The ISME Journal (2015) 9, 1280–1294

www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics

Zheng-Shuang Hua[1,3], Yu-Jiao Han[1,3], Lin-Xing Chen[1,3], Jun Liu[1], Min Hu[1], Sheng-Jin Li[1], Jia-Liang Kuang[1], Patrick SG Chain[2], Li-Nan Huang[1] and Wen-Sheng Shu[1]

[1]*State Key Laboratory of Biocontrol, Key Laboratory of Biodiversity Dynamics and Conservation of Guangdong Higher Education Institutes, College of Ecology and Evolution, Sun Yat-sen University, Guangzhou, PR China and* [2]*Metagenomics Applications Team, Genome Science Group, Los Alamos National Laboratory, Los Alamos, NM, USA*

High-throughput sequencing is expanding our knowledge of microbial diversity in the environment. Still, understanding the metabolic potentials and ecological roles of rare and uncultured microbes in natural communities remains a major challenge. To this end, we applied a 'divide and conquer' strategy that partitioned a massive metagenomic data set (>100 Gbp) into subsets based on K-mer frequency in sequence assembly to a low-diversity acid mine drainage (AMD) microbial community and, by integrating with an additional metatranscriptomic assembly, successfully obtained 11 draft genomes most of which represent yet uncultured and/or rare taxa (relative abundance <1%). We report the first genome of a naturally occurring *Ferrovum* population (relative abundance >90%) and its metabolic potentials and gene expression profile, providing initial molecular insights into the ecological role of these lesser known, but potentially important, microorganisms in the AMD environment. Gene transcriptional analysis of the active taxa revealed major metabolic capabilities executed *in situ*, including carbon- and nitrogen-related metabolisms associated with syntrophic interactions, iron and sulfur oxidation, which are key in energy conservation and AMD generation, and the mechanisms of adaptation and response to the environmental stresses (heavy metals, low pH and oxidative stress). Remarkably, nitrogen fixation and sulfur oxidation were performed by the rare taxa, indicating their critical roles in the overall functioning and assembly of the AMD community. Our study demonstrates the potential of the 'divide and conquer' strategy in high-throughput sequencing data assembly for genome reconstruction and functional partitioning analysis of both dominant and rare species in natural microbial assemblages.
The ISME Journal (2015) **9,** 1280–1294; doi:10.1038/ismej.2014.212; published online 7 November 2014

Introduction

Microorganisms are critical to the functioning of virtually all ecosystems on our planet (Harris, 2009; Jiao *et al.*, 2010), yet we know little about their precise ecological and functional roles in the community (Prosser *et al.*, 2007). This hurdle is mainly caused by the high biodiversity of most microbial assemblages (Tiedje, 1994) and the uncultivable properties of the majority of microbes from the environment (Pace, 1997). With the benefit of cultivation-independent metagenomics approaches, recent studies have successfully reconstructed the genomes of dominant members in the communities (Tyson *et al.*, 2004; Jones *et al.*, 2011; Mason *et al.*, 2012), and advanced our understanding of the metabolic potentials and functional significance of microbes *in situ*. However, natural microbial communities are typically composed of a few dominant species followed by a large number of rare taxa (Sogin *et al.*, 2006). These low-abundance organisms may be representatives of novel microbial lineages (Castelle *et al.*, 2013; Kantor *et al.*, 2013) and play crucial roles in biogeochemical cycles and overall metabolic fluxes (Musat *et al.*, 2008; Wrighton *et al.*, 2012). Moreover, the evenness patterns of low-abundance taxa are important in defining microbial ecosystem dynamics (Huber *et al.*, 2007). Nevertheless, the nature and complexity of information in metagenomic data sets and insufficient sequencing depth and computing resources make it difficult to capture the genomic information and ecological

Correspondence: L-N Huang, College of Ecology and Evolution, Sun Yat-sen University, Guangzhou 510275, PR China
or W-S Shu, College of Ecology and Evolution, Sun Yat-sen University, Guangzhou 510275, PR China.
E-mail: eseshln@mail.sysu.edu.cn or shuws@mail.sysu.edu.cn.
[3]These authors contributed equally to this work.
Received 17 February 2014; revised 21 July 2014; accepted 21 September 2014; published online 7 November 2014

roles of low-abundance populations (Sharon *et al.*, 2013). Furthermore, metagenomics provides no information concerning the dynamic expression and regulation of genes in the environment. Recently, metatranscriptomics approaches have been used to reveal the community-wide gene expression profiles and ecophysiology of natural microbial assemblages (Hewson *et al.*, 2009; Ottesen *et al.*, 2011). However, direct analyses of both DNA and RNA sequence pools from the same communities are few (for example, Frias-Lopez *et al.*, 2008; Shi *et al.*, 2009; Stewart *et al.*, 2011), although coupled community genomic and transcriptomic analysis has the potential to discover and characterize the relative transcriptional levels of a large number of genes (Frias-Lopez *et al.*, 2008), and unravel the functional diversity and ecological partitioning in microbial communities.

Acid mine drainage (AMD) environments have great advantages for the study of microbial community structure and function because of its biological and geochemical simplicity (Baker and Banfield 2003; Denef *et al.*, 2010). Mine tailings represents a major source of AMD via microbially mediated oxidative dissolution of sulfide minerals. The tailings of Fankou Pb/Zn mine in Guangdong, South China, has been intensively studied to reveal the phylogenetic and functional dynamics of microbial communities in the tailings acidification and AMD generation processes (Huang *et al.*, 2011; Chen *et al.*, 2013). As a routine analysis of an ongoing survey, the microbial composition of the AMD sample collected in September of 2012 (Table 1) was assessed using 16S rRNA gene cloning and sequencing (see Supplementary methods). The results showed the Bacteria and Archaea domains of the AMD community was, respectively, dominated by unclassified *Betaproteobacteria* (*Bacteria*) and *Euryarchaeota* (*Archaea*), with several other rare taxa (Supplementary Figure 1). To reveal the ecological roles and functional partitioning (that is, different functions performed by different taxa) of AMD taxa in the community, a novel pipeline was developed to reconstruct genomes for both the dominant and rare taxa from the metagenomic and metatranscriptomic data containing enormous number of read pairs (outlined in Figure 1), and the *in situ* transcriptional profiles of the active populations were analyzed.

Table 1 Physical and chemical characteristics of the AMD sample

pH	*DO*	*TOC*	Fe^{2+}	Fe^{3+}	SO_4^{2-}	*Pb*	*Zn*	*Cu*	*Co*	*Cd*	*Cr*
2.6	6.0	18.5	6.2	193	474	1.5	186	2.3	4.5	4.5	2.2

Abbreviations: AMD, acid mine drainage; DO, dissolved oxygen; TOC, total organic carbon.
Concentrations are given in $mg\,l^{-1}$ except for pH.

Materials and methods

Sampling, physicochemical and community diversity analysis

Sample collection and physicochemical analysis were conducted as previously described (Kuang *et al.*, 2013 and as detailed in Supplementary Methods). Experimental procedures for DNA and RNA extraction, rRNA subtraction, RNA amplification and complementary DNA (cDNA) synthesis are described in Supplementary Methods. Microbial diversity was analyzed by cloning and sequencing PCR-amplified 16S rRNA genes (Supplementary methods). Sequences were compared with those in the Ribosomal Database Project. Barcoded 454 pyrosequencing targeting the hypervariable V4 region of 16S rRNA genes was also conducted to evaluate the microbial community structure. Raw pyrosequencing data were processed and analyzed as previously described (Kuang *et al.*, 2013; for details see Supplementary Methods).

Metagenomic and metatranscriptomic sample preparation, sequencing and de novo assembly

For metagenomic sequencing, two libraries with insert sizes of 500 bp and 2000 bp were independently generated from the total community genomic DNA sample. For metatranscriptomic sequencing, a library with an insert size of 300 bp was generated from the cDNA sample. The three libraries were pair-end sequenced (2×101 bp) on an Illumina Hiseq 2000 instrument (Illumina, Macrogen Inc., Seoul, Korea), producing approximately 105 and 8.7 Gbp of sequences from the DNA and cDNA sample, respectively. The raw reads were quality filtered as detailed in Supplementary Methods and the results are shown in Supplementary Table 1. To reduce the extremely high memory consumption during assembly of high-throughput sequencing-based metagenomic data (Hess *et al.*, 2011), a 'divide and conquer' strategy was used to partition the quality metagenomic reads into low- and high-abundance K-mers groups according to the K-mers depth using Khmer (version 0.3; Pell *et al.*, 2012; depth: 15, K-mer: 31; reads with length less than 63 bp were removed). The high- and low-abundance K-mer reads were then, respectively, assembled at a range of K-mers (47, 51, 55, 59 and 63; Supplementary Table 2) using Velvet (version 1.1.06; Zerbino and Birney, 2008). In addition, *de novo* transcriptome reconstruction was conducted for the quality metatranscriptomic data using Trinity (Grabherr *et al.*, 2011). The 10 scaffold sets generated by the paired-ended assembly using Velvet were combined and broken into contigs by one or more continuous 'N', to stringently avoid potential chimeric scaffolds. Although this step might break valid connections in the scaffolds, we did so to remove duplicated contigs from multiple assembly in the next step, and also to benefit for the gap

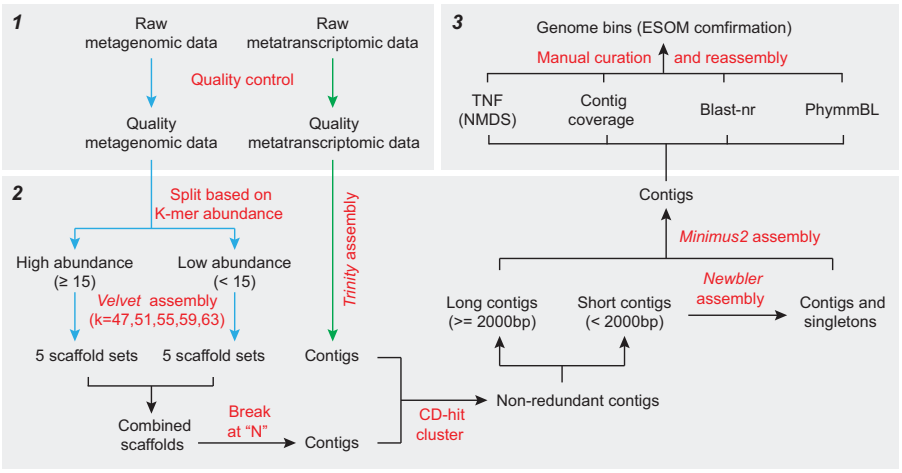

Figure 1 A schematic flow chart depicting the 'divide and conquer' strategy for the genome assembly with combined metagenomic and metatranscriptomic data. Quality control of raw data (1), sequence assembly (2) and genome binning (3) are shown. ESOM, emergent self-organizing map; NMDS, Non-metric Multidimensional scaling; TNF, tetranucleotide frequency. See details in the section of Materials and methods.

closing of continuous Ns. These contigs were then pooled with the contig set from the metatranscriptomic assembly with Trinity. CD-hit algorithms (Li and Godzik, 2006) were applied to cluster the contigs into sequence families (options: −c 0.98, −aS 1, −g 1 and –r 1), for the removal of duplicated contigs from multiple assembly. The resulting non-redundant contigs were then sorted into two pools based on their length, then Newbler and Minimus2 were used for assembly as previously reported (Mason *et al.*, 2012). Contigs with length <2000 bp were assembled using Newbler to generate longer contigs (options: −consed, −mi 98 and −ml 40). The singletons and contigs obtained from the Newbler assembly were combined with the non-redundant contigs with length ⩾2000 bp, and further assembled based on their overlap using Minimus2 (http://sourceforge.net/apps/mediawiki/amos/index.php?title=Minimus2) with a minimum percentage cutoff of 98%.

Genome binning and genome annotation

Both supervised and unsupervised approaches were used for (and to improve the accuracy of) genome binning (Strous *et al.*, 2012). For supervised classification, contigs from the Mininus2 assembly with length ⩾3000 bp (28.9 Mb in total) were compared with the National Center for Biotechnology Information (NCBI) non-redundant (nr) protein database using BLASTx (*e*-value $\leqslant 10^{-5}$). These contigs were also compared against nearly 1900 complete genomes in the NCBI RefSeq database (Pruitt *et al.*, 2007) and nine AMD draft genomes (Tyson *et al.*, 2004; Baker *et al.*, 2006; 2010) via phymmBL program (Brady and Salzberg, 2009). For unsupervised classification, tetranucleotide frequency (TNF) was first calculated for all contigs ⩾3000 bp. Then TNF-based Hierarchical Agglomerative Clustering using Euclidean Distance and *ward* criterion was performed for the contigs. To determine the best number of bins for grouping the contigs, Non-metric Multidimensional scaling analysis based on the contigs' TNF distance matrix was conducted. This resulted in a total of 11 bins (Supplementary Figure 2). Within each bin, contigs with a divergent phylogenetic assignment or coverage were manually removed or merged into another bin. Subsequently, potential chimeric contigs in each bin were detected by BLASTn against themselves. As most of the putative chimeras were from repeat regions, the contigs in each bin were shredded into 500 bp overlapped fake reads and reassembled using phrap (de la Bastide and McCombie, 2007). Finally, 11 draft genomes were obtained and the nucleotide sequences were deposited at MG-RAST under the accession numbers of 4565622.3 – 4565632.3. Emergent self-organizing map-based analysis was conducted for genome binning validation as previously described (Dick *et al.*, 2009). The completeness of the draft genomes was estimated as previously described (Hess *et al.*, 2011). For genome annotation, protein-coding genes were predicted for each draft genome using Genemark (Zhu *et al.*, 2010). Functional annotation of the predicted genes was conducted based on BLASTx analysis (*e*-value $\leqslant 10^{-5}$) against the proteins in the databases of NCBI-nr, Kyoto Encyclopedia of Genes and Genomes (KEGG) and evolutionary genealogy of genes: Non-supervised Orthologous Groups (eggNOG). The matched genes were then assigned to KEGG orthologs (KOs), KEGG pathways, KEGG categories, clusters of orthologous groups of proteins (COG) catalogs and COG categories for further analysis. Details of all analyses are given in the Supplementary Methods.

Statistical analyses

Relative transcriptional activity (RTA) of a given gene, or KO, or COG from a genome was calculated in a normalized way as:

$$\mathrm{RTA}_{ab} = \frac{\mathrm{cDNA}_{ab}}{\mathrm{DNA}_{ab}}$$

where cDNA_{ab} is the relative abundance of cDNA reads matching gene (or KO or COG) a in genome b, and DNA_{ab} is the relative abundance of DNA reads matching gene (or KO or COG) a in genome b. The relative abundance was calculated as the percentage of cDNA (or DNA) reads matching gene (or KO or COG) a in genome b dividing by the total cDNA (or DNA) reads in genome b.

For each of the selected genomes, KOs with cDNA reads counts >1.5 times the interquartile range of cDNA read counts of all the KOs in the genome were labeled as an expression outliers (Gifford *et al.*, 2013). Then the expression outliers from all genomes were combined to make a table of outlier orthologous relationships. The indicator value (IV) of each expression outlier was calculated as:

$$\mathrm{IV} = \frac{\mathrm{cDNA}_{ab}}{\sum_{b=1}^{g} \mathrm{cDNA}_{ab}} \times 100$$

where cDNA_{ab} was calculated as the number of cDNA reads matching the expression outlier KO a in genome b dividing by the total number of cDNA reads in genome b, g = number of genomes. Those KO labeled as expression outliers and with an IV >50 were identified as indicator KO (Gifford *et al.*, 2013).

Statistically significant differences in relative transcriptional activity of genes (or COG or KEGG category) across AMD taxa were determined using the non-parametric Wilcoxon rank-sum test ($P<0.05$).

Results

Physicochemical characteristics and microbial community composition and diversity of the AMD sample

The AMD sample was characterized with low pH (2.6) and total organic carbon and high levels of ferric iron, sulfate and heavy metals, such as Zn, Pb, Cu, Co, Cd and Cr (Table 1). Pyrosequencing analysis of 16S rRNA genes revealed a microbial community of very low diversity, with a total of 81 operational taxonomic units from 10 845 quality sequences assigned to *Bacteria* (99.7%) and *Archaea* (0.3%; Figure 2a). *Proteobacteria* (*Betaproteobacteria,* 94.9%; *Alphaproteobacteria*, 1.2%; *Gammaproteobacteria*, 1.0%), *Firmicutes* and *Nitrospira,* respectively, accounted for 97.1%, 0.9% and 0.6% of the community. These results were in accordance with that revealed by 16S rRNA gene clone library analysis (Supplementary Figure 1). Moreover, mapping the quality metagenomic reads to the operational taxonomic units representative sequences also identified a similar population structure of the AMD community (Figure 2a and Supplementary Figures 1 and 3).

Genome reconstruction and phylogeny

With the benefit of the 'divide and conquer' assembly strategy using metagenomic data, and by integrating with a separate metatranscriptomic assembly (Figure 1), a total of 11 draft genomes were successfully reconstructed, of which 10 were rare members (relative abundance $<1\%$) of the AMD community (Table 2, Figure 2b and Supplementary Figure 4). The taxonomic information of the genomes was then evaluated. Owing to the absence of 16S rRNA gene sequences for seven of the draft genomes (Table 2), a phylogenetic tree was built based on a concatenation of 31 universal and rarely horizontally transferred protein-coding marker

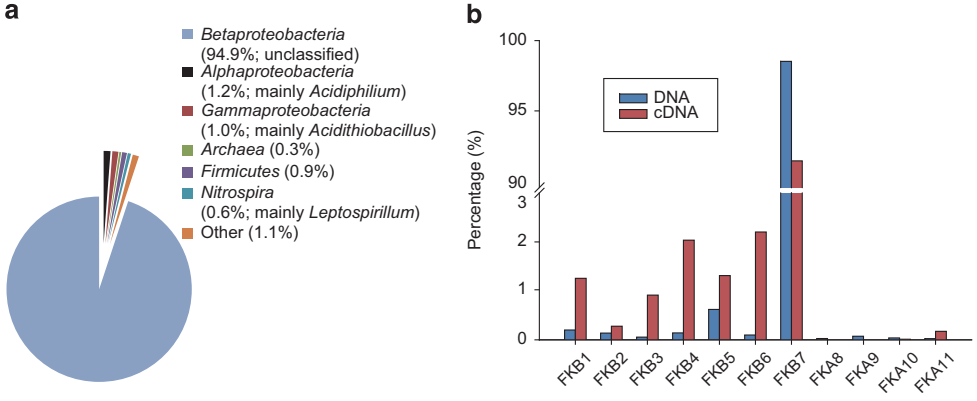

Figure 2 Relative abundance of AMD taxa as revealed by (**a**) 16S rRNA genes analysis based on 454 pyrosequencing, and (**b**) unassembled quality metagenomic and metatranscriptomic reads. See Table 2 for the most closely related organisms for the 11 AMD taxa, and see Supplementary Table 3 for the details of quality reads for each taxon. The relative abundance of a given populations was calculated as the number of quality DNA reads mapped to its corresponding draft genome divided by the total number of quality DNA reads.

Table 2 Information of the 11 draft genomes assembled from metagenomic and metatranscriptomic data of the AMD community

Draft genome	*Closely related organism*	*Total bases (bp)*	*Number of contigs*	*GC content (%)*	*Completeness (%)*[d]	*16 S rDNA sequence*
FKB1	*Acidithiobacillus ferrooxidans*[a]	2 453 291	364	58.0	80.9	No
FKB2	*Acidithiobacillus thiooxidans*[a]	2 643 526	330	53.2	85.5	No
FKB3	*Leptospirillum rubarum* (group II)[a]	2 067 509	295	57.7	77.6	No
FKB4	*Leptospirillum ferrodiazotrophum* (group III)[a]	2 593 804	235	59.0	91.8	No
FKB5	*Acidiphilium cryptum*[a,c]	3 468 182	361	66.7	91.3	Yes
FKB6	*Alicyclobacillus acidocaldarius*[a,b]	2 449 390	286	45.5	75.7	No
FKB7	*Ferrovum* spp.[c]	2 983 188	276	40.1	79.0	Yes
FKA8	*Candidatus Micrarchaeum acidiphilum* ARMAN-2[a,c]	1 266 728	179	43.5	84.9	Yes
FKA9	*Candidatus Parvarchaeum acidiphilum* ARMAN-4[a,c]	1 223 344	203	35.9	79.1	Yes
FKA10	*Candidatus Parvarchaeum acidophilus* ARMAN-5[a]	1 041 772	193	39.9	74.1	No
FKA11	*Picrophilus torridus*[a]	1 113 980	213	40.9	53.2	No

[a]The taxonomic information was achieved by comparing contigs against NCBI-nr database using BLASTx.
[b]The nearest neighbor in maximum-likelihood phylogenies constructed with RAxML (v7.2.7) by combining 31 universal marker genes was identified as the closest organism.
[c]The taxonomic information was obtained by comparing 16 S rDNA sequence against NCBI-nt database using BLASTn.
[d]The completeness of each genome was estimated by the ratio of core genes observed in each genome and the corresponding pan-genome (see Supplementary Methods for details).

genes occurring in 658 fully sequenced genomes in the STRING database (Szklarczyk *et al*., 2011; 1133 genomes initially, with only one genome in each genus selected for the phylogenetic tree construction) and the 11 draft genomes (Supplementary Figure 5). Taxonomic affiliation of the draft genomes was also evaluated by BLASTx analysis against the NCBI-nr database and phymmBL analysis of their contigs. The results showed that FKB1 and FKB2 belonged to *Acidithiobacillus* and were most similar to *At. ferrooxidans* and *At. thiooxidans*, respectively. FKB3 and FKB4 were closely related to *Leptospirillum rubarum* (group II) and *L. ferrodiazotrophum* (group III), respectively. FKB5 was related to *Acidiphilium cryptum*, and FKB6 was most similar to *Alicyclobacillus acidocaldarius*. The unclassified FKB7 harbored a complete 16S rRNA gene sequence sharing 96% similarity with that of *Ferrovum myxofaciens* strain P3G (Johnson *et al*., 2014; Supplementary Figure 6). The archaeal genomes of FKA8, FKA9, FKA10 and FKA11 were most closely related to that of *Picrophilus torridus*, *Candidatus Micrarchaeum acidiphilum* ARMAN-2, *C. Parvarchaeum acidiphilum* ARMAN-4 and *C. Parvarchaeum acidophilus* ARMAN-5, respectively. Based on genomic alignment analysis using MUMmer (Delcher *et al*., 2003), a good matching of nucleotide sequences between the draft genomes and their references was identified (Supplementary Figure 7), suggesting the validity of the assembly and genome binning processes. The taxonomic composition of the AMD community based on the 11 draft genomes was comparable to those revealed by 16S rRNA gene-based analyses (Figure 2a and Supplementary Figures 1 and 3), except for the absence of ARMAN 16S rRNA genes which could not be amplified by the universal primers used (for example, 27F/1492R, 515F/806R). Among the 11 draft genomes, FKB1 and FKB2, FKB3 and FKB4, FKA9 and FKA10 were from the genus of *Acidithiobacillus*, *Leptospirillum* and *C. Parvarchaeum acidiphilum*, respectively. To distinguish their genomic information from the contigs pool, we used the combined information including coverage, TNF and BLAST searches to known databases as detailed in Materials and methods and Supplementary Information. For the three pairs of taxa, genome coverages were 58 vs 37, 19 vs 40 and 42 vs 26, and average GC contents were 58.0% vs 53.2%, 57.7% vs 59.0% and 35.9% vs 39.9%, respectively. Bidirectional BLAST analysis revealed that 1690, 1631 and 979 ortholog genes were shared between FKB1 and FKB2, FKB3 and FKB4, FKA9 and FKA10, respectively, and the identities relative to each other were 63%, 71% and 66% on average based on amino-acid sequence similarity. Comparable similarity values at the amino-acid level have been reported between *At. ferrooxidans* and *At. thiooxidans* (about 69%, Valdés *et al*., 2008; 2011) and the acidophilic archaea ARMAN-4 and ARMAN-5 (about 71%, Baker *et al*., 2010). In contrast, FKB3 and FKB4 showed a higher similarity than that previously reported between known *Leptospirillum* spp. (Goltsman *et al*., 2009), with an average amino-acid identity over 55%.

Global analysis of metabolic potentials and gene expression

To determine the metabolic potentials and gene expression of taxa in the AMD community, gene prediction and functional annotation were conducted for the 11 draft genomes. A range of 1208–3011 genes were predicted from the genomes, of which over 73%, 65% and 64% matched the genes in the NCBI-nr, KEGG and eggNOG databases, respectively (Supplementary Table 3). The gene sets of the 11 taxa were dominated by those involved in the KEGG categories of amino acid, carbohydrate, energy and nucleotide metabolism, and translation (Supplementary Figure 8). At the finer KEGG

metabolic pathway level, different patterns of expressed genes were shown toward different taxa. Many central pathways, including oxidative phosphorylation, amino-acid biosynthesis, pyrimidine and purine metabolism, nitrogen metabolism and carbon metabolism, dominated the transcript pools (Supplementary Figure 9). Other pathways like two-component regulatory systems and ATP-binding cassette transporters were also highly expressed, indicating that the AMD taxa conducted active response and adaptation to the changing environment.

Gene expression profiles of active taxa

To further reveal the ecological roles of active taxa in the AMD community, the 10 genes with the highest relative transcriptional activity in each taxon were characterized (Table 3). As previously reported in marine microbial communities (Gifford *et al.*, 2013), there was a significant positive relationship between the expression level of a KO in a specific taxon and how commonly it was harbored in the other taxa (Supplementary Figure 10; Wilcoxon rank-sum test, $P<0.05$), suggesting that the highly expressed genes were more likely to be shared across multiple taxa. For this reason, the indicator genes representing the processes garnering the most transcriptional effort by that taxon were also characterized. The top 10 indicator genes with the highest relative abundance in the transcript pool of each taxon are shown in Supplementary Table 4. This analysis excluded FKA8, FKA9 and FKA10 due to the low occurrence of associated information in the metatranscriptomic data set (Supplementary Table 3).

The AMD taxa should harbor multiple stress resistance mechanisms to deal with the extreme environmental conditions (for example, low pH and high levels of heavy metals). Transcriptional analysis revealed the expression of enzymes involved in the survival strategies for the resistance of low pH (Figure 3b and Supplementary Table 5), including (i) harboring a highly impermeable cell membranes to protons (hopanoid biosynthesis; COG1657), (ii) generating an inside-positive membrane potential through the uptake of potassium (COG2060, COG2216, COG2156), (iii) pumping out protons using Na^+/H^+ or K^+/H^+ antiporters, (iv) harboring cytoplasmic buffer molecules capable of sequestering protons (for example, lysine, histidine, arginine, H_3PO_4) and (v) degrading organic acids (Baker-Austin and Dopson, 2007). Similarly, the genes associated with heavy metals resistance, including those for Zn, Fe, Pb, Cd, Co, Cu and other heavy metals (Table 1), were also actively expressed in the AMD taxa. Oxidative stress represents another obstacle for survival in AMD environments (Ram *et al.*, 2005). Although no associated parameters were directly determined in this study, the expression of peroxiredoxin, superoxide dismutase and catalase genes indicated such a stress for the analyzed AMD community. On the other hand, chaperones like Hfq, DnaK, GroEL and HSP20 associated with repair of DNA and protein damage caused by the extreme conditions, were fairly highly represented in most of the bacterial taxa (Table 3).

FKB1 and FKB2. The *At. ferrooxidans*-like FKB1 and *At. thiooxidans*-like FKB2 were likely autotrophic for genes encoding the key enzymes involved in Calvin-Benson-Bassham carbon fixation cycle, including RuBisCO and *prkB* were expressed (Supplementary Table 6). The nitrogen fixation gene of *nifH* was highly expressed in FKB1 (Supplementary Tables 4 and 6). Both taxa harbored transcripts associated with glutamine synthetase for ammonium assimilation, and nitrate reductases for dissimilatory functions were highly expressed in FKB2 (Supplementary Tables 4 and 6). Both FKB1 and FKB2 could potentially oxidize sulfur as indicated by the expression of genes encoding sulfite-quinone reductase, tetrathionate hydrolase and thiosulfate:quinone oxidoreductase, whereas sulfur oxidation multienzyme complex (Sox, including *SoxAX*, *SoxB* and *SoxYZ*) system and sulfur dioxygenase were only expressed in FKB2 (Supplementary Tables 4 and 6). The highly expressed rusticyanin in FKB1 may indicate its activity in iron oxidation (Bonnefoy and Holmes, 2012; Supplementary Table 6). Notably, the high occurrence of transcripts associated with motility in FKB2, including genes for flagellar assembly and chemotaxis, suggested a much higher active motility than any other taxa (Supplementary Table 4).

FKB3 and FKB4. The *Leptospirillum*-affiliated FKB3 and FKB4 may perform carbon fixation via the novel reductive tricarboxylic acid cycle (Levicán *et al.*, 2008; Goltsman *et al.*, 2009; Table 3 and Figure 3a). The expression of *nifHDK* genes in the *Leptospirillum* group III-like FKB4 may indicate an important role in nitrogen fixation as FKB1 (Goltsman *et al.*, 2009; Supplementary Table 4). The expression of *nirD* in FKB3 indicated a potential activity of dissimilatory nitrite reduction. Although *Leptospirillum* species were reported to oxidize sulfur (Johnson and Hallberg, 2008), the sulfate assimilation function was active in both FKB3 and FKB4 for thiamine biosynthesis, and the RuBisCO-like genes identified in these two taxa may also indicate an involvement in sulfur metabolism (Ashida *et al.*, 2005). The ferrous iron oxidation-associated genes encoding cytochrome 572 (cyt_{572}) and cytochrome 579 (cyt_{579}) were expressed in FKB3 and cyt_{572} expressed in FKB4 (Supplementary Table 6; Ram *et al.*, 2005). Indicator gene analysis showed that two families of two-component regulatory systems were overexpressed in the FKB4 (Table 3).

FKB5. The FKB5 was most closely related to *A. cryptum*, which has been found to grow on the small amounts of organic carbon originating from chemoautotrophic acidophilic iron/sulfur-oxidizers

Table 3 Top 10 highest expressed genes in the eight transcriptional active AMD taxa

	FKB1	FKB2	FKB3	FKB4	FKB5	FKB6	FKB7	FKA11
S-adenosylmethionine decarboxylase	41.69			2.08	11.63	39.45		
cytochrome c	17.44	18.07			2.86		3.42	
host factor-I protein	13.29	4.86				0.93	5.26	
guanylate kinase	11.56	0.98	1.51	0.46	1.38	1.78	1.40	
preprotein translocase subunit SecY	11.09	1.50		1.54	1.72	0.38	6.58	
omega-6 fatty acid desaturase (delta-12 desaturase)	10.29	3.65				0.98	1.95	
DNA-directed RNA polymerase subunit alpha	8.52			1.44	3.04	0.97	3.28	
localization factor PodJL	7.94							
cytidylate kinase	7.50	1.81	0.20	0.38	0.67	0.71	1.66	2.48
adenylate kinase	6.74	6.64		1.98	3.39	1.28	0.54	4.12
HSP20 family protein	2.81	45.18	35.06	80.20	0.49			
chaperonin GroEL	3.44	10.72	3.78	7.51	8.59	3.40	1.24	
capsular polysaccharide transport system permease protein	2.55	7.62						
molecular chaperone DnaK	1.88	7.38	1.75	3.10	4.63		1.74	2.99
ribulose-bisphosphate carboxylase small chain	1.54	6.82			0.11		3.70	
4-diphosphocytidyl-2-C-methyl-D-erythritol kinase	1.32	6.65	4.47	0.29		1.46	2.34	
molecular chaperone HtpG	1.97	6.24			0.25	1.63	1.10	
UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase	1.94	5.85	10.30	6.94	3.48		1.98	
citryl-CoA synthetase small subunit			15.47	7.68				
cytochrome c oxidase cbb3-type subunit II			8.17	4.57			10.48	
sirohydrochlorin ferrochelatase			6.60	2.38		0.47	0.37	
flagellar hook-associated protein 2		0.67	6.11	0.62				
signal peptidase II		0.29	5.75				0.52	
UDP-glucose 4-epimerase	0.18	0.21	4.93	0.99	0.81	1.11	1.00	0.51
dnaG; DNA primase	1.94		4.69	4.68		5.58	1.27	
undecaprenyl diphosphate synthase	1.49	1.28		12.35	0.49	0.71	1.73	
cysteine synthase A			3.87	8.04	1.04	1.09		1.00
proteasome-associated ATPase			3.76	7.25				
methionyl-tRNA formyltransferase	0.21	0.49		5.37	1.16	0.08	0.49	
proteasome beta subunit			2.30	5.22				
cylphosphatase				4.92				2.53
DNA-directed RNA polymerase subunit omega		1.68			8.56		2.48	
branched-chain amino acid aminotransferase	0.90		2.50	0.91	6.05	1.77	0.34	
formate dehydrogenase subunit delta					5.90			
NADH dehydrogenase (ubiquinone) Fe-S protein 4					5.74			
type VI secretion system protein VasG			0.02	0.14	5.66			
nitrogen regulatory protein P-II 1	1.40	0.25	0.32	1.12	5.60		2.03	
phosphate transport system substrate-binding protein	2.60	1.19	0.98	0.53	5.26	20.26	1.15	
GTP cyclohydrolase I			0.46	0.60		21.14	0.63	0.07
aspartate 1-decarboxylase	2.45	0.49	0.50	0.79		18.02	6.54	
superoxide dismutase, Fe-Mn family		4.66			2.28	13.09	0.57	
4,5-DOPA dioxygenase extradiol			0.32	0.25	0.49	10.31		
glutathione peroxidase		0.33				9.37	1.89	
glucose/mannose transport system substrate-binding protein		0.13				9.19		
lactose/L-arabinose transport system substrate-binding protein						8.88		
galactokinase						8.46		
alkane 1-monooxygenase							10.32	
two-component system, OmpR family, sensor histidine kinase KdpD	0.01	0.21	0.04	0.15	0.59	0.07	9.09	
two-component system, OmpR family, KDP operon response regulator KdpE	0.26			0.35	0.19	0.03	8.96	
putative (di)nucleoside polyphosphate hydrolase	0.61	2.42			1.73		5.51	
cytochrome c oxidase cbb3-type subunit I			0.20	0.28			4.54	
oxygen-independent coproporphyrinogen III oxidase	0.18	0.85	1.37	0.56	0.31	0.33	4.31	
threonine synthase		0.95	1.11	0.93	0.51	0.29	0.47	17.73
S-adenosylmethionine synthetase	0.89	1.01	1.66	2.60	1.16	1.44	0.96	17.43
transcription initiation factor TFIID TATA-box-binding protein								15.94
elongation factor 1-alpha								11.97
transcription initiation factor TFIIB								10.45
DNA-directed RNA polymerase subunit D								9.86
DNA-directed RNA polymerases I, II, and III subunit RPABC5								9.47
nucleoside-diphosphate kinase		2.22	1.29	0.90	2.96	7.91	1.55	8.99
DNA-directed RNA polymerase subunit B								8.44
thioredoxin reductase (NADPH)	1.25	1.57	0.68		1.53	1.58	0.40	7.04

Abbreviations: AMD, acid mine drainage; NADPH, nicotinamide adenine dinucleotide phosphate.
The top 10 genes with highest relative transcriptional activity in each taxon were shown in the cells (in gray). Empty cells indicated either the genome contained no ortholog to the gene or no expression of the ortholog was detected. See Materials and methods for the calculation of relative transcriptional activity.

(Johnson and Hallberg, 2008), and reduce ferric iron coupled to the oxidation of glucose (Küsel *et al.*, 2002). Transcriptional analysis indicated a heterotrophic lifestyle of FKB5, with the expression of various transporting protein-coding genes for dissolved organic carbon resources (Supplementary Table 7). The organic carbon degradation-associated genes like formate dehydrogenase and branched-chain amino-acid aminotransferase, were among the most highly expressed (Table 3), and those for propionyl-CoA synthase and acetyl-CoA synthetases were also expressed. Ammonium assimilation may serve as an important strategy for nitrogen resource of FKB5 as evidenced by the detection of *nrgA*, *gltB* and *gltD* in the transcript pool (Supplementary Table 6). Most of the heterotrophic *Acidiphilium* spp. can oxidize reduced sulfur compounds for energy generation, while showing no nutritional requirement for

them (Johnson and Hallberg, 2008). This was likely the case in the *A. cryptum*-like FKB5 as indicated by the expression of sulfur dioxygenase.

FKB6. The FKB6 showed the highest similarity to *A. acidocaldarius*, a thermoacidophilic heterotroph capable of using multiple sugars as carbon energy source (Lauro *et al.*, 2006). Genes related to sugar and polysaccharide transporters were expressed in FKB6 (Supplementary Table 7), and it devoted more transcriptional effort to phosphate acquisition (via phosphate ATP-binding cassette transporter) than any other taxa (Supplementary Table 4). Genes in the two-component system chemotaxis family, which respond to chemical stimuli by regulating the chemotactic sensitivity and extending the range of signal transduction, were also highly expressed (Supplementary Table 4).

FKB7. The most dominant member of the community FKB7 (relative abundance > 90%, Figure 2b) was determined to be *Ferrovum* spp. (Supplementary Figure 11), a newly discovered group within *Betaproteobacteria* that has been suggested to be widely distributed iron oxidizers (Hallberg *et al.*, 2006; Kuang *et al.*, 2013; Johnson *et al.*, 2014). The RuBisCO and *prkB* genes for carbon fixation were highly expressed in FKB7 (Figure 4 and Supplementary Table 6). No nitrogen fixation-associated genes were detected, whereas the expression of a urease (encoded by ureDABJCEFG operon), and nitrate transporter and assimilatory nitrate reduction genes indicated that this predominant taxon may use urea and nitrate as alternative nitrogen resources (Supplementary Table 6). The expression of *nirB* and *nirD* indicated a potential activity for dissimilatory nitrite reduction. Notably, sulfate reduction genes were highly expressed in an energy-consuming assimilatory pathway (Figure 4), which could provide reduced sulfur for the synthesis of cysteine and methionine and a range of other metabolites. Moreover, the expression of oxidative phosphorylation-associated genes was the highest among the eight taxa (Supplementary Figure 9). The two-component signal transduction system was enriched in the FKB7 transcript pool (KdpD-KdpE; Supplementary Table 4), which might be involved in pH stress through activating the K^+ transporters for K^+ influx to inhibit H^+ influx. Besides, Na^+/H^+ antiporters and arginine decarboxylase may also be used for acid stress, and several genes encoding heavy metal exporting proteins and those for oxidative stress were also expressed (Figure 3b and Supplementary Table 5), indicating multiple stress mechanisms used for FKB7 to adapt to the extreme AMD conditions (Table 1).

To investigate the iron oxidation pathway of FKB7, we searched the draft genome for homolog genes involved in iron oxidation as reported in other iron oxidizers (Bonnefoy and Holmes, 2012) as previously conducted (Liljeqvist *et al.*, 2013), including *cyc*1/*rus*/*cyc*2 in *At. ferrooxidans*, *iro* in *At. ferrivorans*, *cox* operon in *Thiobacillus prosperus*, *fox* gene cluster in *Sulfolobus metallicus*, *foxEYZ* operon of *Rhodobacter capsulatus* SB1003, Cyt_{572}/Cyt_{579} in *Leptospirillum* spp., *PioAB* in *Rhodopseudomonas palustris* TIE-1, and *MtrAB* in *Shewanellas* spp. and *Geobacter* spp. Two genes of FKB7 were similar to *cyc*1 (encoding cytochrome c552; 34% in amino-acid similarity and 84% in query coverage) and *cyc*2 (encoding a high-molecular mass cytochrome; 27% in amino-acid similarity and 97% in query coverage; Supplementary Figure 12), both are involved in the iron oxidation in *Acidithiobacillus* spp. Moreover, in the same contig with the *cyc*1-like gene, we identified a gene with relatively low-sequence similarity to the iron oxidase *iro* in *At. ferrooxidans* and *At. ferrivorans* (32% and 26% in amino-acid similarity, 85% and 92% in query coverage), respectively. All the three putative iron oxidation genes were highly expressed in FKB7 (Figure 4 and Supplementary Table 6). As such, this predominant taxon may gain electrons from ferrous iron via the iron oxidase-like protein and transfer them to the inner membrane through cytochromes encoded by *cyc*1- and *cyc*2-like genes (Figure 4).

FKA11. FKA11, the only active archaeon detected in the AMD system, was closely related to *Picrophilus torridus*, which is well known for its ability to live at pH around 0 and harbors a smaller genome size than any other nonparasitic aerobic microorganisms growing on organic substrates (Futterer *et al.*, 2004). Our indicator gene analysis revealed the genes involved in transcription and replication dominated the transcript pool, including DNA polymerase, RNA polymerase and factors for elongation, transcription and replication (Supplementary Table 4), indicating that FKA11 may be at a rapid growth stage. The central carbon metabolism may also be active as suggested by the gene expression of isocitrate dehydrogenase, a key enzyme in TCA cycle. This corresponded to the expression of several transporters for organic carbon resources, such as oligopeptide, dipeptide, amino acid and sugar (Supplementary Table 7), which are important for heterotrophic *P. torridus*-like species (Futterer *et al.*, 2004). The acid stress resistance was likely conducted by uptake of K^+, using Na^+/H^+ antiporters and also organic acid degradation, whereas peroxiredoxin and thioredoxin reductase (nicotinamide adenine dinucleotide phosphate) genes for the response of oxidative stress were also expressed (Table 3 and Supplementary Table 5), guaranteeing the physiological activity of FKA11 in such an extreme environment (Table 1; Ciaramella *et al.*, 2005).

Discussion

'Divide and conquer' strategy for high-throughput sequencing data analysis and genome reconstruction of dominant and rare taxa

Both dominant and less dominant species are important to the overall function and dynamics of

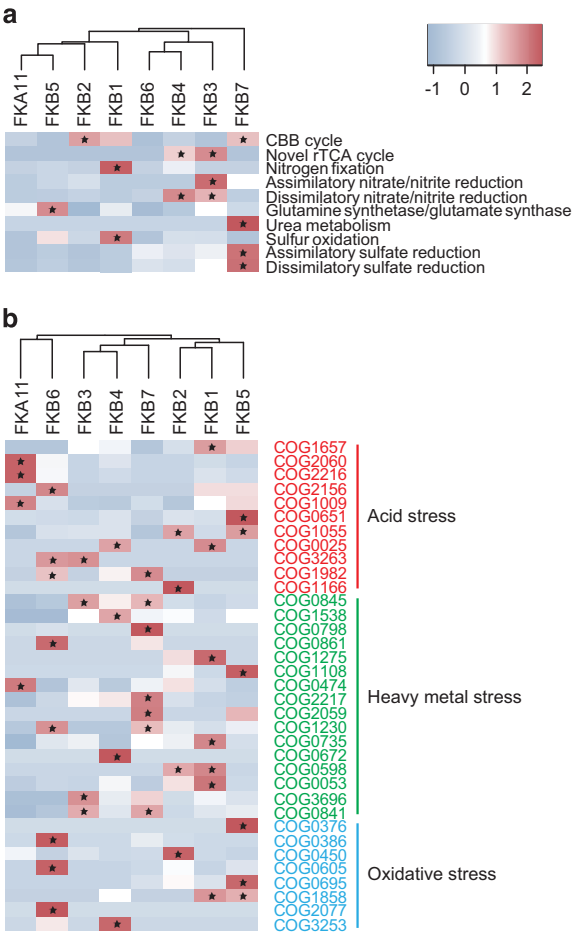

Figure 3 Profiling of (**a**) energy-related metabolisms and (**b**) COGs relevant to acid, heavy metal and oxidative stress in the eight transcriptionally active AMD taxa. The metabolisms and COGs are clustered based on their relative transcriptional activity (z-score normalized across all taxa, indicated by the color key). The relative transcriptional activity of a given metabolism is represented by the average number of relative transcriptional activity of all KOs assigned to this metabolism. Asterisk stands for significant overrepresentation of metabolism or COG in the taxon, determined by non-parametric Wilcoxon rank-sum test ($P<0.05$).

microbial communities (Huber *et al.*, 2007; Musat *et al.*, 2008; Denef *et al.*, 2010). Understanding the potential ecological roles of the rare (and often uncultured) taxa has been particularly challenging, however. Although this issue could now be explored through omics technologies with increased sequencing depth, such approaches are hampered by the subsequent assembling of large short-read sequence data sets (Sharon *et al.*, 2013). To overcome this, several recent studies have attempted targeted genome sequencing of rare taxa after specific enrichment using cell prefiltering and/or biostimulation technologies or single-cell isolation (Wrighton *et al.*, 2012; Castelle *et al.*, 2013; Kantor *et al.*, 2013; Rinke *et al.*, 2013). Although these elegant works have expanded our understanding of the phylogenetic characteristics and functional significance of these lesser known microbes in the environment, only a small fraction of the community is captured in the analyses. To gain a relatively comprehensive look at the community gene content and expression, and to reveal the functional roles of both the dominant and rare taxa in a single community, a naturally low-diversity microbial assemblage from an extreme AMD environment was selected in this study and subjected to parallel metagenomics and metatranscriptomics sequencing, generating a sequence data set of over 110 Gbp. Although lower computing memory assembler like IDBA_ud (Peng *et al.*, 2012) has been used to assemble large metagenomic data sets (Castelle *et al.*, 2013; Hug *et al.*, 2013; Kantor *et al.*, 2013), it was not feasible in our case because of the relatively limited computing resources (512 Gb RAM). Therefore, we attempted a novel 'divide and conquer' strategy using velvet in the assembly of our short-read metagenomic data set, and the results were integrated with a separate assembling of the community cDNA sequences. This has enabled a successful reconstruction of 11 draft genomes, which represent both dominant and rare and/or uncultured taxa, allowing subsequent exploration of their ecological roles and functional partitioning in the AMD community.

Genome construction and gene expression of naturally occurring Ferrovum *spp.*

Molecular investigations have documented that bacteria affiliated with the recently discovered genus '*Ferrovum*' are ubiquitous and thus likely play an important role in various AMD environments (Hallberg *et al.*, 2006; González-Toril *et al.*, 2011; Kuang *et al.*, 2013; Johnson *et al.*, 2014). Despite their wide distribution and high vitality in acidic mine waters, the isolation and cultivation of *Ferrovum* spp. have been difficult, indicating the extremely fastidious nature of these lesser known *Betaproteobacteria* (Johnson *et al.*, 2014). To date, only one pure laboratory isolate (*Ferrovum myxofaciens* strain P3G; Johnson *et al.*, 2014) and several mixed cultures (Hedrich *et al.*, 2009; Kimura *et al.*, 2011) have been reported, and genomic and gene expression information of naturally occurring *Ferrovum* spp. or laboratory isolates are lacking, precluding a comprehensive understanding of their metabolic potentials and *in situ* gene dynamics, which may provide insights into their ecological success in the environment. By deeply sequencing an AMD community predominated by unclassified *Betaproteobacteria*, we reconstructed the first genome of a natural *Ferrovum* population (that is, multiple individuals of FKB7; Figure 4, Table 2 and Supplementary Figure 11). Phylogenetic analysis based on 16S rRNA genes showed that FKB7 was similar to the few cultivated *Ferrovum* representatives (Supplementary Figure 6), and distantly related to the genus of *Nitrosospira* as previously indicated (Johnson *et al.*, 2014). Our transcriptional analysis combining the metagenomic

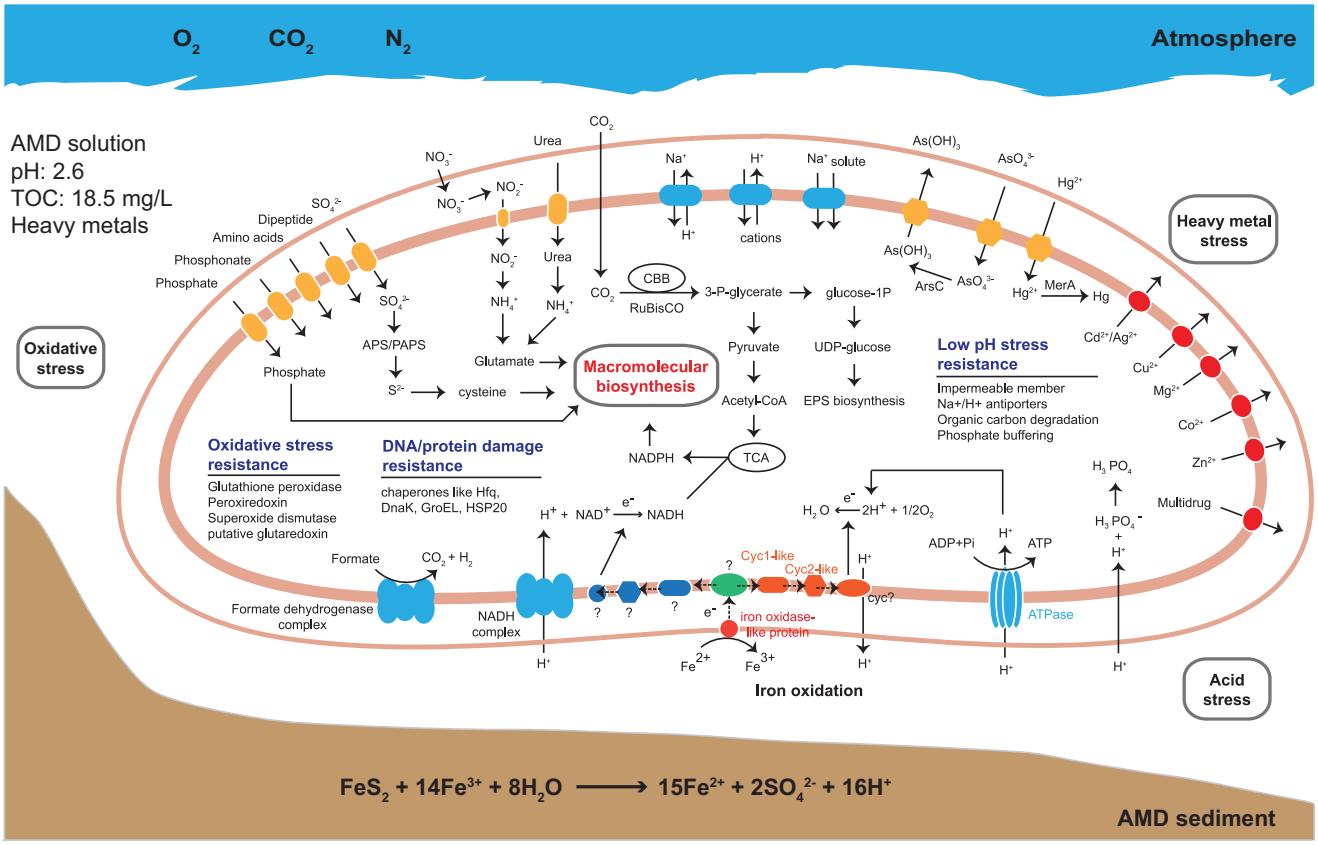

Figure 4 Metabolic abilities of *Ferrovum*-like FKB7 based on the expressed genes, which were predicted from its draft genome assembled from the metagenomic and metatranscriptomic data. Those associated with carbon fixation, nitrogen metabolism, assimilatory sulfur reduction, amino-acid biosynthesis, energy metabolism, transporters, stress response and putative iron oxidation pathway are shown.

and metatranscriptomic data indicated that the *Ferrovum*-like FKB7 could fix carbon via Calvin-Benson-Bassham cycle in the oligotrophic AMD with a low total organic carbon level, and conduct ferrous iron oxidation for energy generation with multiple putative electron transporting proteins (Figure 4). Thus, our study provides the first field evidence at the transcriptomic level for the assumption that these moderately acidophilic bacteria are obligately autotrophic and capable of growth only by ferrous iron oxidation (Rowe and Johnson, 2008; Hallberg, 2010; Johnson *et al.*, 2014). Surprisingly, although the *F. myxofaciens* strain P3G appears to be diazotrophic (Johnson *et al.*, 2014), the nitrogen fixation genes *nifHDK* were absent from the FKB7 genome (Figure 4), and urea and nitrate may instead serve as alternative nitrogen resources as evidenced by the expression of associated genes. These results indicated the high diversity and different evolutionary history of *Ferrovum* spp. Most previous attempts to isolate *Ferrovum* spp. into pure cultures failed because of the contamination of *Acidiphilium* spp. (Hedrich *et al.*, 2009; Kimura *et al.*, 2011; Johnson *et al.*, 2014), which could degrade the organic substances that are otherwise toxic to *Ferrovum* spp. Thus, it is possible that the co-occurring *Acidiphilium*-affiliated FKB5 may facilitate the dominance of FKB7 in the analyzed AMD community (Figure 2).

Ecological roles and functional partitioning of AMD taxa

The phylogenetically distinct taxa in the AMD community showed different ecological roles with specific transcriptional behaviors to coexist in the harsh environment (Table 3 and Supplementary Tables 4 and 5). The availability of carbon and nitrogen resources is vital to the microbial communities populating the oligotrophic and nitrogen-limited AMD systems (Johnson and Hallberg, 2008). The *Acidithiobacillus*-like FKB1 and FKB2 and *Ferrovum*-like FKB7 could fix carbon using the Calvin-Benson-Bassham cycle (Johnson and Hallberg, 2008; Johnson *et al.*, 2014), whereas the *Leptospirillum*-like FKB3 and FKB4 may perform carbon fixation via the novel reductive tricarboxylic acid cycle (Goltsman *et al.*, 2009). In contrast, no carbon fixation genes were expressed in FKB5 (*A. cryptum*-like), FKB6 (*Alb. acidocaldarius*-like) and FKA11 (*P. torridus*-like); these microbes may conduct heterotrophic lifestyle and obtain carbon through different kinds of transporters for DOC resources in the environment (Supplementary

Table 7). The existence of such a distinct lifestyle is crucial for the AMD community, as the coexisting heterotrophic members could consume the DOCs, which are otherwise toxic to the autotrophs (Baker-Austin and Dopson, 2007). Multiple strategies for the utilization of nitrogen resources (for example, nitrogen, ammonium, nitrate, urea and so on) were identified in the AMD taxa, as reported in other AMD environments (Ram *et al.*, 2005; Bertin *et al.*, 2011; Méndez-García *et al.*, 2014). The nitrogen fixation *nif* genes were only expressed in the *At. ferrooxidans*-like FKB1 and *Leptospirillum* group III-like FKB4, which have been found to be the major or even the only nitrogen fixers in some AMD systems (Baker and Banfield, 2003; Goltsman *et al.*, 2009). Other taxa could obtain ammonium or nitrate alternatively, and FKB7 could also use urea as nitrogen resource. All these obtained nitrogen resources could be further used for glutamine and glutamate synthesis (Ertan, 1992). Unlike some subsurface mining environments where external nitrogen is limited (Tyson *et al.*, 2004; Bertin *et al.*, 2011; Méndez-García *et al.*, 2014), the Fankou AMD site is a 'open' system and thus it is possible that part of the fixed nitrogen obtained by most of the taxa comes from external sources.

Many AMD microorganisms conduct energy conservation via the oxidation of reduced inorganic sulfur compounds (RISCs; Baker and Banfield, 2003; Johnson and Hallberg, 2003). In acidic mining environments, RISCs, including elemental sulfur (S^0), thiosulfate ($S_2O_3^{2-}$) and hydrogen sulfide (H_2S), are generated via chemical oxidation of metal sulfides (Schippers and Sand, 1999). With these RISCs in AMD environment, AMD taxa could oxidize them with multiple enzymes, sulfite-quinone reductase can catalyze the oxidation of H_2S to S^0 and thiosulfate:quinone oxidoreductase catalyzes $S_2O_3^{2-}$ to generate tetrathionate ($S_4O_6^{2-}$), whereas tetrathionate hydrolase disproportionates $S_4O_6^{2-}$ to $S_2O_3^{2-}$, SO_4^{2-} and S^0 (Dopson and Johnson, 2012). So far, the Sox system has only been fully characterized in the model organism of *Paracoccus pantotrophus*, in which *SoxXA*, *SoxYZ*, *SoxB* and *SoxCD* together can mediate the oxidation of $S_2O_3^{2-}$, SO_3^{2-}, S^0 and H_2S (Friedrich *et al.*, 2001). In contrast, the partial Sox system without *SoxCD* can only oxidize $S_2O_3^{2-}$ to produce S^0–(Hensen *et al.*, 2006). Our transcriptional analysis revealed the expression of multiple sulfur oxidation genes, including those coding sulfite-quinone reductase, thiosulfate:quinone oxidoreductase and tetrathionate hydrolase in FKB1 and FKB2, and the *Sox* system without *SoxCD* in FKB2. Partial *Sox* system has also been reported in the genus of *Acidithiobacillus*, for example, the *SoxYZ* in *At. ferrivorans* (Liljeqvist *et al.*, 2011) and *SoxAXBYZ* in *At. caldus* (Mangold *et al.*, 2011). With the activities of the sulfur-oxidizing enzymes in FKB1 and FKB2, SO_4^{2-} and S^0 may generate continuously, and the further oxidation of S^0 to SO_4^{2-} is key to the generation of acids because of release of protons in the reactions (Baker and Banfield, 2003). The resulting S^0 could be oxidized to SO_3^{2-} by several enzymes, such as the reverse function of dissimilatory sulfite reductase in green sulfur bacteria (Gregersen *et al.*, 2011), sulfur oxygenase reductase (SOR) in *At. caldus* (Mangold *et al.*, 2011) and *At. ferrivorans* (Liljeqvist *et al.*, 2011), or sulfur dioxygenase in *At. thiooxidans*, *A. acidophilum* and *A. cryptum* (Rohwerder and Sand, 2003). Although no dissimilatory sulfite reductase or SOR transcripts were detected, sulfur dioxygenase gene was expressed in FKB2 and FKB5. This indicated the key role of these two taxa in the acidification of sulfide minerals. Without their capabilities in further oxidation of S^0, deposition of S^0 would occur as previously reported in the extremely acidophilic sulfur-oxidizing biofilms (Johnson *et al.*, 2012) and thus the acid generation process may slow down. Finally, the obtained SO_3^{2-} from sulfur dioxygenase could be oxidized to sulfate via adenylylsulfate by adenylylsulfate reductase and ATP sulfurylase by FKB1 and FKB5. Ferrous iron oxidation represents another energy conservation strategy for AMD microorganisms (Baker and Banfield, 2003), but the electron transporting pathways vary in different species (Bonnefoy and Holmes, 2012). In the Fankou community, several taxa may conduct iron oxidation as indicated by the expression of rusticyanin in FKB1, cyt_{579} and cyt_{572} in FKB3 and cyt_{572} in FKB4. The cyt_{579} was not detected via the functional annotation of the FKB4 genome, but this was likely due to the absence of reference sequences of this protein in the NBCI-nr database (Goltsman *et al.*, 2009). The FKB7 may also oxidize ferrous iron with the electron transporting pathway encompassing several putative proteins. These results suggest a putative role of FKB1, FKB3, FKB4 and FKB7 in iron oxidation in the AMD system, and these iron-oxidizers are further connected with the sulfur oxidizers by providing the effective oxidant ferric iron for the chemical oxidation of metal sulfides (see above).

Our analysis also identified multiple strategies used by the AMD taxa to adapt to the extreme conditions. For acid stress resistance, the taxa (FKB1, FKB2, FKB3, FKB4 and FKB5) may harbor an impermeable membrane via the biosynthesis of hopanoid as previously described in *Acidithiobacillus* and *Acidimicrobiaceae* spp. (Jones *et al.*, 2011), or use Na^+/H^+ antiporters (all eight active taxa) that were widely detected in AMD microorganisms (Tyson *et al.*, 2004). Other acid resistance strategies that consume ATP were also detected, including uptake of K^+ in FKB1, FKB4, FKB5 and FKB6 and the use of proton sequestering molecules in FKB2, FKB3, FKB4, FKB6 and FKB7. Organic acids (for example, acetic, lactic, formic and propionic acid) were reported to be harmful to acidophiles because their protonated form could dissociate a proton in the cell (Ciaramella *et al.*,

2005), thus representing another acid stress. The *A. cryptum*-like FKB5 and *P. torridus*-like FKB11 may degrade organic acids as another stress resistance strategy by expressing the genes encoding propionyl-CoA synthase and acetyl-CoA synthetases (Baker-Austin and Dopson, 2007). Many heavy metal resistance transcripts were detected in all the bacterial taxa (COG3696, COG0841, COG1538 and COG1230), especially those for the efflux of Cd, Zn and Co, which were present at high levels in the AMD. The genes for Fe^{2+}/Zn^{2+} uptake regulation proteins were also highly expressed in all bacterial taxa, likely indicating that uptake of the two meals is carefully balanced because of their high concentrations. Few of the above-mentioned heavy metals resistance genes were expressed in the *P. torridus*-like FKA11, implying that other novel strategies may be used for this archaeon. As previously described (Ram *et al.*, 2005), peroxiredoxin was widely used by all active AMD taxa to remit harm of oxidative stress, although other associated transcripts were also detected. In addition, the highly expressed chaperones genes in all active taxa may indicate a common mechanism for repairing DNA and protein damage because of the harsh conditions.

Concluding remarks

We have attempted a novel pipeline to our short-read metagenomic and metatranscriptomic data sets, and recovered well-curated microbial genomes from both dominant and rare taxa in an extreme, low-diversity AMD system. In particular, we obtained the first genome of a naturally occurring *Ferrovum* population and reconstructed the metabolic pathways of these lesser known but ubiquitous acidophilic microorganisms. The identification and high expression of putative genes involved in iron oxidation indicated its key role in re-generating ferric iron, the primary sulfide oxidant in acidic mining environments. Our transcriptional analysis revealed multiple strategies for resource acquisition (carbon and nitrogen) and energy generation adopted by the AMD taxa through expressing both shared and taxon-specific genes. To survive in such an extreme habitat, these microbes have evolved various mechanisms to tolerate the low pH, high heavy metal concentrations and oxidative stress. Our results provide evidence that the coexistence of species is partly due to the facts that functional partitioning and rare members (for example, the N-fixing FKB1 and FKB4, and the S^0 oxidizing FKB2 and FKB5) may play important ecological roles in the community. Our study highlights the power of the 'divide and conquer' strategy for the assembly of genomes for both dominant and rare taxa from high-throughput sequencing data. With this strategy, the genomic and transcriptomic information of individual species in microbial communities along with environmental gradients or sampled at different time points could be captured, and subsequent comparative taxa transcriptional analyses may lead to a mechanistic understanding of the diverse responses of different taxa to environmental change.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Chien-Chi Lo of Los Alamos National Laboratory for his help in the bioinformatics analyses. We also thank the three anonymous reviewers for providing thoughtful and constructive comments on the manuscript. This work was supported by the National Natural Science Foundation of China (4093212, U1201233 and 31370154), the Guangdong Province Key Laboratory of Computational Science and the Guangdong Province Computational Science Innovative Research Team.

References

Ashida H, Danchin A, Yokota. A. (2005). Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulfur metabolism? *Res Microbiol* **156**: 611–618.

Baker-Austin C, Dopson M. (2007). Life in acid: pH homeostasis in acidophiles. *Trends Microbiol* **15**: 165–171.

Baker BJ, Banfield JF. (2003). Microbial communities in acid mine drainage. *FEMS Microbiol Ecol* **44**: 139–152.

Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD *et al.* (2010). Enigmatic, ultrasmall, uncultivated Archaea. *Proc Natl Acad Sci USA* **107**: 8806–8811.

Baker BJ, Tyson GW, Webb RI, Flanagan J, Hugenholtz P, Allen EE *et al.* (2006). Lineages of acidophilic archaea revealed by community genomic analysis. *Science* **314**: 1933–1935.

Bertin PN, Heinrich-Salmeron A, Pelletier E, Goulhen-Chollet F, Arsène-Ploetze F, Gallien S *et al.* (2011). Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta-and proteo-genomics. *ISME J* **5**: 1735–1747.

Bonnefoy V, Holmes DS. (2012). Genomic insights into microbial iron oxidation and iron uptake strategies in extremely acidic environments. *Environ Microbiol* **14**: 1597–1611.

Brady A, Salzberg SL. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673–676.

Castelle CJ, Hug LA, Wrighton KC, Thomas BC, Williams KH, Wu D *et al.* (2013). Extraordinary phylogenetic diversity and metabolic versatility in aquifer sediment. *Nat Commun* **4**: 1–10.

Ciaramella M, Napoli A, Rossi M. (2005). Another extreme genome: how to live at pH 0. *Trends Microbiol* **13**: 49–51.

Chen LX, Li JT, Chen YT, Huang LN, Hua ZS, Hu M *et al.* (2013). Shifts in microbial community composition

and function in the acidification of a lead/zinc mine tailings. *Environ Microbiol* **15**: 2431–2444.

de la Bastide M, McCombie WR. (2007). Assembling genomic DNA sequences with PHRAP. *Curr Protoc Bioinformatics* **17**: 11.4.1–11.4.15.

Delcher AL, Salzberg SL, Phillippy AM. (2003). Using MUMmer to identify similar regions in large sequence sets. *Curr Protoc Bioinformatics* **2003**: 10.3.1–10.3.18.

Denef VJ, Mueller RS, Banfield JF. (2010). AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **4**: 599–610.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Dopson M, Johnson DB. (2012). Biodiversity, metabolism and applications of acidophilic sulfur-metabolizing microorganisms. *Environ Microbiol* **14**: 2620–2631.

Ertan H. (1992). Some properties of glutamate dehydrogenase, glutamine synthetase and glutamate synthase from *Corynebacterium callunae*. *Arch Microbiol* **158**: 35–41.

Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW *et al.* (2008). Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci USA* **105**: 3805–3810.

Friedrich CG, Rother D, Bardischewsky F, Quentmeier A, Fischer J. (2001). Oxidation of reduced inorganic sulfur compounds by bacteria: emergence of a common mechanism? *Appl Environ Microbiol* **67**: 2873–2882.

Futterer O, Angelow A, Liesegang H, Gottschalk G, Schelper C, Dock C *et al.* (2004). Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. *Proc Natl Acad Sci USA* **101**: 9091–9096.

Gifford SM, Sharma S, Booth M, Moran MA. (2013). Expression patterns reveal niche diversification in a marine microbial assemblage. *ISME J* **7**: 281–198.

Goltsman DSA, Denef VJ, Singer SW, VerBerkmoes NC, Lefsrud M, Mueller RS *et al.* (2009). Community genomic and proteomic analyses of chemoautotrophic iron-oxidizing '*Leptospirillum rubarum*'(group II) and '*Leptospirillum ferrodiazotrophum*'(group III) bacteria in acid mine drainage biofilms. *Appl Environ Microbiol* **75**: 4599–4615.

González-Toril E, Águilera A, Souza-Egipsy V, Pamo EL, España JS, Amils R. (2011). Geomicrobiology of La Zarza-Perrunal acid mine effluent (Iberian Pyritic Belt, Spain). *Appl Environ Microbiol* **77**: 2685–2694.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I *et al.* (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**: 644–652.

Gregersen LH, Bryant DA, Frigaard N-U. (2011). Mechanisms and evolution of oxidative sulfur metabolism in green sulfur bacteria. *Front Microbiol* **2**: 116.

Hallberg KB, Coupland K, Kimura S, Johnson DB. (2006). Macroscopic streamer growths in acidic, metal-rich mine waters in North Wales consist of novel and remarkably simple bacterial communities. *Appl Environ Microbiol* **72**: 2022–2030.

Hallberg KB, González-Toril E, Johnson DB. (2010). *Acidithiobacillus ferrivorans*, sp. nov.; facultatively anaerobic, psychrotolerant iron-, and sulfur-oxidizing acidophiles isolated from metal mine-impacted environments. *Extremophiles* **14**: 9–19.

Harris J. (2009). Soil microbial communities and restoration ecology: facilitators or followers? *Science* **325**: 573–574.

Hedrich S, Heinzel E, Seifert J, Schlömann M. (2009). Isolation of novel iron-oxidizing bacteria from an acid mine water treatment plant. *Adv Mater Res* **71–73**: 125–128.

Hess M, Sczyrba A, Egan R, Kim TW, Chokhawala H, Schroth G *et al.* (2011). Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**: 463–467.

Hensen D, Sperling D, Trüper HG, Brune DC, Dahl C. (2006). Thiosulphate oxidation in the phototrophic sulphur bacterium *Allochromatium vinosum*. *Mol Microbiol* **62**: 794–810.

Hewson I, Poretsky RS, Beinart RA, White AE, Shi T, Bench SR *et al.* (2009). In situ transcriptomic analysis of the globally important keystone N2-fixing taxon *Crocosphaera watsonii*. *ISME J* **3**: 618–631.

Huang LN, Zhou WH, Hallberg KB, Wan CY, Li J, Shu WS. (2011). Spatial and temporal analysis of the microbial community in the tailings of a Pb-Zn mine generating acidic drainage. *Appl Environ Microbiol* **77**: 5540–5544.

Huber JA, Welch DBM, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR *et al.* (2013). Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. *Microbiome* **1**: 22.

Jiao N, Herndl GJ, Hansell DA, Benner R, Kattner G, Wilhelm SW *et al.* (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nat Rev Microbiol* **8**: 593–599.

Johnson DB, Hallberg KB. (2003). The microbiology of acidic mine waters. *Res Microbiol* **154**: 466–473.

Johnson DB, Hallberg KB. (2008). Carbon, iron and sulfur metabolism in acidophilic microorganisms. *Adv Microb Physiol* **54**: 201–255.

Johnson DB, Hallberg KB, Hedrich S. (2014). Uncovering a microbial enigma: isolation and characterization of the streamer-generating, iron-oxidizing acidophilic bacterium, *'Ferrovum myxofaciens'*. *Appl Environ Microbiol* **80**: 672–680.

Johnson DB, Kanao T, Hedrich S. (2012). Redox transformations of iron at extremely low pH: fundamental and applied aspects. *Front Microbiol* **3**: 96.

Jones DS, Albrecht HL, Dawson KS, Schaperdoth I, Freeman KH, Pi Y *et al.* (2011). Community genomic analysis of an extremely acidophilic sulfur-oxidizing biofilm. *ISME J* **6**: 158–170.

Kantor RS, Wrighton KC, Handley KM, Sharon I, Hug LA, Castelle CJ *et al.* (2013). Small Genomes and Sparse Metabolisms of Sediment-Associated Bacteria from Four Candidate Phyla. *mBio* **4**: e00708–e00713.

Kimura S, Bryan CG, Hallberg KB, Johnson DB. (2011). Biodiversity and geochemistry of an extremely acidic, low temperature subterranean environment sustained by chemolithotrophy. *Environ Microbiol* **13**: 2092–2104.

Küsel K, Roth U, Drake HL. (2002). Microbial reduction of Fe(III) in the presence of oxygen under low pH conditions. *Environ Microbiol* **4**: 414–421.

Kuang JL, Huang LN, Chen LX, Hua ZS, Li SJ, Hu M *et al.* (2013). Contemporary environmental variation determines microbial diversity patterns in acid mine drainage. *ISME J* **7**: 1038–1050.

Lauro BD, Rossi M, Moracci M. (2006). Characterization of a β-glycosidase from the thermophilic bacterium *Alicyclobacillus acidocaldarius*. *Extremophiles* **10**: 301–310.

Levicán G, Ugalde JA, Ehrenfeld N, Maass A, Parada P. (2008). Comparative genomic analysis of carbon and nitrogen assimilation mechanisms in three indigenous bioleaching bacteria: predictions and validations. *BMC Genomics* **9**: 581.

Liljeqvist M, Valdes J, Holmes DS, Dopson M. (2011). Draft genome of the psychrotolerant acidophile *Acidithiobacillus ferrivorans* SS3. *J Bacteriol* **193**: 4304–4305.

Liljeqvist M, Rzhepishevska OI, Dopson M. (2013). Gene identification and substrate regulation provide insights into sulfur accumulation during bioleaching with the psychrotolerant acidophile *Acidithiobacillus ferrivorans*. *Appl Environ Microbiol* **79**: 951–957.

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

Mangold S, Valdes J, Holmes DS, Dopson M. (2011). Sulfur metabolism in the extreme acidophile *Acidithiobacillus caldus*. *Front Microbiol* **2**: 17.

Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, Fortney JL *et al.* (2012). Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. *ISME J* **6**: 1715–1727.

Méndez-García C, Mesa V, Sprenger RR, Richter M, Diez MS, Solano J *et al.* (2014). Microbial stratification in low pH oxic and suboxic macroscopic growths along an acid mine drainage. *ISME J* **8**: 1259–1274.

Musat N, Halm H, Winterholler B, Hoppe P, Peduzzi S, Hillion F *et al.* (2008). A single-cell view on the ecophysiology of anaerobic phototrophic bacteria. *Proc Natl Acad Sci USA* **105**: 17861–17866.

Ottesen EA, Marin R, Preston CM, Young CR, Ryan JP, Scholin CA *et al.* (2011). Metatranscriptomic analysis of autonomously collected and preserved marine bacterioplankton. *ISME J* **5**: 1881–1895.

Pace NR. (1997). A molecular view of microbial diversity and the biosphere. *Science* **276**: 734–740.

Pell J, Hintze A, Canino-Koning R, Howe A, Tiedje JM, Brown CT. (2012). Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc Natl Acad Sci USA* **109**: 13272–13277.

Peng Y, Leung HC, Yiu SM, Chin FY. (2012). IDBA-UD: a *de novo* assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**: 1420–1428.

Pruitt KD, Tatusova T, Maglott DR. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**: D61–D65.

Prosser JI, Bohannan BJ, Curtis TP, Ellis RJ, Firestone MK, Freckleton RP *et al.* (2007). The role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5**: 384–392.

Ram RJ, VerBerkmoes NC, Thelen MP, Tyson GW, Baker BJ, Blake RC *et al.* (2005). Community proteomics of a natural microbial biofilm. *Science* **308**: 1915–1920.

Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF *et al.* (2013). Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.

Rohwerder T, Sand W. (2003). The sulfane sulfur of persulfides is the actual substrate of the sulfur-oxidizing enzymes from *Acidithiobacillus* and *Acidiphilium* spp. *Microbiology* **149**: 1699–1710.

Rowe OF, Johnson DB. (2008). Comparison of ferric iron generation by different species of acidophilic bacteria immobilized in packed-bed reactors. *Syst Appl Microbiol* **31**: 68–77.

Schippers A, Sand W. (1999). Bacterial leaching of metal sulfides proceeds by two indirect mechanisms via thiosulfate or via polysulfides and sulfur. *Appl Environ Microbiol* **65**: 319–321.

Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. (2013). Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and phage during infant gut colonization. *Genome Res* **23**: 111–120.

Shi Y, Tyson GW, DeLong EF. (2009). Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**: 266–269.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Stewart FJ, Sharma AK, Bryant JA, Eppley JM, DeLong EF. (2011). Community transcriptomics reveals universal patterns of protein sequence conservation in natural microbial communities. *Genome Biol* **12**: R26.

Strous M, Kraft B, Bisdorf R, Tegetmeyer HE. (2012). The binning of metagenomic contigs for microbial physiology of mixed cultures. *Front Microbiol* **3**: 410.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P *et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res* **39**: D561–D568.

Tiedje JM. (1994). Microbial diversity: of value to whom? *ASM News* **60**: 524–525.

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

Valdés J, Ossandon F, Quatrini R, Dopson M, Holmes DS. (2011). Draft genome sequence of the extremely acidophilic biomining bacterium *Acidithiobacillus thiooxidans* ATCC 19377 provides insights into the evolution of the *Acidithiobacillus* genus. *J Bacteriol* **193**: 7003–7004.

Valdés J, Pedroso I, Quatrini R, Dodson RJ, Tettelin H, Blake R *et al.* (2008). *Acidithiobacillus ferrooxidans* metabolism: from genome sequence to industrial applications. *BMC Genomics* **9**: 597.

Wrighton KC, Thomas BC, Sharon I, Miller CS, Castelle CJ, VerBerkmoes NC *et al.* (2012). Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661–1665.

Zerbino DR, Birney E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* **18**: 821–829.

Zhu W, Lomsadze A, Borodovsky M. (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res* **38**: e132.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)