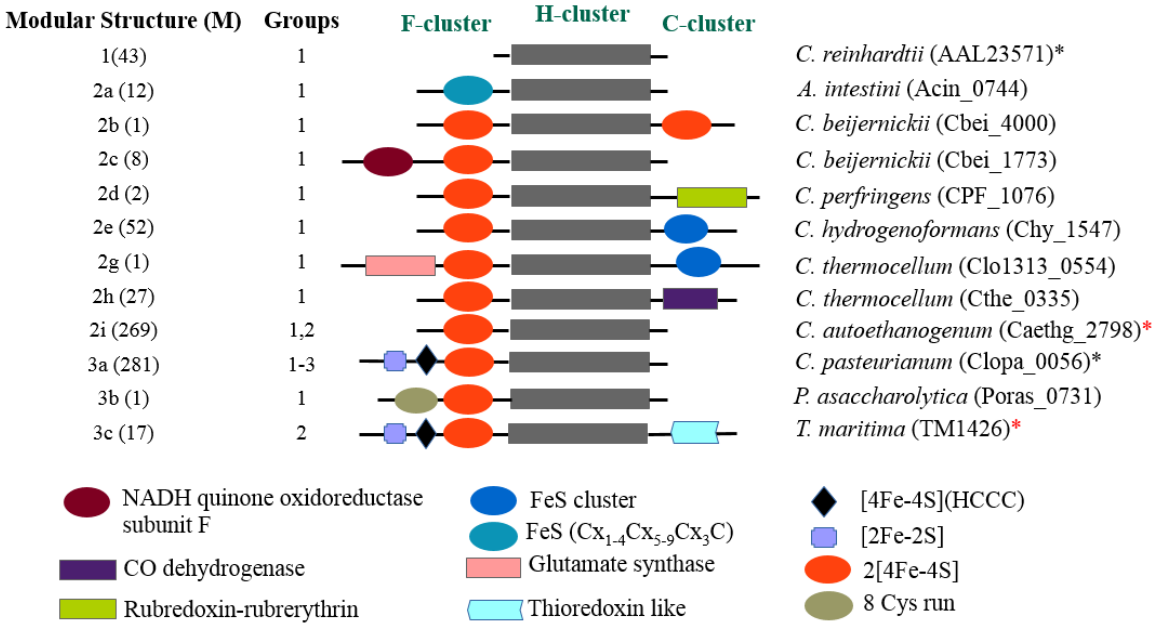


Fig. 6. Inferred domain composition of the F- and C-clusters associated with HydA homologs. Homologs and modular structures are organized by Hyd group designations given on the left side of the figure. The number in the parenthesis next to the modular structure designation represents the number of HydA homologs out of 714 that have the specified modular structure. An example of a HydA in an organism and its associated locus tag with each modular structure is indicated on the right side of the figure. Hyd homologs that have been biochemically characterized are denoted with an asterisk. Hyd homologs with a red asterisk have been biochemically shown to bifurcate while those with a black asterisk have been biochemically characterized but are not known to bifurcate.

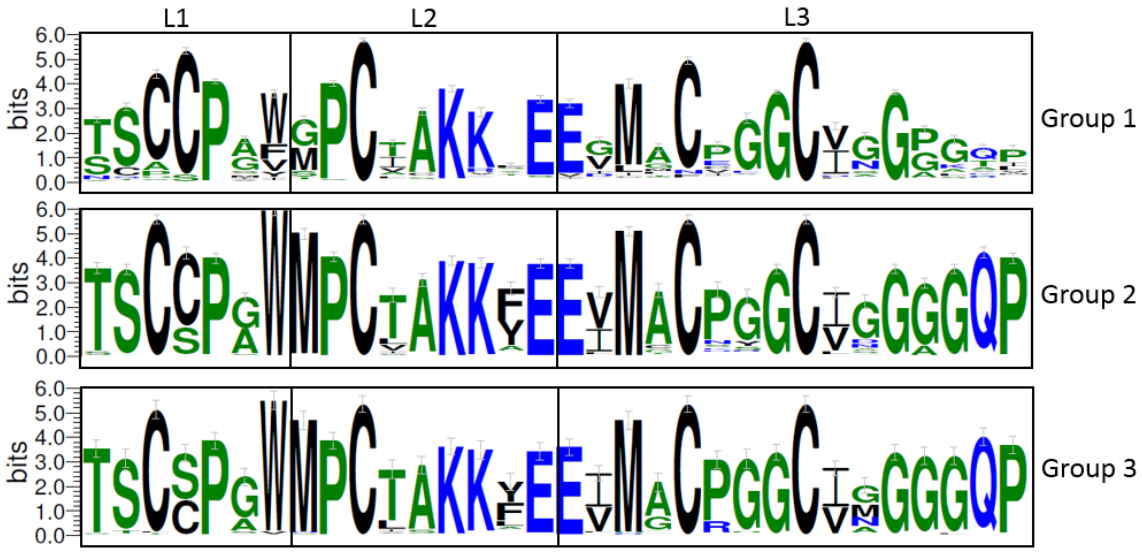


Fig. 7. Schematic diagram indicating the extent of conservation in residues comprising the H-cluster motifs (L1, L2, and L3, as outlined by [11]) in HydA as organized by Hyd group

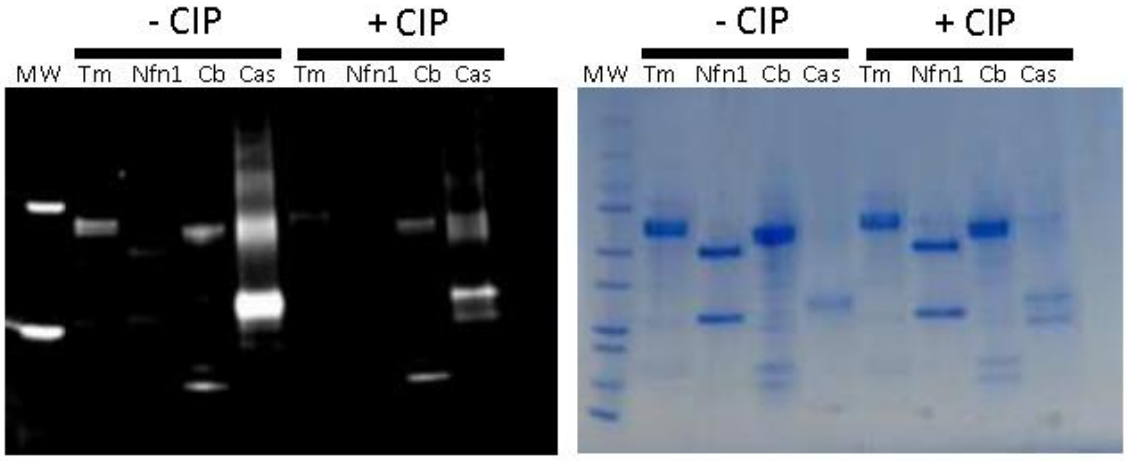


Fig. 8. Phosphoprotein-specific staining. Purified Hyd were separated using SDS-PAGE and stained with a fluorescent dye that is specific for phosphoproteins (left panel) followed by Coomassie staining (right panel) to observe protein abundance. Samples were run with and without treatment with calf intestinal phosphatase (CIP). The G2 Hyd from *Thermotoga maritima* (Tm) and the G3 Hyd from *Caldicellulosiruptor bescii* (Cb) were positive for phosphorylation. Casein (Cas) from *Bos taurus* and transhydrogenase (Nfn1) from *Pyrococcus furiosus* served as positive and negative controls for phosphorylation, respectively.

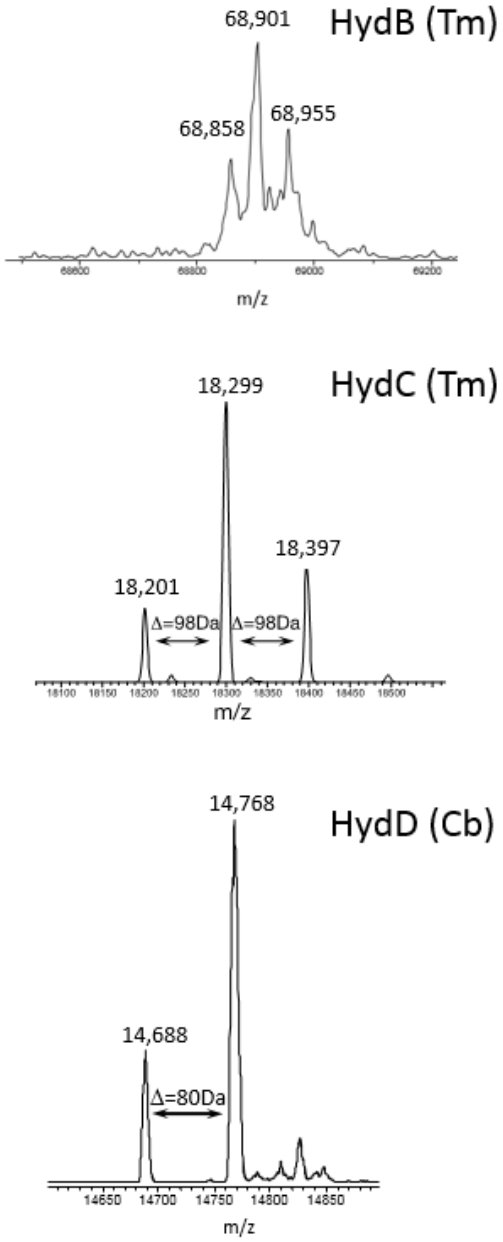
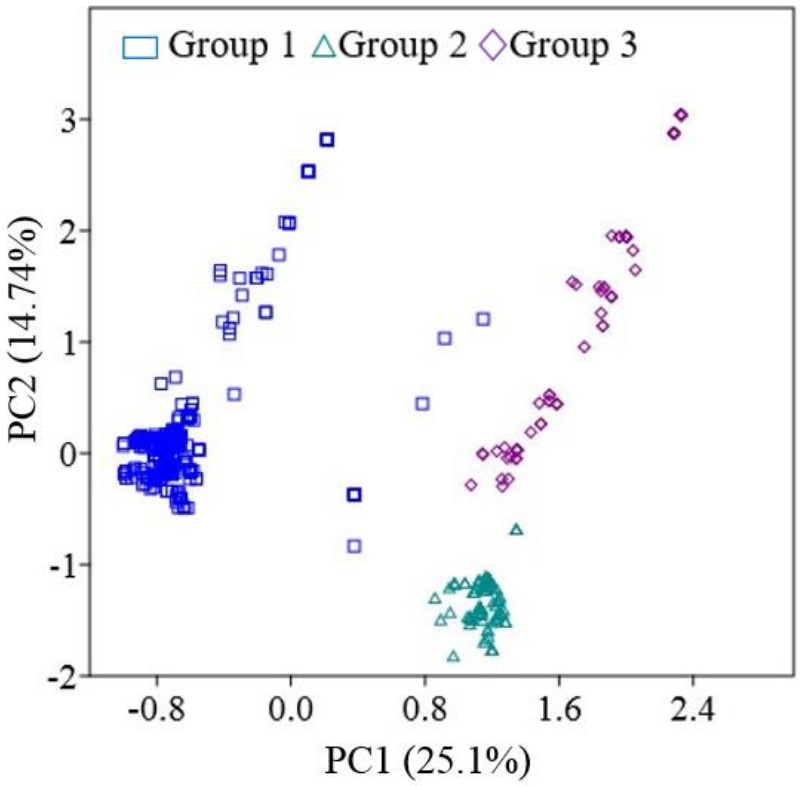
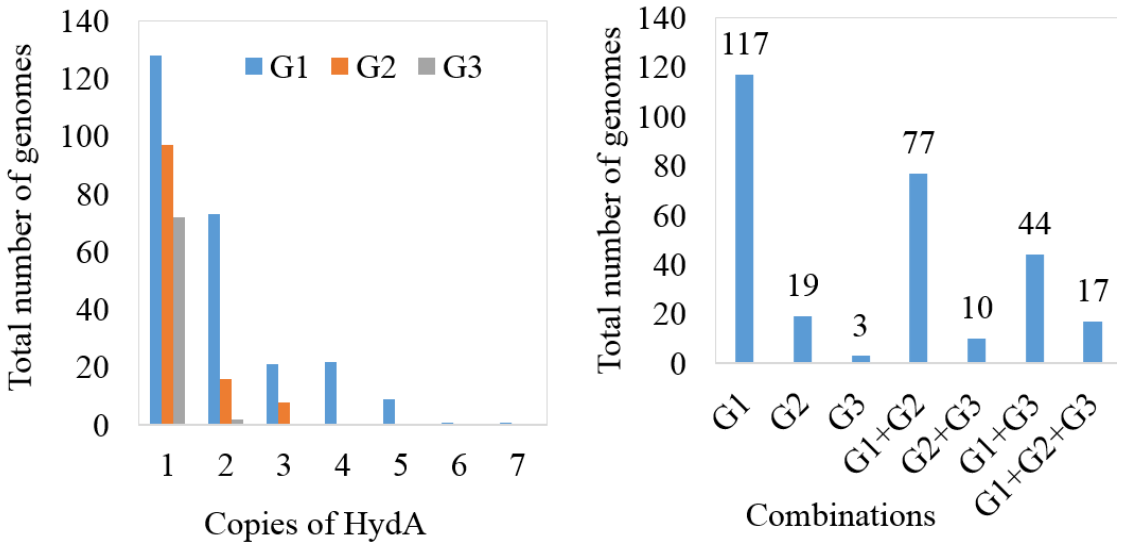


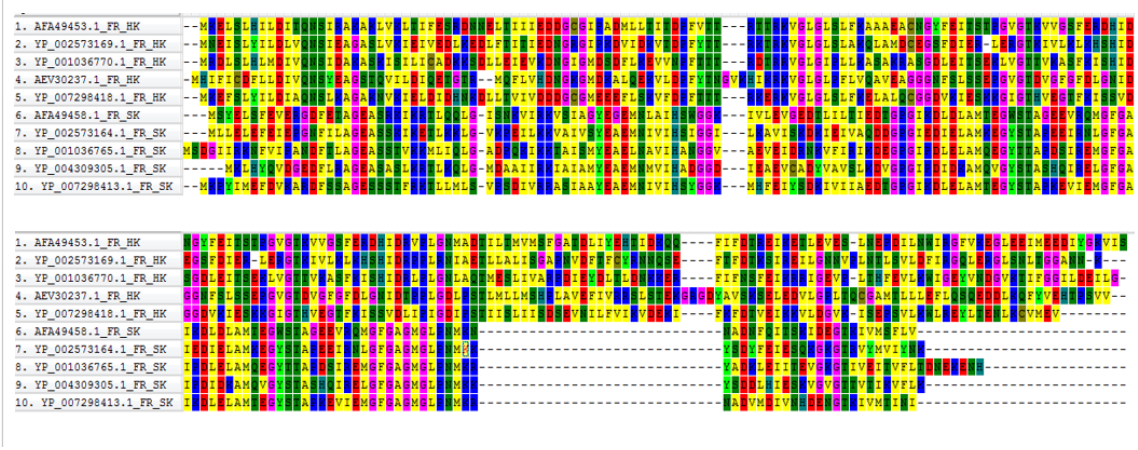
Fig. 9. LCMS analysis of Hyd subunits from *T. maritima* (Tm) and *C. bescii* (Cb). Intact protein analysis of G2 and G3 Hyd separated by SEC-ESI-Q-TOF shows heterogeneity of proteins. The HydB (G2) and HydC (G2) subunits have three prominent forms, expected MW = 68,551 Da (top panel) and expected MW = 18,193 Da (middle panel), respectively. The HydD (G3) subunit (expected MW = 14,688 Da, bottom panel) was found in two major forms spaced by 80 Da.



Supplemental Figure 1. Principal Component Analysis (PCA) ordination of a matrix describing the dissimilarity in the proteins encoded by genes flanking hydA (+/- 10 genes). Only proteins (n = 40) that were encoded in the flanking regions of >2% of the hydA homologs were considered in this analysis. The amount of variation explained by each axis is given in parentheses.

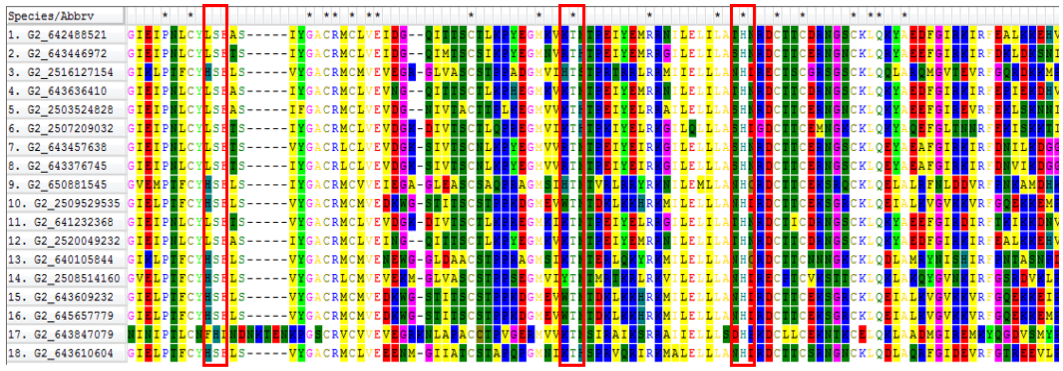


Supplemental Figure 2. Distribution of 714 Hyd groups among the genomes that encode for HydA (n=287). Total number of genomes that encode for one, two, three, four or five isoforms of HydA in each of the three specified groups (A). Different combinations of Hyd groups encoded in Hyd encoding genomes (B).

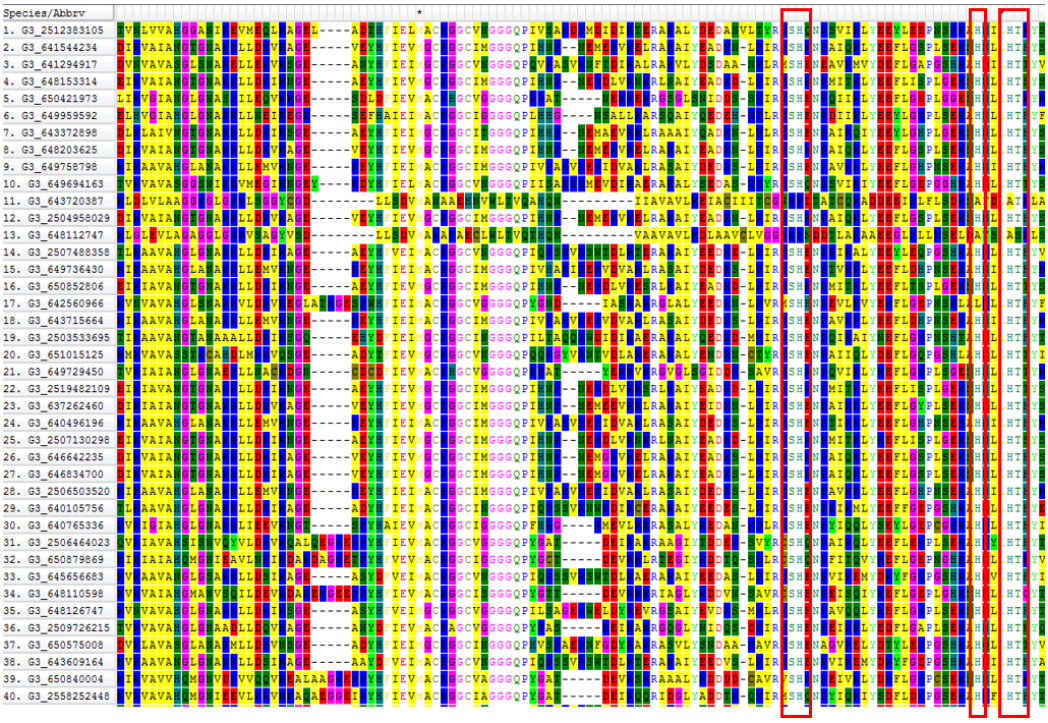


Supplemental Figure 3. Multiple sequence alignment of a segment of histidine kinase-like (Hkl) and serine/threonine kinase (Stk) proteins confirms they are different proteins. The top five sequences represent Htk proteins (represented as HK at the end of each sequence id) while the bottom 5 sequences represent Stk protein (represented as SK at the end of each sequence id) in both top and bottom box.

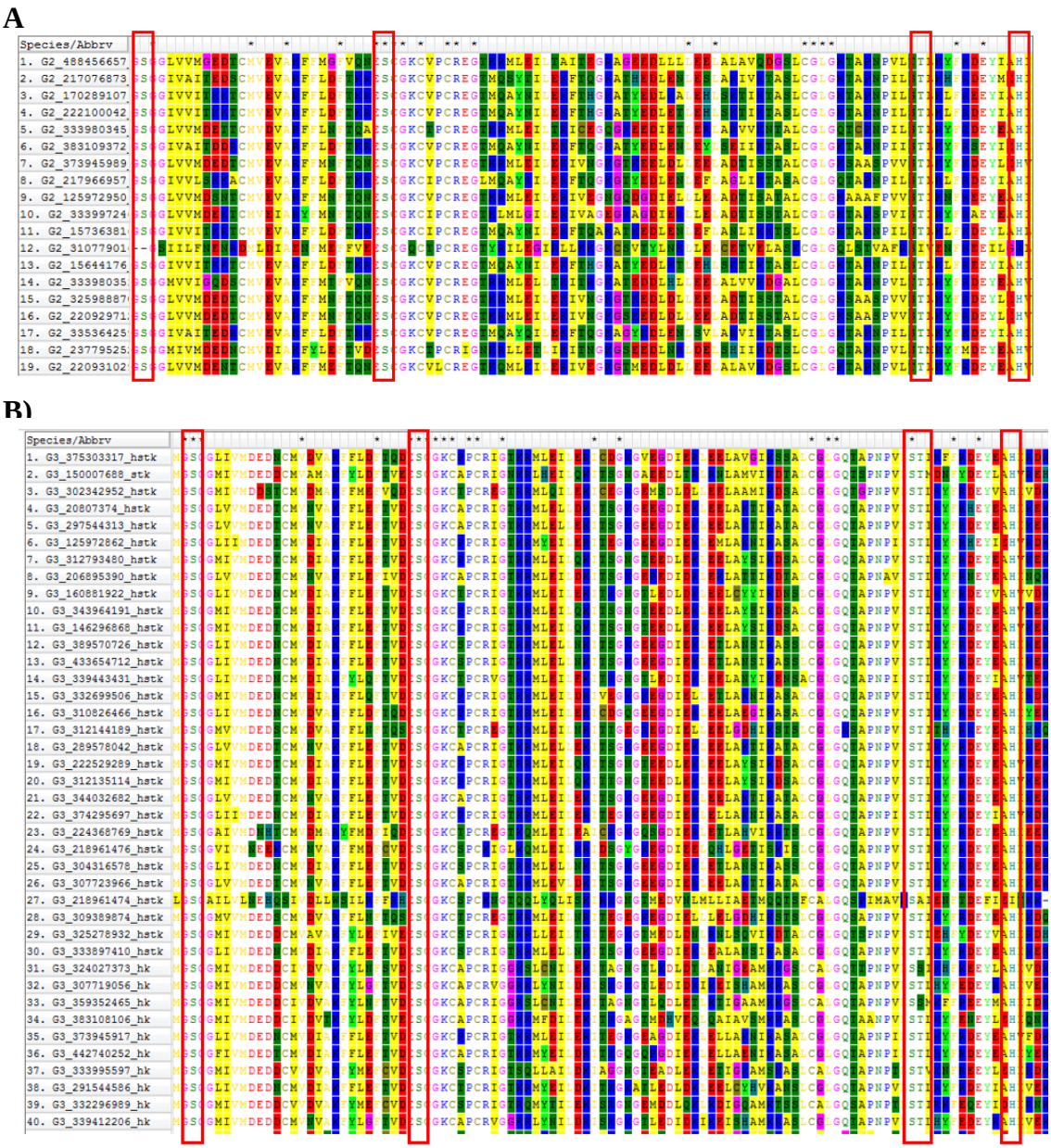
A)



B)

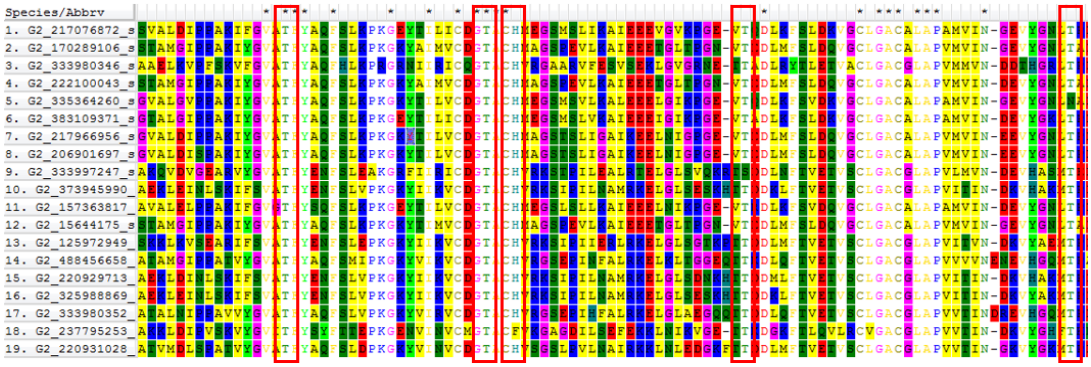


Supplemental Figure 4. Multiple sequence alignment of HydA of two multimeric Hyd groups. **A.** G2 HydA (represented as G2 at the beginning of each sequence id) with serine phosphatases (Sp) in the flanking region. **B.** G3 HydA (represented as G3 at the beginning of each sequence id) with serine/threonine kinase (Stk) or histidine kinase-like (Hkl) in the flanking region. For simplicity, only 18 and 40 aligned sequence for G2 HydA and G3 HydA, respectively, are shown here. Conserved (>90%) serine (S), threonine (T) and histidine (H) residues are highlighted with red boxes.

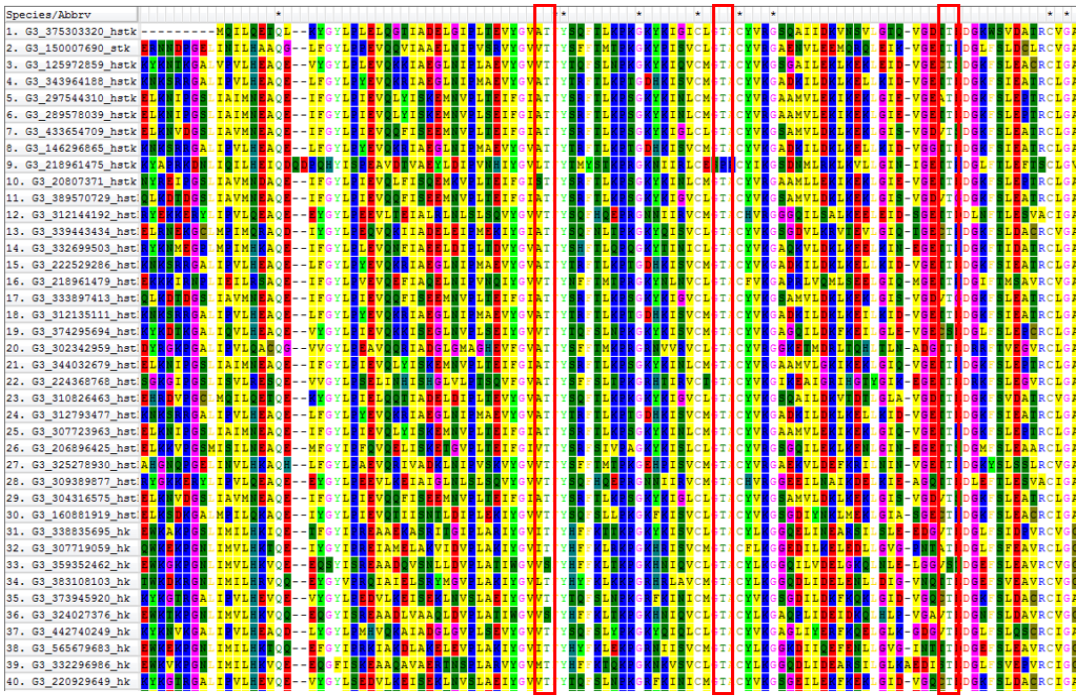


Supplemental Figure 5. Multiple sequence alignment of HydB from both groups that encoded for HydB (i.e. G2 and G3). **A.** HydB encoded by G2 Hyd (represented as G2 at the beginning of each sequence id) that have serine phosphatase (Sp) in the flanking region. **B.** HydB encoded by G3 Hyd (represented as G3 at the beginning of each sequence id) that have serine/threonine kinase (Stk) or histidine kinase-like (Hkl) in the flanking region. For simplicity, only 19 and 40 aligned sequences for G2 HydB and G3 HydB, respectively, are shown here. Conserved (>90%) serine (S), threonine (T) and histidine (H) residues are highlighted with red boxes.

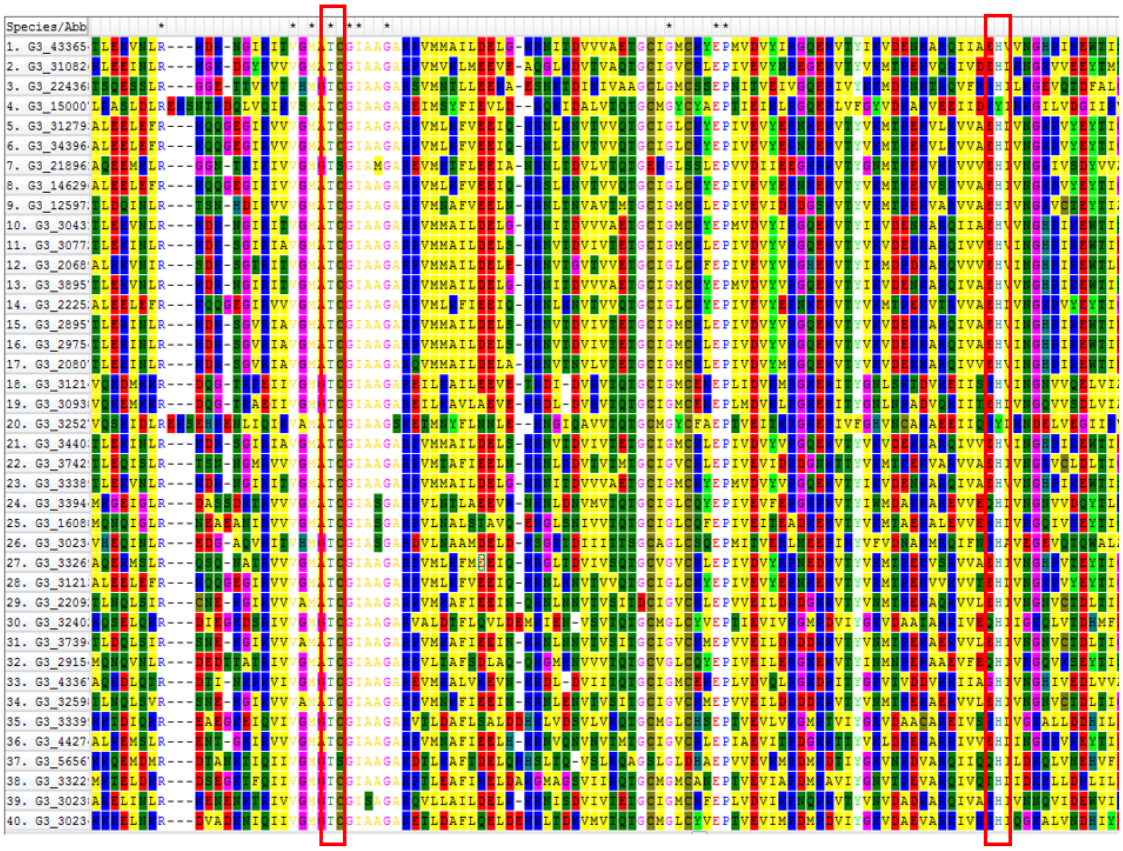
A)



B



Supplemental Figure 6. Multiple sequence alignment of HydC from both groups that encoded for HydC (i.e. G2 and G3). **A)** HydC encoded by G2 Hyd (represented as G2 at the beginning of each sequence id) that have serine phosphatase (Sp) in the flanking region. **B)** HydC encoded by G3 Hyd (represented as G3 at the beginning of each sequence id) that have serine/threonine kinase (Stk) or histidine kinase-like (Hkl) in the flanking region. For simplicity, only 19 and 40 aligned sequences for G2 HydC and G3 HydC, respectively, are shown here. Conserved (>90%) threonine (T) and histidine (H) residues are highlighted with red boxes.



Supplemental Figure 7. Multiple sequence alignment of HydD from the groups that encoded for HydD i.e. G3 (only a subset of the sequence alignment i.e. n=40 are shown here for simplicity). Conserved (>90%) threonine (T) and histidine (H) residues are highlighted with red boxes. : Fast algorithms for mining association rules. *Proceedings of the 20th Very Large Databases (VLDB) Conference*; Santiago, Chile.

REFERENCES

[1] D.W. Mulder, E.S. Boyd, R. Sarma, R.K. Lange, J.A. Endrizzi, J.B. Broderick, J.W. Peters, Stepwise [FeFe]-hydrogenase H-cluster assembly revealed in the structure of HydAΔEFG, Nature, 465 (2010) 248-251.

[2] J.W. Peters, W.N. Lanzilotta, B.J. Lemon, L.C. Seefeldt, X-ray crystal structure of the Fe-only hydrogenase (Cpl) from Clostridium pasteurianum to 1.8 angstrom resolution, Science, 282 (1998) 1853-1858.

[3] J.A. Stapleton, J.R. Swartz, A cell-free microtiter plate screen for improved [FeFe]-hydrogenases, PLOS ONE, (2010).

[4] J.A. Stapleton, J.R. Swartz, Development of an in vitro compartmentalization screen for high-throughput directed evolution of [FeFe]-hydrogenases, PLOS ONE, (2010).

[5] R. Cammack, Michel F, and Robert R., Hydrogen as a Fuel: Learning from Nature, Taylor and Francis Inc., New York, USA, 2001.

[6] W. Lubitz, H. Ogata, O. Rudiger, E. Reijerse, Hydrogenases, Chemical Reviews, 114 (2014) 4081-4148.

[7] T.M. VanderSpek, A.F. Arendsen, R.P. Happe, S.Y. Yun, K.A. Bagley, D.J. Stufkens, W.R. Hagen, S.P.J. Albracht, Similarities in the architecture of the active sites of Ni-hydrogenases and Fe-hydrogenases detected by means of infrared spectroscopy, European Journal of Biochemistry, 237 (1996) 629-634.

[8] S. Shima, R.K. Thauer, U. Ermler, Carbon monoxide as intrinsic ligand to iron in the active site of [Fe]-hydrogenase, Metal ions in life sciences, 6 (2009) 219-240.

[9] S. Shima, O. Pilak, S. Vogt, M. Schick, M.S. Stagni, W. Meyer-Klaucke, E. Warkentin, R.K. Thauer, U. Ermler, The crystal structure of [Fe]-hydrogenase reveals the geometry of the active site, Science, 321 (2008) 572-575.

[10] S. Shima, R.K. Thauer, A third type of hydrogenase catalyzing H2 activation, The Chemical Record, 7 (2007) 37-46.

[11] M.W.W. Adams, The structure and mechanism of iron-hydrogenases, Biochimica Et Biophysica Acta-Bioenergetics, 1020 (1990) 115-145.

[12] R. Cammack, Bioinorganic chemistry - Hydrogenase sophistication, Nature, 397 (1999) 214-215.

[13] K.A. Vincent, A. Parkin, F.A. Armstrong, Investigating and exploiting the electrocatalytic properties of hydrogenases, Chemical Reviews, 107 (2007) 4366-4413.

[14] M. Calusinska, T. Happe, B. Joris, A. Wilmotte, The surprising diversity of clostridial hydrogenases: a comparative genomic perspective, Microbiology, 156 (2010) 1575-1588.

[15] D.S. Horner, B. Heil, T. Happe, T.M. Embley, Iron hydrogenases - ancient enzymes in modern eukaryotes, Trends in Biochemical Sciences, 27 (2002) 148-153.

[16] J. Meyer, [FeFe] hydrogenases and their evolution: a genomic perspective, Cellular and Molecular Life Sciences, 64 (2007) 1063-1084.

[17] J.W. Peters, G.J. Schut, E.S. Boyd, D.W. Mulder, E.M. Shepard, J.B. Broderick, P.W. King, M.W.W. Adams, FeFe - and NiFe -hydrogenase diversity, mechanism, and maturation, Biochimica Et Biophysica Acta-Molecular Cell Research, 1853 (2015) 1350-1369.

[18] S.A. Caffrey, H.S. Park, J.K. Voordouw, Z. He, J. Zhou, G. Voordouw, Function of periplasmic hydrogenases in the sulfate-reducing bacterium Desulfovibrio vulgaris Hildenborough, Journal of Bacteriology, 189 (2007) 6159-6167.

[19] J.M. Odom, H.D. Peck, Hydrogen cycling as a general mechanism for energy coupling in the sulfate-reducing bacteria, Desulfovibrio sp., Fems Microbiology Letters, 12 (1981) 47-50.

[20] J.M. Odom, H.D. Peck, Hydrogenase, electron-transfer proteins, and energy coupling in the sulfate-reducing bacteria Desulfovibrio, Annual Review of Microbiology, 38 (1984) 551-592.

[21] K. Schuchmann, V. Muller, A bacterial electron-bifurcating hydrogenase, Journal of Biological Chemistry, 287 (2012) 31165-31171.

[22] M. Verhagen, T. O'Rourke, M.W.W. Adams, The hyperthermophilic bacterium, Thermotoga maritima, contains an unusually complex iron-hydrogenase: amino acid sequence analyses versus biochemical characterization, Biochimica Et Biophysica Acta-Bioenergetics, 1412 (1999) 212-229.

[23] Y. Nicolet, C. Piras, P. Legrand, C.E. Hatchikian, J.C. Fontecilla-Camps, Desulfovibrio desulfuricans iron hydrogenase: the structure shows unusual coordination to an active site Fe binuclear center, Structure, 7 (1999) 13-23.

[24] G. Voordouw, V. Niviere, F.G. Ferris, P.M. Fedorak, D.W.S. Westlake, Distribution of hydrogenase genes in Desulfovibrio spp. and their use in identification of species from

the oil field environment, Applied and Environmental Microbiology, 56 (1990) 3748-3754.

[25] Y. Nicolet, B.J. Lemon, J.C. Fontecilla-Camps, J.W. Peters, A novel FeS cluster in Fe-only hydrogenases, Trends in Biochemical Sciences, 25 (2000) 138-143.

[26] E.S. Boyd, T.L. Hamilton, K.D. Swanson, A.E. Howells, B.K. Baxter, J.E. Meuser, M.C. Posewitz, J.W. Peters, [FeFe]-hydrogenase abundance and diversity along a vertical redox gradient in Great Salt Lake, USA, International Journal of Molecular Sciences, 15 (2014) 21947-21966.

[27] M. Forestier, P. King, L.P. Zhang, M. Posewitz, S. Schwarzer, T. Happe, M.L. Ghirardi, M. Seibert, Expression of two [Fe]-hydrogenases in Chlamydomonas reinhardtii under anaerobic conditions, European Journal of Biochemistry, 270 (2003) 2750-2758.

[28] M.L. Ghirardi, M.C. Posewitz, P.C. Maness, A. Dubini, J.P. Yu, M. Seibert, Hydrogenases and hydrogen photoproduction in oxygenic photosynthetic organisms, Annual Review of Plant Biology, 58 (2007) 71-91.

[29] T. Happe, A. Kaminski, Differential regulation of the Fe-hydrogenase during anaerobic adaptation in the green alga Chlamydomonas reinhardtii, European Journal of Biochemistry, 269 (2002) 1022-1032.

[30] T. Happe, B. Mosler, J.D. Naber, Induction, localization and metal content of hydrogenase in the green alga Chlamydomonas reinhardtii, European Journal of Biochemistry, 222 (1994) 769-774.

[31] D.W. Mulder, E.M. Shepard, J.E. Meuser, N. Joshi, P.W. King, M.C. Posewitz, J.B. Broderick, J.W. Peters, Insights into [FeFe]-hydrogenase structure, mechanism, and maturation, Structure, 19 (2011) 1038-1052.

[32] J.E. Meuser, E.S. Boyd, G. Ananyev, D. Karns, R. Radakovits, U.M.N. Murthy, M.L. Ghirardi, G.C. Dismukes, J.W. Peters, M.C. Posewitz, Evolutionary significance of an algal gene encoding an [FeFe]-hydrogenase with F-domain homology and hydrogenase activity in Chlorella variabilis NC64A, Planta, 234 (2011) 829-843.

[33] J.S. Chen, L.E. Mortenson, Purification and properties of hydrogenase from Clostridium pasteurianum W5, Biochimica et Biophysica Acta-Protein Structure, 371 (1974) 283-298.

[34] P.M. Vignais, B. Billoud, J. Meyer, Classification and phylogeny of hydrogenases, Fems Microbiology Reviews, 25 (2001) 455-501.

[35] G.J. Schut, M.W.W. Adams, The iron-hydrogenase of Thermotoga maritima utilizes ferredoxin and NADH synergistically: a new perspective on anaerobic hydrogen production, Journal of Bacteriology, 191 (2009) 4451-4457.

[36] S. Wang, H. Huang, J. Kahnt, R.K. Thauer, A reversible electron-bifurcating ferredoxin and NAD-dependent [FeFe]-hydrogenase (HydABC) in Moorella thermoacetica, Journal of Bacteriology, 195 (2013) 1267-1275.

[37] Y. Zheng, J. Kahnt, I.H. Kwon, R.I. Mackie, R.K. Thauer, Hydrogen formation and its regulation in Ruminococcus albus: involvement of an electron-bifurcating [FeFe]-hydrogenase, of a non-electron-bifurcating [FeFe]-hydrogenase, and of a putative hydrogen-sensing [FeFe]-hydrogenase, Journal of Bacteriology, 196 (2014) 3840-3852.

[38] G. Herrmann, E. Jayamani, G. Mai, W. Buckel, Energy conservation via electron-transferring flavoprotein in anaerobic bacteria, Journal of Bacteriology, 190 (2008) 784-791.

[39] F. Li, J. Hinderberger, H. Seedorf, J. Zhang, W. Buckel, R.K. Thauer, Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from Clostridium kluyveri, Journal of Bacteriology, 190 (2008) 843-850.

[40] S. Wang, H. Huang, J. Kahnt, A.P. Mueller, M. Köpke, R.K. Thauer, NADP-specific electron-bifurcating [FeFe]-hydrogenase in a functional complex with formate dehydrogenase in Clostridium autoethanogenum grown on CO, Journal of Bacteriology, 195 (2013) 4373-4386.

[41] S. Wang, H. Huang, J. Kahnt, R.K. Thauer, Clostridium acidurici electron-bifurcating formate dehydrogenase, Applied and Environmental Microbiology, 79 (2013) 6176-6179.

[42] W. Buckel, R.K. Thauer, Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation, Biochimica Et Biophysica Acta-Bioenergetics, 1827 (2013) 94-113.

[43] H.Y. Huang, S.N. Wang, J. Moll, R.K. Thauer, Electron bifurcation involved in the energy metabolism of the acetogenic bacterium Moorella thermoacetica growing on glucose or H2 plus CO2, Journal of Bacteriology, 194 (2012) 3689-3699.

[44] S. Malki, I. Saimmaime, G. Deluca, M. Rousset, Z. Dermoun, J.P. Belaich, Characterization of an operon encoding an NADP-reducing hydrogenase in Desulfovibrio fructosovorans, Journal of Bacteriology, 177 (1995) 2628-2636.

[45] J.T. Henry, S. Crosson, Ligand-binding PAS domains in a genomic, cellular, and structural context, Annual Review of Microbiology, 65 (2011) 261-286.

[46] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, D.G. Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology, 7 (2011) 539-544.

[47] G.E. Crooks, G. Hon, J.M. Chandonia, S.E. Brenner, WebLogo: A Sequence Logo Generator, Genome Research, 14 (2004) 1188-1190.

[48] I. Dondoshansky, Y. Wolf, Blastclust (ncbi software development toolkit), NCBI, Bethesda, Md, (2002).

[49] A. Marchler-Bauer, M.K. Derbyshire, N.R. Gonzales, S. Lu, F. Chitsaz, L.Y. Geer, R.C. Geer, J. He, M. Gwadz, D.I. Hurwitz, C.J. Lanczycki, F. Lu, G.H. Marchler, J.S. Song, N. Thanki, Z. Wang, R.A. Yamashita, D. Zhang, C. Zheng, S.H. Bryant, CDD: NCBI's conserved domain database, Nucleic Acids Research, 43 (2015) D222-D226.

[50] The Universal Protein Resource (UniProt), Nucleic Acids Research, 36 (2008) D190-D195.

[51] M. Hahsler, S. Chelluboina, K. Hornik, C. Buchta, The arules R-package ecosystem: Analyzing interesting patterns from large transaction data sets, Journal of Machine Learning Research, 12 (2011) 2021-2025.

[52] R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th Very Large Databases (VLDB) Conference, vol. 1215, Santiago, Chile, 1994, pp. 487-499.

[53] O. Hammer, D.A.T. Harper, P.D. Ryan, PAST: Paleontological statistics software package for education and data analysis, Palaeontologia Electronica, 4 (2001) 1-9.

[54] M.E. Smoot, K. Ono, J. Ruscheinski, P.L. Wang, T. Ideker, Cytoscape 2.8: New features for data integration and network visualization, Bioinformatics, 27 (2011) 431-432.

[55] Y. Assenov, F. Ramirez, S.E. Schelhorn, T. Lengauer, M. Albrecht, Computing topological parameters of biological networks, Bioinformatics, 24 (2008) 282-284.

[56] K. Ma, M.W.W. Adams, Sulfide dehydrogenase from the hyperthermophilic archaeon Pyrococcus furiosus: a new multifunctional enzyme involved in the reduction of elemental sulfur, Journal of Bacteriology, 176 (1994) 6509-6517.

[57] D.A. Ravcheev, X.Q. Li, H. Latif, K. Zengler, S.A. Leyn, Y.D. Korostelev, A.E. Kazakov, P.S. Novichkov, A.L. Osterman, D.A. Rodionov, Transcriptional regulation of central carbon and energy metabolism in bacteria by redox-responsive repressor Rex, Journal of Bacteriology, 194 (2012) 1145-1157.

[58] E. Hidalgo, B. Demple, An iron-sulfur center essential for transcriptional activation by the redox-sensing SoxR protein, The EMBO Journal, 13 (1994) 138-146.

[59] M. Zheng, G. Storz, Redox sensing by prokaryotic transcription factors, Biochemical Pharmacology, 59 (2000) 1-6.

[60] S.N. Wang, H.Y. Huang, J. Moll, R.K. Thauer, NADP+ reduction with reduced ferredoxin and NADP+ reduction with NADH are coupled via an electron-bifurcating enzyme complex in Clostridium kluyveri, Journal of Bacteriology, 192 (2010) 5115-5123.

[61] P.M. Vignais, B. Billoud, Occurrence, classification, and biological function of hydrogenases: An overview, Chemical Reviews, 107 (2007) 4206-4272.

[62] L. Feinberg, J. Foden, T. Barrett, K.W. Davenport, D. Bruce, C. Detter, R. Tapia, C. Han, A. Lapidus, S. Lucas, J.F. Cheng, S. Pitluck, T. Woyke, N. Ivanova, N. Mikhailova, M. Land, L. Hauser, D.A. Argyros, L. Goodwin, D. Hogsett, N. Caiazza, Complete genome sequence of the cellulolytic thermophile Clostridium thermocellum DSM1313, Journal of Bacteriology, 193 (2011) 2906-2907.

[63] X. Brazzolotto, J.K. Rubach, J. Gaillard, S. Gambarelli, M. Atta, M. Fontecave, The [Fe-Fe]-hydrogenase maturation protein HydF from Thermotoga maritima is a GTPase with an iron-sulfur cluster, Journal of Biological Chemistry, 281 (2006) 769-774.

[64] E.M. Shepard, E.S. Boyd, J.B. Broderick, J.W. Peters, Biosynthesis of complex iron-sulfur enzymes, Current Opinion in Chemical Biology, 15 (2011) 319-327.

[65] E.M. Shepard, B.R. Duffus, S.J. George, S.E. McGlynn, M.R. Challand, K.D. Swanson, P.L. Roach, S.P. Cramer, J.W. Peters, J.B. Broderick, [FeFe]-hydrogenase maturation: HydG-catalyzed synthesis of carbon monoxide, Journal of the American Chemical Society, 132 (2010) 9247-9249.

[66] E.M. Shepard, S.E. McGlynn, A.L. Bueling, C.S. Grady-Smith, S.J. George, M.A. Winslow, S.P. Cramer, J.W. Peters, J.B. Broderick, Synthesis of the 2Fe subcluster of the [FeFe]-hydrogenase H-cluster on the HydF scaffold, Proceedings of the National Academy of Sciences of the United States of America, 107 (2010) 10448-10453.

[67] J.K. Rubach, X. Brazzolotto, J. Gaillard, M. Fontecave, Biochemical characterization of the HydE and HydG iron-only hydrogenase maturation enzymes from Thermatoga maritima, Febs Letters, 579 (2005) 5055-5060.

[68] S. Bellon, J.D. Parsons, Y.Y. Wei, K. Hayakawa, L.L. Swenson, P.S. Charifson, J.A. Lippke, R. Aldape, C.H. Gross, Crystal structures of Escherichia coli topoisomerase IV ParE subunit (24 and 43 Kilodaltons): a single residue dictates differences in novobiocin

potency against topoisomerase IV and DNA gyrase, Antimicrobial Agents and Chemotherapy, 48 (2004) 1856-1864.

[69] T. Tanaka, S.K. Saha, C. Tomomori, R. Ishima, D.J. Liu, K.I. Tong, H. Park, R. Dutta, L. Qin, M.B. Swindells, T. Yamazaki, A.M. Ono, M. Kainosho, M. Inouye, M. Ikura, NMR structure of the histidine kinase domain of the E. coli osmosensor EnvZ, Nature, 396 (1998) 88-92.

[70] L. Aravind, C.P. Ponting, The cytoplasmic helical linker domain of receptor histidine kinase and methyl-accepting proteins is common to many prokaryotic signalling proteins, Fems Microbiology Letters, 176 (1999) 111-116.

[71] A. Daddaoua, C. Molina-Santiago, J. de la Torre, T. Krell, J.L. Ramos, GtrS and GltR form a two-component system: The central role of 2-ketogluconate in the expression of exotoxin A and glucose catabolic enzymes in Pseudomonas aeruginosa, Nucleic Acids Research, 42 (2014) 7654-7665.

[72] A. Marina, C.D. Waldburger, W.A. Hendrickson, Structure of the entire cytoplasmic portion of a sensor histidine kinase protein, The EMBO Journal, 24 (2005) 4247-4259.

[73] P.M. Wolanin, P.A. Thomason, J.B. Stock, Histidine protein kinases: key signal transducers outside the animal kingdom, Genome Biology, 3 (2002).

[74] A.M. Stock, V.L. Robinson, P.N. Goudreau, Two-component signal transduction, Annual Review of Biochemistry, 69 (2000) 183-215.

[75] K.E. Swain, J.J. Falke, Structure of the conserved HAMP domain in an intact, membrane-bound chemoreceptor: a disulfide mapping study, Biochemistry, 46 (2007) 13684-13695.

[76] S.B. Williams, V. Stewart, Functional similarities among two-component sensors and methyl-accepting chemotaxis proteins suggest a role for linker region amphipathic helices in transmembrane signal transduction, Molecular Microbiology, 33 (1999) 1093-1102.

[77] B. Boxma, R.M. de Graaf, G.W.M. van der Staay, T.A. van Alen, G. Ricard, T. Gabaldon, A.H.A.M. van Hoek, S.Y. Moon-van der Staay, W.J.H. Koopman, J.J. van Hellemond, A.G.M. Tielens, T. Friedrich, M. Veenhuis, M.A. Huynen, J.H.P. Hackstein, An anaerobic mitochondrion that produces hydrogen, Nature, 434 (2005) 74-79.

[78] B. Boxma, G. Ricard, A.H. van Hoek, E. Severing, S.Y. Moon-van der Staay, G.W.M. van der Staay, T.A. van Alen, R.M. de Graaf, G. Cremers, M. Kwantes, N.R. McEwan, C.J. Newbold, J.P. Jouany, T. Michalowski, P. Pristas, M.A. Huynen, J.H.P. Hackstein, The [FeFe] hydrogenase of Nyctotherus ovalis has a chimeric origin, BMC Evolutionary Biology, 7 (2007) 230-242.

[79] R.M. de Graaf, G. Ricard, T.A. van Alen, I. Duarte, B.E. Dutilh, C. Burgtorf, J.W.P. Kuiper, G.W.M. van der Staay, A.G.M. Tielens, M.A. Huynen, J.H.P. Hackstein, The organellar genome and metabolic potential of the hydrogen-producing mitochondrion of Nyctotherus ovalis, Molecular Biology and Evolution, 28 (2011) 2379-2391.

[80] J. Cieśla, T. Frączyk, W. Rode, Phosphorylation of basic amino acid residues in proteins: important but easily missed, ACTA Biochimica Polonica, 58 (2011) 137-147.

[81] J. Deutscher, M.H. Saier, Ser/Thr/Tyr protein phosphorylation in bacteria - for long time neglected, now well established, Journal of Molecular Microbiology and Biotechnology, 9 (2005) 125-131.

[82] S.F.F. Pereira, L. Goss, J. Dworkin, Eukaryote-like serine/threonine kinases and phosphatases in Bacteria, Microbiology and Molecular Biology Reviews, 75 (2011) 192-212.

[83] J.B. Stock, A.J. Ninfa, A.M. Stock, Protein phosphorylation and regulation of adaptive responses in Bacteria, Microbiological Reviews, 53 (1989) 450-490.

[84] M. Schnolzer, W.D. Lehmann, Identification of modified peptides by metastable fragmentation in MALDI mass spectrometry, International Journal of Mass Spectrometry, 169 (1997) 263-271.

[85] A. Tholey, J. Reed, W.D. Lehmann, Electrospray tandem mass spectrometric studies of phosphopeptides and phosphopeptide analogues, Journal of Mass Spectrometry, 34 (1999) 117-123.

[86] D.T. McLachlin, B.T. Chait, Analysis of phosphorylated proteins and peptides by mass spectrometry, Current Opinion in Chemical Biology, 5 (2001) 591-602.

[87] K.F. Medzihradszky, N.J. Phillipps, L. Senderowicz, P. Wang, C.W. Turck, Synthesis and characterization of histidine-phosphorylated peptides, Protein Science, 6 (1997) 1405-1411.

[88] R.S. Annan, S.A. Carr, Phosphopeptide analysis by matrix-assisted laser desorption time-of-flight mass spectrometry, Analytical Chemistry, 68 (1996) 3413-3421.

[89] S. Metzger, R. Hoffmann, Studies on the dephosphorylation of phosphotyrosine-containing peptides during post-source decay in matrix-assisted laser desorption/ionization, Journal of Mass Spectrometry, 35 (2000) 1165-1177.

APPENDIX B

COPY OF PERMISSION TO REPRINT