Comparative Genomic Analysis Reveals 2-Oxoadic Dehydrogenase Complex Lipoylation Correlation with Aerobiosis in Archaea

Kirill Borziak¹, Mareike G. Posner², Abhishek Upadhyay², Michael J. Danson²,³, Stefan Bagby², Steve Dorus¹*

¹Department of Biology, Syracuse University, Syracuse, New York, United States of America, ²Department of Biology & Biochemistry, University of Bath, Claverton Down, United Kingdom, ³Centre for Extremophile Research, University of Bath, Claverton Down, United Kingdom

Abstract

Metagenomic analyses have advanced our understanding of ecological microbial diversity, but to what extent can metagenomic data be used to predict the metabolic capacity of difficult-to-study organisms and their abiotic environmental interactions? We tackle this question, using a comparative genomic approach, by considering the molecular basis of aerobiosis within archaea. Lipoylation, the covalent attachment of lipoic acid to 2-oxoacid dehydrogenase multienzyme complexes (OADHCs), is essential for metabolism in aerobic bacteria and eukarya. Lipoylation is catalysed either by lipoate protein ligase (LplA), which in archaea is typically encoded by two genes (Lpla-N and Lpla-C), or by a lipoyleoctanoyl) transferase (Lpb or LipM) plus a lipoic acid synthetase (LipA). Does the genomic presence of lipoylation and OADHC genes across archaea from diverse habitats correlate with aerobiosis? First, analyses of 11,826 biotin protein ligase (BPL)-LplA-LipB transferase family members and 147 archaeal genomes identified 85 species with lipoylation capabilities and provided support for multiple ancestral acquisitions of lipoylation pathways during archaeal evolution. Second, with the exception of the Sulfolobales order, the majority of species possessing lipoylation systems exclusively retain LplA, or either LipB or LipM, consistent with archaeal genome streamlining. Third, obligate anaerobic archaea display widespread loss of lipoylation and OADHC genes. Conversely, a high level of correspondence is observed between aerobiosis and the presence of LplA/LipB/LipM, LipA and OADHC E2, consistent with the role of lipoylation in aerobic metabolism. This correspondence between OADHC lipoylation capacity and aerobiosis indicates that genomic pathway profiling in archaea is informative and that well characterized pathways may be predictive in relation to abiotic conditions in difficult-to-study extremophiles. Given the highly variable retention of gene repertoires across the archaea, the extension of comparative genomic pathway profiling to broader metabolic and homeostasis networks should be useful in revealing characteristics from metagenomic datasets related to adaptations to diverse environments.

Introduction

Culture-independent, metagenomic analyses have been particularly successful in advancing our knowledge of microbial abundance across diverse ecological niches (reviewed by [1]). Nonetheless, few studies have leveraged the wealth of genomic data across diverse archaeal taxa to explore adaptation to extreme archaean environments although this must have a functional basis in genomic diversification [2,3,4,5]. Recent experimental studies have begun to utilize metagenomic data to decipher evolutionary processes [6] but substantial obstacles remain in applying such approaches to the complex biotic and abiotic interactions of natural populations (reviewed by [7]). To what extent can comparative genomic approaches inform our understanding of the evolution and functional capacity of organisms that cannot be cultured or studied in the laboratory? Further, can abiotic characteristics of extremophile habitats be inferred directly from the analysis of metagenomic data?

Archaeal evolution has been dominated by reductions in genome complexity and the retention of highly variable genetic architectures across lineages ([8] and reviewed by [9]). Recent analyses reveal two distinct phases of archaean genome evolution. The first, the innovation phase, is associated with an increase in genome complexity and an associated increase in gene families to an average of approximately 2500 gene families. The second, the reductive phase, is characterized by genome streamlining and the retention of a more minimal, and potentially heterogeneous, gene repertoire (1400–1800 gene families) [10]. This persistent genomic streamlining has radically altered the repertoires of even the most highly conserved gene classes, including those involved in translation, replication, cell division and DNA repair, and is thus central to functional diversity across the domain [11]. In addition to the diversifying impact of differential gene loss across taxa, archaean genome analyses have revealed notable exceptions where horizontal gene transfer (HGT) has been a prevalent force. For example, gene flow from eubacteria to Halobacteriales has
contributed to the absence of reductive genome evolution in this archaean order [10]. We therefore propose that gene repertoire heterogeneity, particularly associated with metabolism and homeostasis, may reflect archaean adaptation to, and exploitation of, a remarkable diversity of environments. We assess this possibility by considering aerobicis within archaea because (i) archaean display tremendous diversity in their utilization and tolerance of aerobic environments and (ii) aerobicis pathways have been well characterized biochemically. Lipoylation, the covalent attachment of lipoic acid to the dihydrolipoyl acyltransferase (E2) subunit of 2-oxoacid dehydrogenase multienzyme complexes (OADHCs), is essential for metabolism in aerobic bacteria and eukarya (reviewed by [12,13]). Specifically, OADHC lipoylation is required for channeling substrates between the active sites of the three protein subunits of OADHCs: 2-oxoacid decarboxylase (E1), E2 and dihydrolipoiapamide dehydrogenase (E3). The lipoyl domain of E2 (E2lipD) is the post-translational modification target. The mechanisms of lipoylation have been studied to varying extents in all domains of life [14,15,16,17]. In *Escherichia coli*, lipoylation is catalyzed by two routes: lipoic acid synthetase (LipA) and lipoic/octanoyl transferase (LipB), or lipoate protein ligase (LplA) [18]. LipB and LipA work in tandem: LipB catalyzes the covalent attachment of octanoic acid to the E2 lipoyl domain, and then LipA introduces sulpher atoms at the C6 and C8 positions. Alternatively, LipA can catalyze both conversion of lipoic acid to lipoyl-AMP and subsequent covalent attachment of the lipoyl moiety to E2lipD [19,20]. It is noteworthy that greater diversity in lipoyl biosynthesis has been observed in other eubacteria, including an alternative octanoyl transferase, LipM, and a lipoyl-scavenging protein, LipL, in Firmicutes [17,21,22]. In eukaryotes and most bacteria, LipA is encoded by a single gene, whereas studies in the archaean *Thermoplasma acidophilum* revealed distinct genes, LplA-N and LplA-C, encoding proteins that correspond to the N- and C-terminal domains of *E. coli* LplA and that are both required for E2 lipoylation [23,24,25]. The distribution and genomic characteristics of lipoylation systems have yet to be studied across archaea.

Based on their well characterized biochemical interaction, we propose that genomic retention of the components of the OADHC lipoylation pathway, including lipoylation enzymes and E2, may serve as a diagnostic marker for aerobic metabolism. We have therefore examined their evolutionary retention across available archaean genomes in the context of the following predictions. First, co-retention of LplA, LipB or LipM is unexpected given the widespread genomic streamlining observed in archaea. Second, the octanoyl transferases, LipB and LipM, would appear to be unlikely to be the pervasive archaean lipoylation system given their enzymatic preference for octanoic acid, a product of fatty acid (FA) biosynthesis. FA biosynthesis was believed to be completely absent from archaea [26], although archaean FA synthase pathways have recently been identified [27]. Although the prevalence of archaean FA biosynthesis has yet to be carefully examined, we suggest that the genomic presence of octanoyl transferases may be a reliable indicator of this biochemical capacity. Third, evolutionary loss of lipoylation, including lipoylation enzymes and their E2 substrates, may be widespread in anaerobic archaean, particularly those that are obligate anaerobes or display poor oxygen tolerance. Targeting this well-characterized metabolic pathway also provides a general assessment of the robustness of genomic inferences about the metabolic regimes of difficult-to-study microbes whose genomes are highly represented in environmental metagenomic studies [28,29,30].

**Materials and Methods**

**Lipoylation System Classification**

Lipoylation systems across the three domains of life were surveyed to assess the presence of each lipoylation system amongst archaea. To do so we characterized the genomic composition of lipoylation systems and OADHC lipoic acid acceptor protein (E2) in 147 archaean species, including 43 Crenarchaeota, 96 Euryarchaeota, 5 Thaumarchaeota, 1 Korarchaeum, 1 Nanoarchaeum and 1 Aigarchaeum of which 20 are genome sequences from metagenomic environmental samples. First, an analysis of all 11,826 protein domains within the Pfam BPL, LipA-LipB cofactor transferase family protein domain [PF03099] [31] was conducted. Domain protein sequences were aligned using the MAFFT iterative refinement method [32], and a neighbor-joining phylogenetic tree was constructed with the NJ,PAUP, using the default parameters [33]. The resultant phylogeny resolved clades that corresponded to LipA, LipB, LipL, and LipB on the existing biochemical characterization for proteins within each clad [16,17,19,34,35,36] This Pfam analysis thus provided a preliminary catalogue of archaean lipoylation.

**Comparative Genomic Analysis**

To address the possible incomplete annotation of archaean lipoylation proteins in the Pfam PF03099 database, homology-based approaches were used to confirm and expand the identification of LipA, LipM and LipB in the 147 archaean genomes (Table S1). *T. acidophilum* LplA-N (Q9HKT1), *F. acidarmanus* LipB (P0AQU0) and *M. avorzae* LipM (Q0W151) protein sequences were obtained from UniProt and searched against annotated archaeal protein databases (NCBI Microbial Genomes) using BLASTp (E-threshold = 1E-10) to identify a representative sequence with the highest homology in each of the thirteen taxonomical groups analyzed. These “best-hit” representative sequences were then searched against available genome sequences using BLASTn within their respective taxonomical group to determine the presence and copy number of each gene. In species where no homologous genes were identified, PSI-BLAST (E-threshold = 0.001; 2 iterations maximum) was also used to confirm the absence of any related sequence. Both BLASTp and tBLASTn results were manually assessed to ensure identification of lipoylation proteins and exclusion of biotinylation proteins, based on the Pfam phylogenetic classification.

A similar approach was used to assess the presence of the lipoic acid adenyltransferase lipoylation domain LipA-C (using *T. acidophilum* Q9HKT2), the octanoyl synthase LipA (using *F. pruexulicas* S0AQU0), the eubacterial octanoyl transferase LipL (using *B. subtilis* P35411), and lipoylation substrates, including the dihydrolipoyl transferase (E2) subunit of the OADHC (using *T. acidophilum* Q9HA5). The *B. subtilis* LipL sequence was used because no annotated archaean LipL exists. BLASTp and tBLASTn were conducted on these sequences as described above. Again, manual curation was employed to exclude proteins with the LipA-C domain that exist as part of lipoylation and biotinylation proteins, non-LipA radical SAM proteins, and biotinylation targets. In order to identify lipoylation targets exhaustively, the lipoyl domain of *T. acidophilum* E2 was used as a PSI-BLAST query. Using an E-value cutoff of 0.001, PSI-BLAST was iterated until convergence (four iterations). Due to the abundance of biotinyl domains in the results, maximum likelihood phylogenetic analyses were employed to differentiate between the two targets (see below). The lipoyl domains were also differentiated from biotinyl domains based on protein domain architecture and sequence annotation. The resultant lipoyl domain-containing proteins were classified based
on their domain architectures, revealing three distinct classes: true dihydrolipoamide transferase proteins (based on the presence of the acyltransferase catalytic domain, PF00190), glycine cleavage protein H (GcvH) (based on annotation and high homology with biochemically characterized bacterial GcvH), and single domain proteins (containing only the lipoyl domain). Confirming our Pfam results, no archaeal LipL proteins were identified in either BLASTp or tBLASTn analyses.

Additionally, sequence motif analysis was conducted as a validation step. To confirm LplA-N identification we examined conservation amongst two essential motifs to confirm LplA-N sequences (Fig. S1). The first is RRXTGGG(A/S/T)(A/I/V)(I/F/Y)HD with the second R and first two Gs forming the core. In the T. acidophilum LplA-N:LplA-C complex structure [25], this motif lies at the functional interface of LplA-N with LplA-C and includes the lipoate binding loop. The second conserved motif is G(R/K)K(I/L/V)SGX(A/G)Q, with occasional substitution of the first G, the S and Q. This motif, corresponding to residues 143–151 of T. acidophilum LplA-N, forms part of B9 in T. acidophilum LplA-N:LplA-C. B9 is located adjacent to the lipoate binding loop [25] and in lipoyl-AMP-bound structures the conserved K and G at the third and sixth positions respectively of the motif are involved in interactions with the adenine and lipoyl parts of lipoyl-AMP [37].

Maximum Likelihood Phylogenetic Analysis

The phylogenetic relationships between LplA, LipM and LipB in archaea, eubacteria and eukaryotes, as well as all archaeal proteins containing biotinyl and lipoyl domains were analyzed using a maximum likelihood phylogenetic approach. Protein sequences were retrieved from UniProt for all archaea identified in the previous analyses, major eukaryotic species (S. cerevisiae, D. melanogaster, M. musculus and H. sapiens) and eubacteria representing Actinobacteria (S. coelicolor), Bacteroidetes (B. thetaiotaomicron), Firmicutes (B. subtilis and S. aureus) and Proteobacteria (E. coli and B. pseudomallei). Multiple sequence alignment was conducted using the L-INS-i algorithm of MAFFT [32]. Bootstrapped maximum likelihood phylogenetic analyses were done with empirical amino acid frequencies, sub-tree pruning and regrafting topology search, and a parsimony starting tree using the PhyML package [38].

Horizontal Gene Transfer (HGT) Analysis

The codon based approach of Davis and Olsen (2009) was used to calculate modal codon usage for the 147 archaeal species surveyed above and to detect significant codon usage outliers as putative HGT events [39]. Protein coding sequences were downloaded for all species and genes were deemed as recent horizontal acquisition events if the codon usage was significantly different from the whole genome modal frequency using a threshold of p<0.01, as suggested by Davis and Olsen (2009), and a more conservative threshold (p<0.05). Statistical comparison of the frequency of HGT between gene sets was conducted using a two-tailed Chi-square test with Yates correction.

Archaean Metabolic Environments

The categorization of archaean metabolic environments, particularly relating to aerobicism and oxygen tolerance, was based on a detailed curation of available literature and the Genomes Online Database (GOLD v4.0; [40]). The availability of phenotypic and habitat information is highly variable amongst archaea, particularly given the expanded use of metagenomic environmental sampling. In some archaean orders, relevant data were limited to a subset of the member species.

Results

Archaean Lipoylation Pathway Heterogeneity

To broadly characterize the distribution of lipoylation pathways across archaean we conducted a comprehensive analysis of proteins within the biotin and lipoate B/A ligase and octanoyl carrier domain family (Pfam03099). Neighbor-joining tree construction for all domain sequences within this family (n = 11,826) resulted in five broad clades, encompassing the biotin ligase, LplA, LipB, LipL and LipM protein groups (Fig. 1). Archaean representatives were identified amongst all these clades with the exception of LipL, which has been previously identified only in bacterial Firmicutes [17]. In total, 126 archaeal lipoylation proteins were identified and formed the basis of our subsequent assessment of the prevalence of lipoylation proteins across 147 archaean species. These searches revealed 16 additional archaean proteins in 14 taxa, resulting in a total of 142 lipoylation proteins in 85 species (Table S1).
The LipB lipoylation system was found to be the least prevalent (11 genes within 9 species) and was straightforward to distinguish given the substantial sequence divergence between LipB and LipM/LplA proteins (Fig. 1). It is noteworthy that all species possessing LipB also possess the lipoyl synthase LipA and E2 and thus have a complete E2 lipoylation pathway. Due to the higher levels of homology, a full maximum likelihood phylogenetic analysis was conducted to distinguish between LipM and LplA sequences. Of the 131 protein sequences analyzed, 63 genes in 44 taxa were identified within two closely related monophyletic clades that were associated with Actinobacterial and Firmicute LipM, respectively (Fig. 2; Clade III and IV). The remaining 68 sequences in 50 taxa consist of putative LipA-N proteins. Amongst these, the proteins found in Clade I are LplA based upon (i) their phylogenetic relationship with the biochemically characterized *T. acidophilum* LplA, (ii) manual inspection of diagnostic amino acid residues within the catalytic domain (see Methods) and (iii) the genomic presence of the other part of this bipartite system, LplA-C. Despite the closer phylogenetic relationship of Clade II to LipM, we propose that proteins in this clade are LipA-N, and not octanoyl transferases, based on the correlated presence of LplA-C and the absence of LipA in the genomes of these taxa (Table S1); however this classification should be considered provisional in the absence of additional biochemical data. Based on the fact that the octanoic acid transferred to E2 by LipM requires LipA for conversion to a lipoyl group, the genomic presence of LipA was used to independently confirm LipM in Clades III and IV. As expected, LipA was identified in 38 of the 44 genomes that possess LipM, thus confirming the complete lipoylation pathway in those taxa. Additionally, LipA was absent from all species with LipA as a sole transferase system, with the exception of *A. pernix*. Thus, the combination of complementary phylogenetic and genomic approaches provides a substantive basis for differentiating between LipM and LplA.

**Origins of Archaeal Lipoylation Pathways**

Our phylogenetic analysis provides support for ancestral monophyletic origins of lipoylation systems prior to the divergence of species within archaean orders. A pattern consistent with this is observed in several clades, including LipA in Thermococcales, LipM in Methanococcales, Halobacteriales and Sulfurospirales (Fig. 2), and LipB in Thermoproteales and the closely related *Aigarchaeota* (represented by *C. subterraneum*) (Fig. S3). In several cases it is also possible to infer the likely source of lipoylation system acquisition. For example, the inclusion of *Staphylococcus* within LipM Clade III, which includes diverse Halobacteriales and Methanococcales, is consistent with acquisition from an ancestral Actinobacterial species. Similarly, the presence of *Staphylococcus* and *Bacillus* within Clade IV is indicative of an independent ancestral gene transfer of LipM from an ancestral Firmicute. Lastly, in Thermoproteales and Aigarchaeota, LipB displays a monophyletic relationship with *E. coli* and *B. subtilis* (Proteobacteria) LipB (Fig. S3). These observations are most parsimoniously explained by ancestral acquisition events although it is difficult to exclude the possible effects of historical HGT amongst archaean taxa on the contemporary phylogenetic distribution of these systems. The evolutionary origins of the LplA system in archaea are more difficult to reconstruct as archaean LplA sequences are very distantly related to LplA in eubacteria and eukaryotes. However, this observation in conjunction with the deep evolutionary branches across numerous LplA clades is most consistent with ancient origins of archaean LplA systems and potential loss during archaean evolution.

To explicitly examine the prevalence of HGT in the evolution of archaean lipoylation systems, we performed codon usage bias analysis of all 147 archaean genomes in our dataset. The modal codon usage method revealed that 14.1% (20 of 142; p<0.10) of lipoylation genes show significantly different codon usage from the genomic mode, consistent with horizontal gene transfer events (Table S1). The frequency of significant HGT events involving lipoylation genes is significantly lower than the observed genome average (34.5% across 147 genomes; p<0.0001) and recently published estimates [41]. Although putative HGT events were distributed across many archaean orders (Fig. 2), these putative events were significantly concentrated amongst the 20 species possessing multiple copies of lipoylation genes in their genomes ($X^2 = 3.84$, $p = 0.025$). Amongst the 20 putative HGT events, 13 also occur amongst closely related taxa as evidenced by their phylogenetic proximity (Fig. 2). For example, three HGT events involving LipB were found to be concentrated in Thermoproteales (T. acidophilum str. IM2, *P. calidolerum*, JCM 11548, and *P. aestuarii* TE7) and these genes are closely related to all other Thermoproteales LipB genes (Fig. S3). A second possible hallmark of HGT events between closely related taxa would be the acquisition of a second copy of the same lipoylation gene from a closely related sister taxon. This was observed in 9 out of 10 events in species with multiple lipoylation genes (Table S2). Taken together, HGT is not particularly prevalent amongst lipoylation genes (in comparison to the genome average) and the enrichment of recent HGT events in taxa with multiple genes is more consistent with transient increases in copy number (and the potential establishment of functional redundancy), which are subsequently returned to a single-gene state by gene loss.

**Lipoylation Pathway Retention in Archaea**

Genomic streamlining in archaea has been well documented [8,9] and may extend to ancestrally acquired lipoylation pathways in archaea. Consistent with this prediction, the majority of archaea capable of lipoylation (79%, 67 of 85 species) exclusively retain only one transferase system: either LplA or LipM or LipB (Table S2). Sulfurospirales species, of which 8 out of 16 are isolates of *S. islandicus*, are the primary exception, possessing multiple copies of both LipA and LipM. As mentioned previously, only 9 species (excluding Sulfurospirales) retain multiple lipoyl transferase genes. None of these cases includes the retention of multiple distinct lipoylation transferase systems (for example, *LipM* and *LplA* or *LipM* and *LipB* and 6 of these display evidence of being the result of a recent HGT event. These observations, in conjunction with our phylogenetic analyses, are consistent with ancestral lipoylation system acquisition events in archaea. Furthermore, the marginally greater retention of *LipM* and *LipB* (53 species) relative to *LipA* (50 species) is also noteworthy as it was previously thought that FA biosynthesis, the source of octanoic acid, was absent (or taxonomically restricted) in archaea. Contrary to this view, the widespread identification of octanoyl transferases provides strong complementary support for FA biosynthesis across diverse archaea [27].

**Correspondence between Lipoylation Pathways and OADHC Substrate E2**

Addition of a lipoyl moiety to the E2 subunit of OADHCs is essential for aerobic metabolism. We therefore catalogued the presence of E2 across archaean genomes and, as expected, our results show a widespread correlation between the presence of E2 and lipoylation systems. Specifically, all species (9 out of 9) possessing *LipB* have an intact lipoylation pathway, also possessing lipoyl synthetase LipA and E2. Similarly, 82% of species (36 of 44)
possessing LipM also possess LipA and E2. Exceptions to this include 7 Sulfolobales (A. hospitalis W1, all three Metallosphaera species, S. acidocaldarius DSM 639, S. islandicus REY15A, and S. tokodaii str. 7) and C. haloredivivus sp. G17. In contrast, only 34% of species (17 of 50) with LplA possess E2; 10 Sulfolobales, A. pernix, 3 Thermococcus species, and both Thermoplasma species. The absence of E2 in most Thermococcus species is perhaps explained by the presence of an alternative lipoylation target, the glycine cleavage system protein H (GcvH). A comprehensive bioinformatic search for potential lipoylation domains in the genomes of the remaining 20 species without E2 or GcvH revealed a complete absence in 19 of their genomes (the exception being S. acidocaldarius DSM 639). This observation is likely explained by a transferase function involving substrates with cryptic lipoylation domains, although it is possible that these represent obsolete lipoylation systems that may be subject to loss through genome reduction mechanisms. Overall, the strong correspondence between E2 and lipoylation systems, particularly in the case of LipM and LipB, suggests a conserved aerobic metabolic functionality of lipoylation systems in these taxa.

Absence of Lipoylation Systems in Anaerobes

The absence of a lipoylation pathway in 62 of the 147 species surveyed raised the possibility that loss of lipoylation pathways might be concentrated amongst anaerobes. In support of this assertion, correlated lipoylation pathway absence (including transferase/ligase enzyme and its substrate) was observed in all Thaumarchaeota, Nanoarchaeota, Methanopyrales, Methanobacterales, Methanococcales, Archaeoglobales, Methanosarcinales, and Methanomicrobiales, all of which are characterized as obligate anaerobes (Fig. 3). Strikingly, amongst methanogens, all sequenced Methanocellales retain LipM and LipA and have demonstrated oxygen stress tolerance [42,43,44]. This raises the possibility that this lipoylation pathway may not be solely associated with energy metabolism per se, but rather may be part of a pathway to survive periodic exposure to aerobic conditions. A particularly compelling example that supports selective retention of lipoylation is LipM Clade III where the presence and phylogenetic proximity of LipM in Halobacteriales and Methanocellales is consistent with acquisition of LipM prior to the divergence between Methanogen Class II and Halobacterial lineages, followed by loss in other obligately anaerobic Class II Methanogens (Methanosarcinales and Methanomicrobiales). A similar pattern is observed in LplA Clade I and Clade II (Fig. 2) where LplA sequences from Desulfurococcales, Sulfolobales, and Thermococcales (and Thermoplasma in the case of Clade I) cluster monophyletically. This is most parsimoniously explained by presence of LplA in the ancestor of Crenarchaeota and
Euryarchaeota with subsequent loss in lineages leading to the Methanogens, Archaeoglobales and Halobacteriales (Fig. 3).

Lipoylation in Aerotolerant Archaea

Four archaeal orders displayed variable retention of lipoylation systems and, amongst these, there was a strong correspondence between retention of a lipoylation enzyme and its E2 substrate, as would be predicted by their biochemical relationship. In Thermoproteales, all six species retaining $E_2$ also possess $\text{LipB}$ (Fig. 4a), with five of these from the *Pyrobaculum* genus. *Pyrobaculum* species are metabolically versatile and grow under both aerobic and anaerobic conditions [45,46], with the sole exception of *P. islandicum*, a strict anaerobe in which the absence of lipoylation capability is most parsimoniously explained by gene loss [47].

Halobacteriales can generally tolerate aerobic conditions, consistent with the widespread retention of $\text{LipM}$ and $E_2$ among Halobacteriales (Fig. 4b). A correlated loss of both $\text{LplM}$ and $E_2$ was observed in two Halobacteriales: *Halorhabdus tiamatea*, an anaerobe that inhabits anoxic deep sea brine [48], and *Haloquadratum walsbyi*, which inhabits essentially anoxic environments due to their extremely high salinity [49].

Despite meta-level correspondence between anaerobic metabolism and $E_2$ lipoylation loss, notable exceptions were identified. As mentioned previously, *LplA* appears to have been retained in 9 of the 10 sequenced anaerobic Desulfurococcales species, despite widespread $E_2$ loss (Fig. S4). Strikingly, *A. pernix*, the only sequenced strictly aerobic Desulfurococcales [50], retains *LplA* and the $\text{OADHC}$ operon. *LplA* has been retained in all sequenced species of Thermococcales but only three possess $E_2$, a finding consistent with anaerobic conditions for most Thermococcales (Table S1). Unlike the better studied Crenarchaeota and Euryarchaeota, restricted genomic and environmental data exist for the more recently identified Korarchaeota, Aigarchaeota and Thaumarchaeota [51,52,53 and reviewed by 54]. Consistent with a general correspondence between $E_2$ lipoylation and aerobic metabolism, analysis of the strictly anaerobic *Korarchaeum cryptofilum* [51], the only Korarchaeota species with an available genome, revealed an absence of lipoylation and $\text{OADHC}$ genes. *Caldarchaeum subterraneum*, the only representative of the proposed phylum Aigarchaeota, has features distinguishing it from the Thaumarchaeota, such as a ubiquitin-like protein modifier system [52] and genes encoding LipB and OADHC components. Previous analysis of the *Caldarchaeum subterraneum* genome suggested versatile energy metabolism [52] including an almost complete Emden-Meyerhof pathway and a complete citric acid cycle, and our identification of an $E_2$ lipoylation pathway is consistent with aerobic metabolism. In contrast to these Korarchaeota and Aigarchaeota examples, Thaumarchaeota species lack lipoylation and $\text{OADHC}$ genes despite the fact that they inhabit diverse environments, ranging from aerated soils to oxygen-depleted marine sediment [54]. This observation is explained by their ability to oxidize ammonia (and potentially related substrates) and the adaptation of Thaumarchaeota ecotypes to diverse abiotic conditions, including low ammonia and low oxygen environments [55,56]. As such, this largely autotrophic basal archaeal clade has the unique biological ability to oxidize reduced nitrogen species and presumably has no evolutionary reliance on aerobic metabolic pathways associated with OADHC complexes.

Genomic Heterogeneity of the LplA Lipoylation System

In contrast to widespread reductions of genome complexity in archaea, including $E_2$ lipoylation loss across diverse anaerobes,
our analyses also revealed substantial \textit{LplA} copy expansion and heterogeneity across a restricted set of species possessing this gene. As a likely result of gene duplication events, \textit{LplA} and \textit{LipM} copy number varies across Sulfolobales and \textit{A. pernix}. This variation may ultimately prove to be an adaptive response relating to the availability of exogenous lipoic acid and endogenous octanoic acid. Our analysis also confirmed the previous identification of a single Sulfolobale \textit{LplA} copy encoding both N- and C-terminal domains of \textit{LplA} and identified similar genes in a Thermococcale (\textit{gammatolerans EJ3}), a Halobacteriale (two copies in \textit{turkmenica DSM5511}) and a Desulfurococcale (\textit{A. pernix}). Previous phylogenetic analyses have supported the proposition that a bipartite gene system, comprising \textit{LplA-N} and \textit{LplA-C}, predates the origin of the \textit{LplA} gene found in most bacteria and eukaryotes that encode both \textit{LplA} domains [23]. The widespread presence of the bipartite \textit{LplA-N/C} system in archaea, and evidence presented supporting its ancient evolutionary origins, are consistent with this scenario and are further supported by our identification of numerous bacterial species, such as \textit{Bordetella}, \textit{Achromobacter} and \textit{Rhodanobacter}, which also possess a bipartite \textit{LplA} gene system (data not shown). It is therefore possible that distinct genes encoding the N- and C-terminal domains have formed chimeric proteins in several archaeal lineages. Relevant to the possibility of chimeric fusions is the observation that genomic rearrangements have resulted in co-localization of \textit{LplA-N} and \textit{LplA-C} multiple times during archaeal evolution. Transcriptional coupling of \textit{LplA-N} to \textit{LplA-C} is present in \textit{T. acidophilum} (supported by out-of-frame coding sequence overlap (1 base pair), a readily identifiable TATA box upstream of the \textit{LplA-C} gene, and the absence of identifiable cis-regulatory sequences proximal to the 5’ end of \textit{LplA-N} [24]) and our analysis revealed an independent origination of transcriptional coupling of \textit{LplA-N} to \textit{LplA-C} genes in five Desulfurococcales. The monophyletic relationship amongst these species is consistent with the co-localization of these genes in their common ancestor (Fig. S4). Therefore, \textit{LplA} and \textit{LipM} gene duplications appear to occur in a restricted set of taxa and targeted experiments will be necessary to assess a possible association of this with differential oxygen tolerance capacities amongst these species.

**Discussion**

Metagenomic sequencing of microbial communities has progressed beyond the initial goals of assessing species composition to the more penetrating proposition that biotic and abiotic
interactions can be modelled based on metagenomic data. Given the overwhelming complexity of such ecological and environmental interactions, the accuracy of such inferences needs to be initially assessed using relatively straightforward interactions, mediated by well-studied pathways, in a set of organisms likely to exhibit marked diversity in the relevant interactions. We have therefore investigated aerobiciosis capacity across a diverse set of archaeal genomes using the well-characterized enzymes responsible for OADHC lipoylation. The potential of using metagenomic data to establish links between metabolic capabilities and environmental conditions is of particular importance to archaea, which are often difficult to culture in the laboratory and therefore remain refractory to direct analysis [57].

OADHC lipoylation is essential for metabolism in aerobic bacteria and eukarya, making it a compelling candidate system to assess the potential for more expansive metagenomic analyses across the archaea. Our analysis revealed three broad trends, which together suggest that metagenomic inferences have the potential to be informative when well-understood pathways are interrogated in organisms possessing relevant environmental/ecological diversity. First, the retention of a single lipoylation pathway (LplA, LipB or LipM) in species capable of lipoylation is consistent with genome streamlining during archaeal evolution. As such, the presence of genes, pathways or networks within archaeal genomes (and the concomitant absence of redundancy) can be generally attributed to the selective retention of essential functions. Second, the rather widespread presence of a LipM/LipB-LipA system provides support for the presence of FA biosynthesis and endogenous octanoic acid across a surprisingly diverse range of archaea. Third, OADHC E2 lipoylation has been consistently lost in obligate anaerobes and may therefore serve as a diagnostic metagenomic marker for aerobicosis. Similarly, the presence of the lipoylation/OADHC system in organisms previously thought to be strict anaerobes may indicate the existence of mechanisms for oxygen tolerance, but may also reflect previously unrecognized aerobic respiration capabilities. In addition to our observation, aerotolerance has been attributed to superoxide reductase in some species of Methanosarcinales, indicating that distinct mechanisms leading to oxygen stress adaptation may exist [58]. As metagenomic approaches often result in fragmented genome sequences, inferring gene or pathway absence may be difficult, making arguments based on gene presence (in this case indicating aerobicosis or oxygen tolerance) more reliable. It should be noted that our analysis relied largely upon complete genomes, although uncertainty associated with the analysis of incomplete sequences may still apply for species derived from ecological samples. In conclusion, given highly variable retention of gene repertoires across the archaea, extension of comparative genomic approaches to broader metabolic and homeostasis networks should be useful in revealing genome-wide characteristics related to archaean adaptation to diverse environments.

Our analysis demonstrates that the evolution of archaean lipoylation systems is generally in agreement with major trends identified in recent reconstructions of archaean genome evolution [9]. An increase in genomic complexity (the innovation phase) is evidenced by multiple lipoylation system acquisitions that have involved all the primary lipoyl-octanoyl transferase systems (LplA, LipM and LipB). It is noteworthy that, based on our phylogenetic analyses, these systems are inferred to be largely cubacterial in origin and appear to have been acquired from a diverse range of bacterial phyla, including Firmicutes, Actinobacteria and Proteobacteria. This period of increased complexity in lipoylation genetics was then followed by a reductive phase where lipoylation systems were lost across a diverse range of archaean species, notably those that have become adapted to an obligately anaerobic life history. Gene loss may therefore have played a prominent role in the functional diversification of archaea during their adaptation to, and exploitation of, diverse and often extreme habitats.

Supporting Information

Figure S1 LplA-N protein alignment. LplA-N sequences were included from one representative species for each of the six archaeal orders in Fig. 1 that retain LplA (P. horikoshii, Pyrococcus horikoshii, a Sulfolobale; M. arvoryzae, Methanocellale; H. butylicus, Hyperthermus butylicus, a Desulfurococcale; S. solfataricus, Sulfolobus solfataricus, a Sulfolobale; T. acidophilum, Thermoplasma acidophilum, a Thermoplasmatale; N. pharaonis, Natronomonas pharaonis, a Halobacterales). Conserved amino acid residues are highlighted. Secondary structure elements (α-helices $\alpha_1$ to $\alpha_8$, β-strands $\beta_1$ to $\beta_{10}$, and 310-helices $\eta_1$ to $\eta_3$) from T. acidophilum LplA-N in the structure of the T. acidophilum LplA-N-LplA-C complex (PDB code 3R07) are shown.

Figure S2 Maximum Likelihood Phylogeny of LplA and LipM. Phylogenetic tree including 132 LplA and LipA sequences from archaea, LplA sequences from major eukaryotic species (S.cerevisiae, D. melanogaster, M. musculus and H. sapiens) and LipA and LipM sequences from euabacteria representing actinobacteria (S. coelicolor), bacteroidetes (B. thetaotaomicron), firmicutes (B. subtilis and S. aureus) and proteobacteria (E. coli and B. pseudomallei). Putative horizontal gene transfer events are indicated (asterisks). Bootstrap values are provided and species abbreviations can be found in Supplemental Table S1.

Figure S3 Maximum Likelihood Phylogeny of LipB. Phylogenetic tree including 13 LipB sequences from archaea and LipB sequences from major eukaryotic species (S.cerevisiae, D. melanogaster, M. musculus and H. sapiens) and euabacteria representing actinobacteria (S. coelicolor), bacteroidetes (B. thetaotaomicron) and proteobacteria (E. coli and B. pseudomallei). Putative horizontal gene transfer events are indicated (asterisks). Bootstrap values are provided and species abbreviations can be found in Supplemental Table S1.

Figure S4 Desulfurococcales retention of LplA and OADHC operons. The ancestral lineage of the five monophyletic species (red cross) where LplA-N and LplA-C are inferred to have become co-localized and transcriptionally coupled. (EPS)

Table S1 Lipoylation systems and substrates in Archaea. A complete inventory of the species analyzed, the lipoylation related proteins identified and evidence supporting horizontal gene transfer. (XLS)

Acknowledgments

We would like to thank Navroz Bhopti and Natalie Ring for analytical assistance during the preparation of this manuscript and several anonymous reviewers for their insightful comments.

Author Contributions

Conceived and designed the experiments: SD KB SB. Performed the experiments: KB MP AU. Analyzed the data: SD KB SB MD. Wrote the paper: SD SB.
References

1. Wilmes P, Simmons SL, Denev VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev 33: 109–132.

2. Tyson GW, Chapman J, Hagiholz P, Allen EE, Ram RJ, et al. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature 428: 37–43.

3. Tully BJ, Nelson WC, Heidelberg JF (2012) Metagenomic analysis of a complex marine planktonic thaumarchaeal community from the Gulf of Maine. Environ Microbiol 14: 254–267.

4. Hallam SJ, Putnam N, Preston CM, Dungan MJ (2009) A unique mechanism of aminoacyl tRNA synthesis in Escherichia coli. PLoS Biol 7: e1000147.

5. Marx CJ (2009) Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol 26: 2087–2095.

6. Wolf YI, Kosuri N (2013) Parallel evolutionary dynamics of adaptively divergent archaea. Genetica 143: 213–226.

7. Schonauer MS, Kastaniotis AJ, Kursu VAS, Hiltunen JK, Dieckmann CL (2012) Distinct origins of the protein ligase A gene of rice (Oryza sativa). Mol Biol Evol 29: 1851–1859.

8. Posner MG, Upadhyay A, Bagby S, Hough DW, Danson MJ (2009) Lipoic acid synthesis and attachment in yeasts. Mol Microbiol 72: 605–616.

9. Reche PA (2000) Lipoylating and biotinylating enzymes contain a homologous multidomain lipoyl domain: implications for the evolution of protein ligases. Biochem Soc Trans 28: 9–13.

10. Perham RN (2000) Swinging arms and swinging domains in multifunctional enzymes: catalytic machines for multistep reactions. Annu Rev Biochem 69: 995–1030.

11. Christensen QH, Martin N, Mansilla MC, de Mendoza D, Cronan JE (2011) A novel amidotransferase required for lipoic acid cofactor assembly in Bacillus subtilis. Mol Microbiol 80: 350–363.

12. Christensen QH, Martin N, Mansilla MC, de Mendoza D, Cronan JE (2012) Lipoic acid metabolic and transcriptomic analysis of OAHDC lipoylation in Archaea. PLoS Biol 10: e1001390.

13. Christensen QH, Cronan JE (2012) Lipoic Acid Synthesis: A New Family of Octanoyltransferases. Adv Microb Physiol 59: 1–147.

14. Pester M, Schleper C, Wagner M (2011) The Thaumarcheota: an emerging view of their phylogeny and ecophysiology. Curr Opin Microbiol 14: 300–306.

15. Schleper C, Nicol GW (2010) Amphitrichous archaea: physiology, ecology and evolution. Adv Microb Physiol 57: 1–41.

16. Elkins JG, Podar M, Graham DE, Makarova KS, Wolf Y, et al. (2008) A large phylogenies by maximum likelihood. Syst Biol 52: 696–704.

17. Chen Y, Conrad R, Lu Y (2011) Distinct gene set in two different lineages of ammonia-oxidizing archaea: implications for the evolution of ammonia-oxidizing archaea. Mol Microbiol 80: 334–346.

18. Chapman-Smith A, Cronan JE, Jr. (1999) The enzymatic biotinylation of the archaeal lipoamide dehydrogenase. J Biol Chem 274: 36345–36350.

19. Reche PA, Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P (2008) Mesophilic and thermophilic archaea: proposal for a third archaeal phylum, the Thaumarcheota. Nat Rev Microbiol 6: 619–627.

20. Kratzer C, Welte C, Dorner K, Friedrich T, Deppenmeier U (2011) Pseudomonas stutzeri: an emerging view of their phylogeny and ecophysiology. Curr Opin Microbiol 14: 300–306.

21. Homberg J, Lopez-Garcia P, Moreira D (2012) A CDP-independent fatty acid synthesis pathway in archaea: implications for the origin of phospholipids. Mol Microbiol 85: 326–335.

22. Grogan K, Atomi H, Ezaki S, Yano K, Kimura T, et al. (2004) Crystal structure of lipoate-protein ligase A from Pyrobaculum aerophilum. J Biol Chem 279: 52013–52019.

23. Bolhuis H, Barrientos C, Huber R, Noël MB, et al. (2008) Thaumarchaeota: a new class of superoxide reductase containing an iron-sulfur cluster. FEMS J 278: 442–451.

24. Hosp M, Sczyrba A, Egan R, Kim TW, Chokhavala H, et al. (2011) Metagenomic discovery of biomasses-degrading genes and genomes from cow manure. Science 331: 463–466.

25. Fuchs M, Corggi PC, Eberhardt RY, Mistry J, Tate J, et al. The Pfam protein families database. Nucleic Acids Res 40: D290–301.

26. Katoh K, Asimenos G, Toh H (2009) Multiple alignment of DNA sequences with MAFFT. Methods Mol Biol 517: 39–64.

27. Wheeler T: Large-scale neighbor-joining with NJ: In: Proceedings of the 9th Workshop on Algorithms in Bioinformatics. Salzberg S, Warnow T, editors; 2009. Berlin. Springer. 373–389.

28. Christensen QH, Cronan JE (1995) Lipoic acid metabolism in Escherichia coli: the lipA and lipB genes define redundant pathways for ligation of lipoyl groups to apoirotin. J Bacteriol 177: 1–10.

29. Denef VJ, Banfield JF, Simonis A, Cronan JE, Jr. (1999) The enzymatic biotinylation of the archaeal lipoamide dehydrogenase: a post-translational modification of exceptional specificity. Trends Biochem Sci 24: 359–363.

30. Christensen QH, Cronan JE (2010) Lipoic Acid Synthesis: A New Family of Octanoyltransferases Generally Annnotated as Lipoate Protein Ligases. Biochem Soc Trans 38: 1002–1006.

31. Fujisawa K, Toma S, Okamura-Ikebuchi K, Motokawa Y, Nakagawa A, et al. (2005) Crystal structure of lipoate-protein ligase A from Escherichia coli: determination of the lipoic acid-binding site. J Biol Chem 280: 33645–33651.

32. Guignard S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.

33. Davis J, Olsen GJ (2009) Modul coding usage: assessing the typical codon usage of a genome. Mol Microbiol 72: 800–810.

34. Morris TW, Reed KE, Cronan JE (1994) Identification of the gene encoding the archaeal lipoamide dehydrogenase. J Biol Chem 269: 5357–5364.

35. Chapman-Smith A, Cronan JE, Jr. (1999) Lipoic acid metabolism in Archaea: a post-translational modification of exceptional specificity. Trends Biochem Sci 24: 359–363.

36. Christensen QH, Cronan JE (2010) Lipoic Acid Synthesis: A New Family of Octanoyltransferases Generally Annnotated as Lipoate Protein Ligases. Biochem Soc Trans 38: 1002–1006.