OPEN

The ISME Journal (2016), 1–13

www.nature.com/ismej

ORIGINAL ARTICLE

Predicting taxonomic and functional structure of microbial communities in acid mine drainage

Jialiang Kuang[1,2], Linan Huang[1], Zhili He[2], Linxing Chen[1], Zhengshuang Hua[1], Pu Jia[1], Shengjin Li[1], Jun Liu[1], Jintian Li[1], Jizhong Zhou[2,3,4] and Wensheng Shu[1]

[1]*State Key Laboratory of Biocontrol, Guangdong Key Laboratory of Plant Resources and Conservation of Guangdong Higher Education Institutes, College of Ecology and Evolution, Sun Yat-sen University, Guangzhou, People's Republic of China;* [2]*Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK, USA;* [3]*Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA and* [4]*State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, People's Republic of China*

Predicting the dynamics of community composition and functional attributes responding to environmental changes is an essential goal in community ecology but remains a major challenge, particularly in microbial ecology. Here, by targeting a model system with low species richness, we explore the spatial distribution of taxonomic and functional structure of 40 acid mine drainage (AMD) microbial communities across Southeast China profiled by 16S ribosomal RNA pyrosequencing and a comprehensive microarray (GeoChip). Similar environmentally dependent patterns of dominant microbial lineages and key functional genes were observed regardless of the large-scale geographical isolation. Functional and phylogenetic β-diversities were significantly correlated, whereas functional metabolic potentials were strongly influenced by environmental conditions and community taxonomic structure. Using advanced modeling approaches based on artificial neural networks, we successfully predicted the taxonomic and functional dynamics with significantly higher prediction accuracies of metabolic potentials (average Bray–Curtis similarity 87.8) as compared with relative microbial abundances (similarity 66.8), implying that natural AMD microbial assemblages may be better predicted at the functional genes level rather than at taxonomic level. Furthermore, relative metabolic potentials of genes involved in many key ecological functions (for example, nitrogen and phosphate utilization, metals resistance and stress response) were extrapolated to increase under more acidic and metal-rich conditions, indicating a critical strategy of stress adaptation in these extraordinary communities. Collectively, our findings indicate that natural selection rather than geographic distance has a more crucial role in shaping the taxonomic and functional patterns of AMD microbial community that readily predicted by modeling methods and suggest that the model-based approach is essential to better understand natural acidophilic microbial communities.

The ISME Journal advance online publication, 4 March 2016; doi:10.1038/ismej.2015.201

Introduction

Given the critical importance of species biogeography for biological conservation and climate change management, the development and application of statistic models for predicting the species distribution are an essential issue in community ecology (Elith and Leathwick, 2009). In the past two decades, the number of studies involved in species distribution models of plants and animals has increased markedly, providing ecological insights into the assessment of impacts and consequences of environmental changes on natural communities and ecosystems (Guisan and Thuiller, 2005; Guisan *et al.*, 2006; Austin, 2007; Pearman *et al.*, 2008).

Microorganisms are arguably the most diverse and abundant group of organisms on Earth (Fierer and Jackson, 2006), driving the bulk of biogeochemical cycles on the planet and influencing the functioning of virtually all ecosystems. During the last few years, a large number of phylogeny/taxonomy-based surveys have focused on the spatio-temporal dynamics and biogeographic patterns of microbial communities, revealing environmental variations (that is,

Correspondence: W Shu, State Key Laboratory of Biocontrol, Guangdong Key Laboratory of Plant Resources and Conservation of Guangdong Higher Education Institutes, College of Ecology and Evolution, Sun Yat-sen University, Guangzhou 510275, People's Republic of China.
E-mail: shuws@mail.sysu.edu.cn
Received 1 May 2015; revised 27 September 2015; accepted 2 October 2015

contemporary environmental conditions) (Lozupone and Knight, 2007; Lauber *et al.*, 2009) or spatial isolation (that is, historical events and disturbances) (Whitaker *et al.*, 2003; Martiny *et al.*, 2011) are the major factors shaping the large-scale ecological breadth of microbes. However, these studies are mainly limited to descriptive approaches rather than predictive model-based analyses (Gonzalez *et al.*, 2012). With the recent development of high-throughput molecular technologies and advanced bioinformatics tools, there have been increasing attempts to predict the biogeographic distributions of microbes across diverse ecosystems (King *et al.*, 2010; Larsen *et al.*, 2012; Bokulich *et al.*, 2013; Ladau *et al.*, 2013; Szabo *et al.*, 2013). These pioneering studies demonstrate that it is now possible to obtain more comprehensive understanding of microbial communities and their connections with climate change and biogeochemical cycling using vastly increased data sets.

Although the novel predictive strategies based on phylogenetic/taxonomic profiles have significantly advanced the study of microbial communities from the descriptive nature to the predictive science, the underlying mechanisms of how changes in these spatio-temporal variations of biogeographic pattern affect the processes of ecosystem functioning remain largely unknown, especially in a predictive scheme. As a broad range of functional variation may occur among closely related organisms, taxonomic distributions are assumed to be ambiguous in assessing the response of microbial communities to environmental changes (Green *et al.*, 2008) and may be of little value in predicting the functional dynamics in ecosystems. Thus, the functional traits (for example, gene content and metabolic potential), which determine the habitat-related attributes of a specific microbial species, have recently received a great deal of attention. Recent studies have highlighted the critical importance of trait-based approaches for studying microbial biogeography (Green *et al.*, 2008; Raes *et al.*, 2011; Barberán *et al.*, 2014). The investigation of functional traits distribution across spatial/temporal scales and along geochemical gradients will help elucidate how natural communities and their ecological functions respond to environmental changes (Green *et al.*, 2008; Bryant *et al.*, 2012; Fierer *et al.*, 2012) and subsequently identify the interaction of ecological processes affecting biogeographic patterns (Hanson *et al.*, 2012). Consequently, by combining the advanced modeling strategy and trait-based approaches, never has there been a greater opportunity for investigating the dynamics of functional community structure in space and time.

Community assembly is previously suggested to be deterministic in trait-based functional structure but historically contingent in taxonomic composition, indicating that environmental conditions would determine the types of ecological niches available for specific functional groups, whereas species compositions with similar physiological fitness are stochastically influenced by the history (Fukami *et al.*, 2005). Accordingly, the responses of functional traits specifically associated with the habitat-related attributes of microbial taxa may be more deterministic to environmental changes compared with those of taxonomic community composition. Thus, we hypothesized that natural microbial assemblages may be better predicted at the functional genes level rather than species. Here, we use the acid mine drainage (AMD) model system to test this hypothesis. These acidic, metal-rich drainages arise largely from the microbially mediated oxidative dissolution of sulfide minerals (for example, pyrite) and represent a major environmental problem worldwide (Baker and Banfield, 2003; Johnson and Hallberg, 2003). The microbial and geochemical simplicity of AMD systems makes them ideal targets for a quantitative, genomic-based test of our assumption. We applied a comprehensive functional gene array (GeoChip 4.0) (Tu *et al.,* 2014) and a recently developed modeling approach (Larsen *et al.*, 2012) to 40 environmental samples that were previously collected from diverse AMD sites across Southeast China with detailed microbial community composition and associated geochemical properties (Kuang *et al.*, 2013). Our results demonstrated that the patterns of taxonomic and functional community structure were environmentally dependent and readily predictable with significant higher prediction accuracies of metabolic potentials compared with relative microbial abundances. These findings provide ecological important insights into the adaptive strategies of how these microorganisms can survive and thrive in the extreme AMD environment.

Materials and methods

Sample description

A total of 40 AMD samples that are distinct with respect to environmental characteristics were previously collected across Southeast China. Sampling procedure, physicochemical analyses, DNA extraction, bar-coded 16S ribosomal RNA pyrosequencing and data processing were described previously (Kuang *et al.*, 2013) (also see a brief description in Supporting Information). The sequences reported in this paper have been deposited in the European Nucleotide Archive database (accession no. PRJEB9908). Total community genomic DNAs were profiled with the comprehensive microarray GeoChip 4.0 (Tu *et al.*, 2014) and subsequently analyzed with the corresponding data of geochemistry and microbial community composition (see below).

GeoChip analysis and data processing

The general pipeline of DNA labeling, GeoChip processing and data normalization was described

previously (He *et al*., 2007; Tu *et al*., 2014). The details of GeoChip analysis are available in Supporting Information. The GeoChip 4.0 covers major functional genes involved in biogeochemical processes and stress toleration and adaptation, and generates standard and comparable data sets appropriate for subsequent tests of ecological theories and biogeographic hypotheses (He *et al*., 2012). A suite of genes targeting a specific pathway of biogeochemical process are grouped into a functional gene category. The diversity of a given gene can be estimated by the detection of various 50-mer oligonucleotide probes, whereas its functional metabolic potential (that is, gene abundance) is reflected by the sum of signal intensities of these detected probes. In this study, we specifically focused on the commonly detected genes involved in the key biogeochemical and ecological processes in AMD ecosystems, including C, N, S cycling, P utilization, energy processes, stress responses, heavy metal and antibiotic resistance. To increase the confidence of GeoChip hybridization intensity data, only those probes detected in at least half of the total samples were retained for subsequent analyses. Totally, 114 genes that met these criteria were selected (Supplementary Table S1) and the GeoChip data set reported in this paper is publicly available at http://ieg.ou.edu/4download/.

Prediction model of microbial assemblages and functional metabolic potentials

A modeling approach based on the artificial neural networks (ANN) was applied to predict the microbial community composition and functional metabolic potentials in response to the environmental changes. This method was developed to capture and model the complex interactions between microbial taxa and their environment, and was demonstrated to be able to accurately predict natural microbial assemblages (Larsen *et al*., 2012). Here, we further apply this strategy to predict the functional metabolic potentials. Significant relationships of the interactions between nodes (that is, environmental parameters, microbial taxa or functional genes) were estimated using Bayesian network inference with Java Objects (BANJO v2.2.0) (Smith *et al*., 2006; Larsen *et al*., 2012) (for example, see Supplementary Figure S1). The relationships revealed by the consensus network generated from the output of BANJO highest-scoring networks could be expressed as a set of formulas such that the value of every node is a function of the value of its parent nodes (Larsen *et al*., 2012). Selected nodes were subsequently incorporated into the nonlinear equation modeling and these ANN-based functions were derived using Eureqa v 0.99.9 beta software (Schmidt and Lipson, 2009). The best-fitted equations based on the optimality criteria were then used for the prediction. In the formula search, data from 30 randomly selected samples were used for model training. After the generation and selection of the best-fitting equation, the data of the remaining 10 samples were imported to validate this equation. Statistical significance of the model was tested by a randomized permutation-based approach (reshuffled 10 000 times) as described previously (Larsen *et al*., 2015). In addition, two null models were performed to test whether the predicted model has better correlation with biological observation than these null models: (i) setting all taxa's predicted relative abundance/metabolic potentials equal to the average taxa abundance/metabolic potentials across all samples, (ii) setting all taxa abundances/metabolic potentials equal to the minimum observed values across all samples (Larsen *et al*., 2015). Details of the modeling method are available in Supporting Information.

Comparison of prediction accuracies between different biotic levels

To test our hypothesis that the dynamics of functional metabolic potentials in response to environmental changes are more predictable than those of microbial taxa, permutation-based Bray–Curtis similarities between predicted and observed values at different biotic levels were calculated to provide a statistic estimation of prediction accuracy (Larsen *et al*., 2012). The differences of these prediction accuracies were subsequently analyzed by *t*-test (pairwise *t*-test) and the statistical significance (*P*-value) was adjusted by the Bonferroni correction and the false discovery rate (Benjamini algorithm), respectively, to deal with the non-independent data sets (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). In this study, three microbial taxonomic/phylogenetic levels including phylum, order and operational taxonomic unit (defined at the 97% 16S ribosomal RNA similarity level) were chosen. Specifically, those microbial taxa with high relative abundance (that is, phyla $>1\%$, orders $>0.1\%$) or wide distribution (that is, operational taxonomic units observed in at least half of the total samples) were used for the analyses. For functional metabolic potentials, models were respectively fitted using original signal intensities and relative values that were normalized between 1 and 100 according to this formula:

$$\text{Func norm}_{i}^{j} = \left(1 - \frac{\text{Max}(\text{Func}^{j}) - \text{Func}_{i}^{j}}{\text{Max}(\text{Func}^{j}) - \text{Min}(\text{Func}^{j})}\right) \times 99 + 1$$

where Func norm_{i}^{j} is the normalized value for the metabolic potential of gene j at sample i, Func_{i}^{j} is the observed value for the metabolic potential of gene j at sample i, and Max and Min give the maximum and minimum values for the metabolic potential of gene j across all samples. Given that many microbial taxa and a large number of GeoChip probes were not widely detected across all samples, the difference of predictive power at different biotic levels may be affected by the

rarely observed taxa/probes. Thus, we further investigated the predictive power at various taxa/probes occurrences (that is, percentage of the total samples where a given taxa/probe was detected).

Statistical analyses

Various packages were used for the implementation of statistical analyses in R (R Core Team, 2014). Environmental variables were standardized to zero mean and unit variance using 'decostand' function in the vegan package (version 2.3-0, Oksanen *et al.*, 2015). Bray–Curtis distances were used to construct the dissimilarity matrices for microbial community composition and functional community structure, whereas Euclidean distances were calculated using standardized environmental variables and geographical locations (vegan 2.3-0). Permutational multivariate analysis of variance ('Adonis' function), analysis of similarity (ANOSIM, 'anosim' function) and multi-response permutation procedure analysis (MRPP, 'mrpp' function) were conducted to test the statistical significance of difference between three *a priori* classified groups of samples based on the clustering of geochemical data (see below) (vegan 2.3-0). Mantel tests were performed to reveal the correlation between two dissimilarity matrices (vegan 2.3-0). Principal component analysis was used to link the general pattern of functional community structure to distinct environmental conditions and convert a set of variables from possibly correlated to linearly uncorrelated (vegan 2.3-0). To quantitatively evaluate the relative influence of environmental properties, geographical distribution and microbial community composition to the diversities (Simpson index) and metabolic potentials of functional genes (De'ath, 2007; Kuang *et al.*, 2013), aggregated boosted tree analysis was applied using 'gbm' function with 5000 trees used for the boosting, 10-fold cross-validation and three-way interactions within the gbm package (version 2.1.1). Using individual variables assumed to be significantly correlated with each other would result in a large number of unplanned comparisons and so severely inflate Type I errors (John *et al.*, 2007). To reduce these impacts, orthogonal composite variables including PCs_{Env}, $PC_{Location}$ and PCs_{Taxa} derived from principal component analysis were computed respectively before the aggregated boosted tree analyses and only PCs with an eigenvalue >1 were retained. The correlations of functional metabolic potentials with the environmental properties and relative microbial abundances were fitted with generalized linear models (glm). Hierarchical cluster analysis was performed based on average linkage method (*hclust* argument) with Euclidean distance measure (*dist* argument) and visualized with heatmap.2 function (gplots 2.17.0, Warnes *et al.*, 2015).

Results

Patterns of taxonomic and functional community structure among distinct environmental conditions

Prior to adopting the prediction procedures, we first investigated whether the patterns of microbial community composition and functional community structure share mathematically describable relationships with environmental conditions (Larsen *et al.*, 2012). The environmental properties of our samples (Supplementary Table S2 and also see Supplementary Table S5 in Kuang *et al.*, 2013) represented the typical range of geochemical conditions in AMD environments (Johnson and Hallberg, 2003), although more extreme conditions (for example, extremely low solution pH between 0.3 and 1.2) have been reported in the Richmond Mine in California (Druschel *et al.*, 2004; Denef *et al.*, 2010; Mueller *et al.*, 2010). Hierarchical cluster analysis based on the geochemical data showed that the analyzed AMD samples were well separated into three groups qualitatively owing to the distinct environmental conditions (Figure 1a). Specifically, samples defined in *Group 1* were associated with more extremely acidic conditions ($pH = 2.2 \pm 0.07$, mean ± s.e.) and contained significantly higher concentrations of total organic carbon ($22 \pm 4.8\ mg\ l^{-1}$, *t*-test, comparing *Group 1* with *Group 2*/*3*, respectively, with both $P<0.05$), total phosphorus (P) ($6.5 \pm 2.5\ mg\ l^{-1}$, $P<0.05$) and heavy metals such as arsenic (As) ($24 \pm 9.9\ mg\ l^{-1}$, $P<0.05$) and cadmium (Cd) ($1.5 \pm 0.59\ mg\ l^{-1}$, $P<0.05$), whereas samples defined in *Group 2* were characterized by relatively moderate pH levels ($pH = 2.4 \pm 0.08$) and significant higher dissolved oxygen concentration ($6.2 \pm 1.3\ mg\ l^{-1}$, $P<0.05$). In comparison, the samples defined in *Group 3* were apparently featured by higher pH values ($pH = 3.0 \pm 0.09$), significantly lower electrical conductivity ($2605 \pm 330\ \mu S\ cm^{-1}$, $P<0.05$) and sulfate concentration ($2034 \pm 384\ mg\ l^{-1}$, $P<0.05$). Correspondently, the taxonomic microbial community composition (Figure 1b, Supplementary Table S3) and functional community structure (Figure 1c) were likely shaped by the distinct geochemical properties, and the differences of these abiotic and biotic structures among the sample groups were significantly different as revealed by three complementary non-parametric multivariate statistical tests (Table 1). Moreover, a significant correlation was found between functional and taxonomic β-diversities (Mantel test, $R = 0.32$, $P = 0.035$, Sorenson dissimilarities using profiles of 16S ribosomal RNA and functional genes). Additional analyses were also applied to assess whether geographical distance or local site characteristics (for example, climate and mineralogy) affected functional community structure as the AMD samples were collected across a wide range of distance (up to over 1600 km) and patchily located in different mining areas. Similar to the pattern of microbial community composition observed in previous study

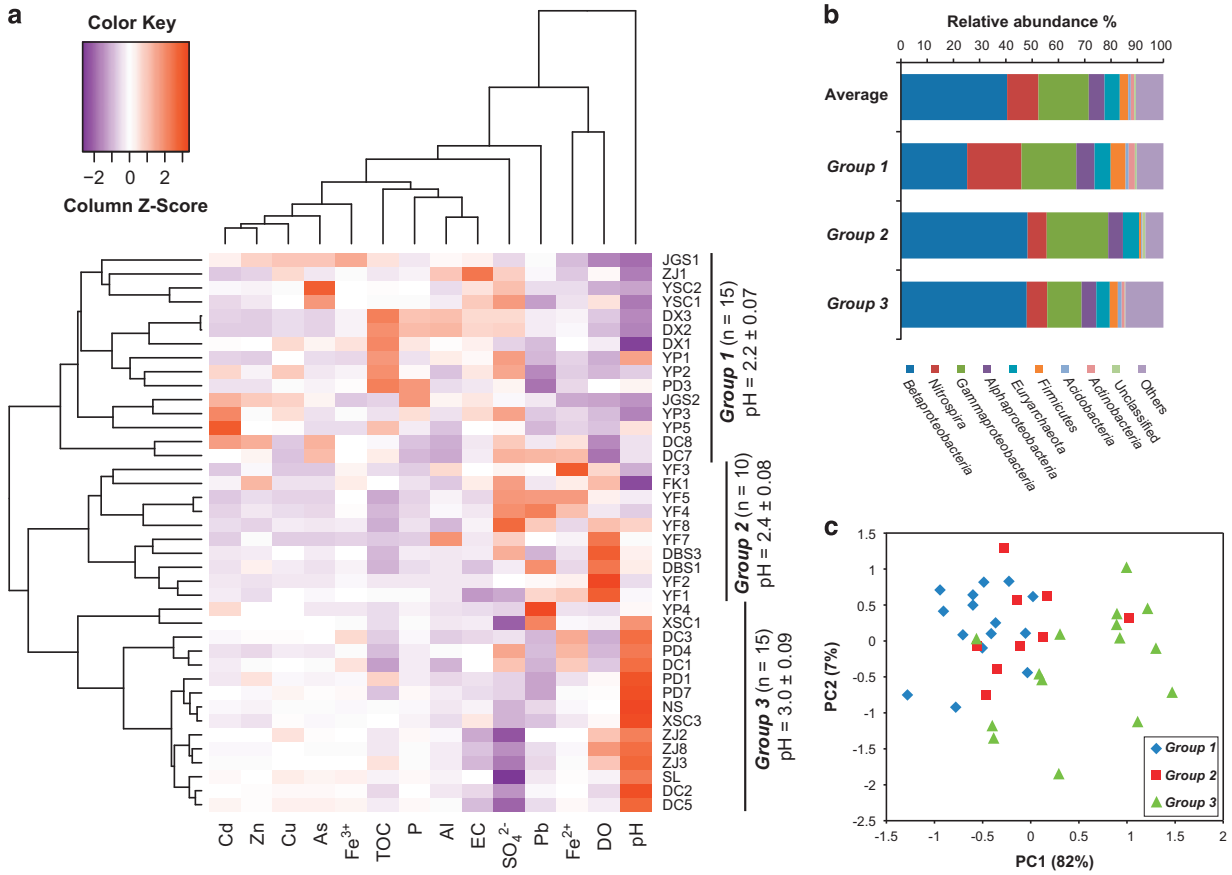

Figure 1 Hierarchical cluster analysis of geochemical data of the 40 acid mine drainage (AMD) samples (**a**) and the distribution patterns of microbial community composition (**b**) and functional community structure (**c**) among distinct environmental conditions. Geochemical data including pH, electrical conductivity (EC), dissolved oxygen (DO), total organic carbon (TOC), total phosphorus (P) and the concentrations of sulfate (SO_4^{2-})/ ferric (Fe^{3+})/ ferrous (Fe^{2+})/ aluminum (Al)/ arsenic (As)/ cadmium (Cd)/copper (Cu)/lead (Pd) and zinc (Zn) were standardized before hierarchical clustering (see details in Materials and methods). Relative abundances (%) of dominant lineages (phylum level) were shown in overall communities (average) and in different groups of AMD samples. Principal component analysis (PCA) was used to link the pattern of functional community structure to distinct environmental conditions based on the overall functional profiles (that is, selected probes of all key functional genes).

Table 1 Results of significant differences of the geochemical properties, the microbial community composition and the functional community structure between the sample groups

Dissimilarity method	*Geochemical properties (Euclidean distance)*		*Microbial community composition (Bray–Curtis)*[a]		*Functional community structure (Log, Bray–Curtis)*[b]	
	Statistic	P-*value*	*Statistic*	P-*value*	*Statistic*	P-*value*
Adonis[c]	0.029	0.001	0.098	0.018	0.207	0.001
ANOSIM[d]	0.244	0.001	0.041	0.015	0.238	0.001
MRPP[e]	0.082	0.001	0.032	0.031	0.106	0.001

[a]Bray–Curtis distance of microbial community composition is calculated based on the OTUs defined at the 97% similarity level. [b]The signal intensity of each probe was log-transformed before Bray–Curtis distance calculation. [c]Permutational multivariate analysis of variance. [d]Analysis of similarity. [e]Multi-response permutation procedure analysis.

(Kuang *et al*., 2013), there were no significant correlations between geographical distance and functional community dissimilarity (Mantel test, $P > 0.05$) and no significant differences of functional structure between most pairs of mining areas (Supplementary Table S4), implying a limited influence of spatial variation on the functional community structure. These results suggested that functional community structure as well as taxonomic community composition was better predicted

by environmental variation rather than spatial variations, consistent with the assumption of the modeling method.

Responses of metabolic potentials to the dynamics of environmental properties and microbial taxonomy

To further identify the driving forces for the patterns of diversity and metabolic potential of each functional gene, the relative influences of environmental properties (that is, PCs_{Env}), geographical distribution ($PC_{Location}$) and microbial community composition (PCs_{Taxa}) were interpreted by using the aggregated boosted tree models. The PCs with an eigenvalue greater than one collectively accounted for $>70\%$ of the variations and were chosen for the aggregated boosted tree analyses (Supplementary Table S5). Generally, the PCs_{Taxa} were identified as the major factors affecting the patterns of gene diversity, whereas the metabolic potentials of functional genes were influenced by the PC_{Env} (that is, E1) or PCs_{Taxa} (that is, T1, T3 and T4) (Figure 2). In contrast, spatial distribution was found to contribute less to both gene diversity and metabolic potential. We further addressed the responses of metabolic potentials to the changes of environmental properties and microbial taxonomy. The most dominant (top-50%) variables of each important PC were selected based on the PC loadings (Supplementary Table S5) and incorporated into the multiple linear regression analyses. Among the metabolic potentials of 114 genes analyzed, 23 and 20 were significantly correlated with environmental properties and relative microbial abundances, respectively (Supplementary Table S6). In most cases of E1, solution pH was indicated as a strong predictor of and negatively correlated with metabolic potentials as revealed by the best models. When considering the relative microbial abundances, *Euryarchaeota* and *Gammaproteobacteria* were commonly found to

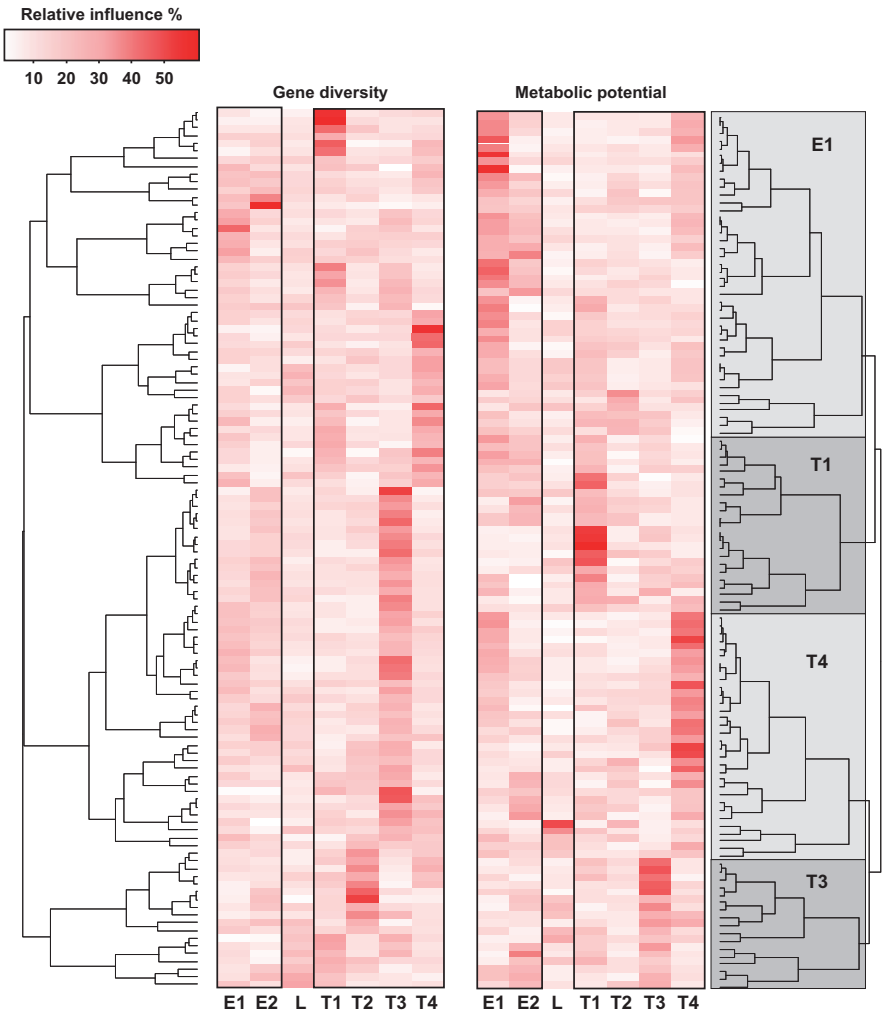

Figure 2 Relative influence (%) of environmental properties (PCs_{Env}, E1, E2), spatial distribution ($PC_{Location}$, L) and microbial community composition (PCs_{Taxa}, T1–T4) for gene diversity (Simpson index) and functional metabolic potential that evaluated by ABT models. Data profiles (that is, the relative influence of different PCs for each gene) were clustered with average clustering method based on the Pearson correlation. The 'boxes' qualitatively reveal the major factors (including E1, T1, T3 and T4) for the metabolic potentials of different functional genes according to the results of hierarchical clustering.

be significantly related to the functional metabolic potentials. These results suggested that there were clear patterns of gene diversity and metabolic potential of various key biogeochemical processes and stress responses, and solution pH and some dominant microbial lineages were the major factors determining the functional metabolic potentials in the AMD ecosystem.

Prediction of microbial community composition and functional metabolic potential

An ANN-based modeling approach was applied to predict the interactions among environmental properties, microbial community composition and functional metabolic potentials according to their relationships. Our results indicated that there were significant differences of prediction accuracy (that is, Bray–Curtis similarity between predicted and observed values) between different biotic levels, and the prediction accuracies of functional metabolic potential were significantly higher than those of relative microbial abundance (Figure 3a). Consistently, the cross-validation results showed that the functional metabolic potentials ($R^2_{(\text{orig. signal})} = 0.97$) were better predicted than those of relative microbial abundances ($R^2_{(\text{Phylum})} = 0.70$, $R^2_{(\text{Order})} = 0.62$, $R^2_{(\text{operational taxonomic unit})} = 0.52$) (Figure 3b, also see Supplementary Figures S2 and S3). A clear trend was found that the prediction accuracies and the coefficients (R^2) of relative microbial abundance decreased at lower microbial taxonomic levels. In order to assess whether this decrease in predictive power with increasing taxonomic resolution was largely a result of an increased number of rarely observed taxa, we further investigated the patterns of predictive power at various taxa occurrences. Our results suggested that there was no significant difference in predictive power across various occurrence levels (Figure 3a). The lowest prediction accuracy at operational taxonomic unit level implied that different microbial species might have similar responses to environmental changes and that our measured environmental parameters could not definitely simulate their natural dynamics. Higher predictive power was observed when modeling the functional metabolic potentials with relatively lower accuracies for models using normalized data, which might be due to the higher dependency of the overall data set especially for the data points of minimal and

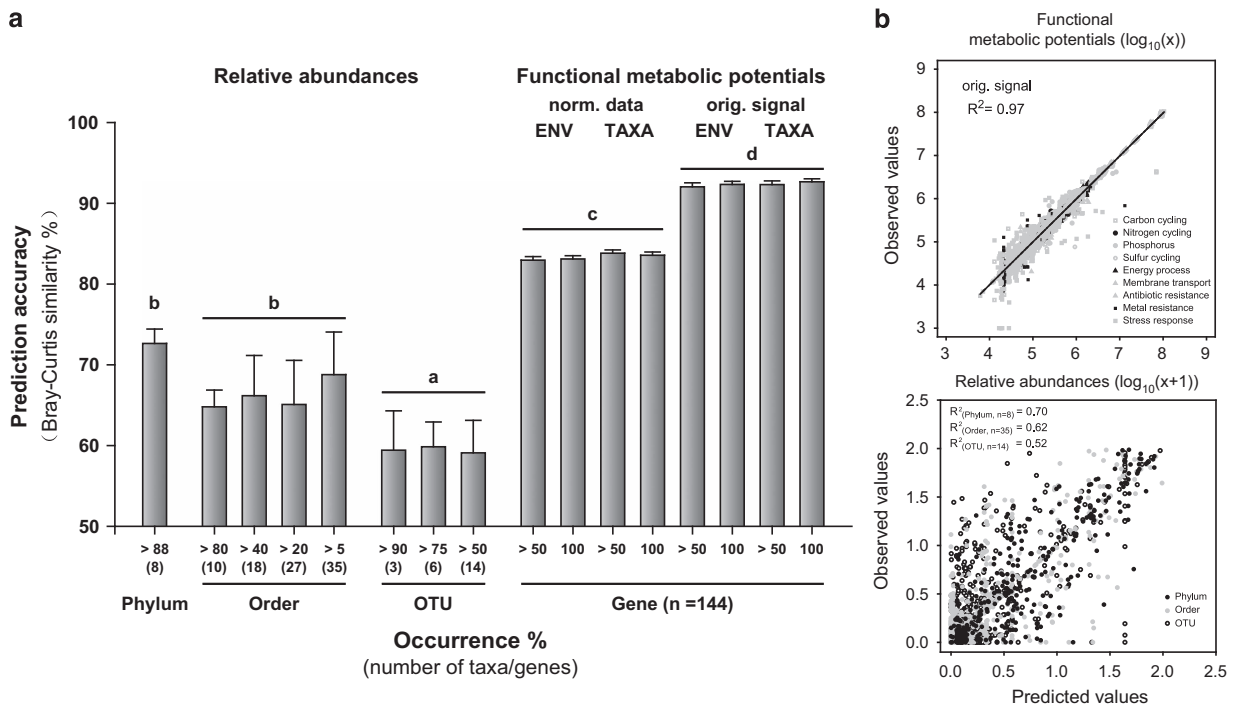

Figure 3 The comparison of prediction accuracies (**a**) and cross-validation results (**b**) between different biotic levels. Bray–Curtis similarity between predicted and observed values was used to represent the prediction accuracy according to the data sets for training and validation (that is, average values) (**a**). Values were mean ± s.e. and all these similarities were significant ($P<0.05$) that tested by a randomized permutation-based approach (reshuffled 10 000 times). The differences of prediction accuracies were subsequently analyzed by *t*-test (pairwise *t*-test) and the statistical significance (*P*-value) was adjusted by the Bonferroni correction and the false discovery rate (FDR), respectively, to deal with the non-independent data sets and consistent results were found by using these two *P*-value adjustment methods. The number of samples (*n*) for each level was listed. For metabolic potentials of functional genes, similarities were calculated based on the models using original signals (orig. signal) and normalized data (norm. data). In addition, these models were constructed with and without the information of microbial abundances of dominant phyla (TAXA versus ENV, see Supplementary Methods). For instance, predictive models of relative abundance at phylum level and predictive models of metabolic potential of functional genes with information of microbial abundances were available in Supplementary Tables S11 and S12. The scatter plots show the cross-validation of predicted and observed values for functional metabolic potentials and relative microbial abundances (**b**).

maximal values. Notably, modeling with relative abundances of microbial phyla could significantly predict the validation data points in 92.1% (105/114) of the functional genes as compared with 78.9% (90/114) of them without such taxonomic information, implying that the interaction of microbial species was necessary for predicting functional community structure that are not used to train the model. Subsequently, we also estimated the effect of rarely detected GeoChip probes on the predictive power, that is, a comparison using a data set with probes detected in at least 50% of samples versus all probes, and similarly limited effect was found for these rare probes (Figure 3a). In addition, two null models were performed to validate these ANN-based models and all of our predictive models at different biotic levels were better correlated with biological observation than the null models (Supplementary Tables S7–10), providing useful biological insight into the interactions. Collectively, these findings supported our hypothesis, suggesting that functional traits are more predictable by environmental conditions than microbial community composition.

Finally, we explored how microbial community composition and functional metabolic potential responded to the changes of pH, which was previously identified as the primary determinant of microbial diversity in extreme AMD systems (Kuang *et al.*, 2013). The most accurate models of relative microbial abundances (phylum level, Supplementary Table S11) and functional metabolic potentials (orig. signal with TAXA, occurrence $>50\%$, Supplementary Table S12), which generated significant Bray–Curtis similarity ($P<0.05$) of 72.6 ± 1.8 and 92.3 ± 0.6 for all training and validation data points (Figure 3a and Supplementary Figure S4), were used for the subsequent simulations. Functional metabolic potential was modeled based on the environmental parameters and relative microbial abundances as the consensus network revealed that these variables could be directly or indirectly predicted by solution pH (Supplementary Figure S1 and Supplementary Table S13). Thus, we could extrapolate the dynamics of microbial community composition and functional metabolic potential along a wider pH gradient even though the pH of our observed samples mainly ranged from 2.0 to 3.0. Finally, a pH range of 1.8–4.4 (at a 0.01-unit interval), which covers most of the pH values previously reported in AMD around the globe (Kuang *et al.*, 2013), was chosen for modeling. The predicted microbial community composition exhibited a consistent trend with the observed pattern especially for *Euryarchaeota*, *Nitrospira* and *Gammaproteobacteria* (Figure 4), corroborating the high accuracy of this modeling strategy. The pH-dependent distribution of these predominant lineages was possibly attributed to their remarkable environmental preferences and supposed to

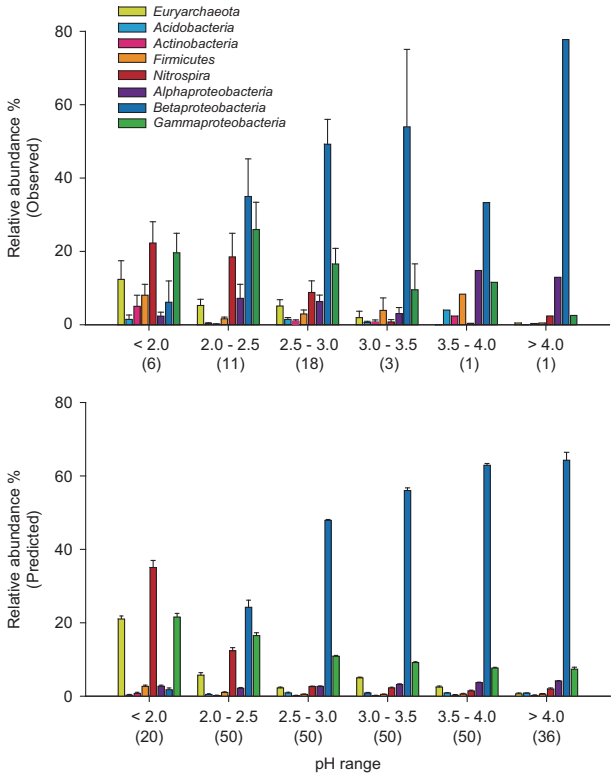

Figure 4 The comparison of predicted and observed relative abundances (%) of dominant lineages along the gradient of pH levels. Values were mean ± s.e. and the numbers in bracket indicate the number of samples that considered in each pH group.

contribute to the dynamics of functional metabolic potentials. Indeed, although the relative metabolic potentials of 52 genes kept consistent or fluctuant revealing in the predictive models with non-significant relationship ($P>0.05$, linear regression) between relative metabolic potentials and pH values (Supplementary Table S14), the relative metabolic potentials of the remaining 62 genes (significantly related to pH, $P<0.05$) showed clear patterns along the pH gradient (Figure 5 and Supplementary figure S5), and generally the observed values could be accurately predicted (Supplementary Figures S6–11). Specifically, the changes of relative metabolic potentials of some key genes related to biogeochemical processes (for example, nitrogen, phosphorus and sulfur cycling, Figures 5a and b and Supplementary Figure S5a) indicated the dynamics of resources utilization and energy transformation associated with the acidification process in AMD ecosystem, whereas the increase of relative metabolic potentials of genes involved in environmental adaptation such as heavy metal resistance (Figure 5c) and stress response (Supplementary Figure S5b) in more extreme conditions implied the adaptive strategies for these extremophilic communities.

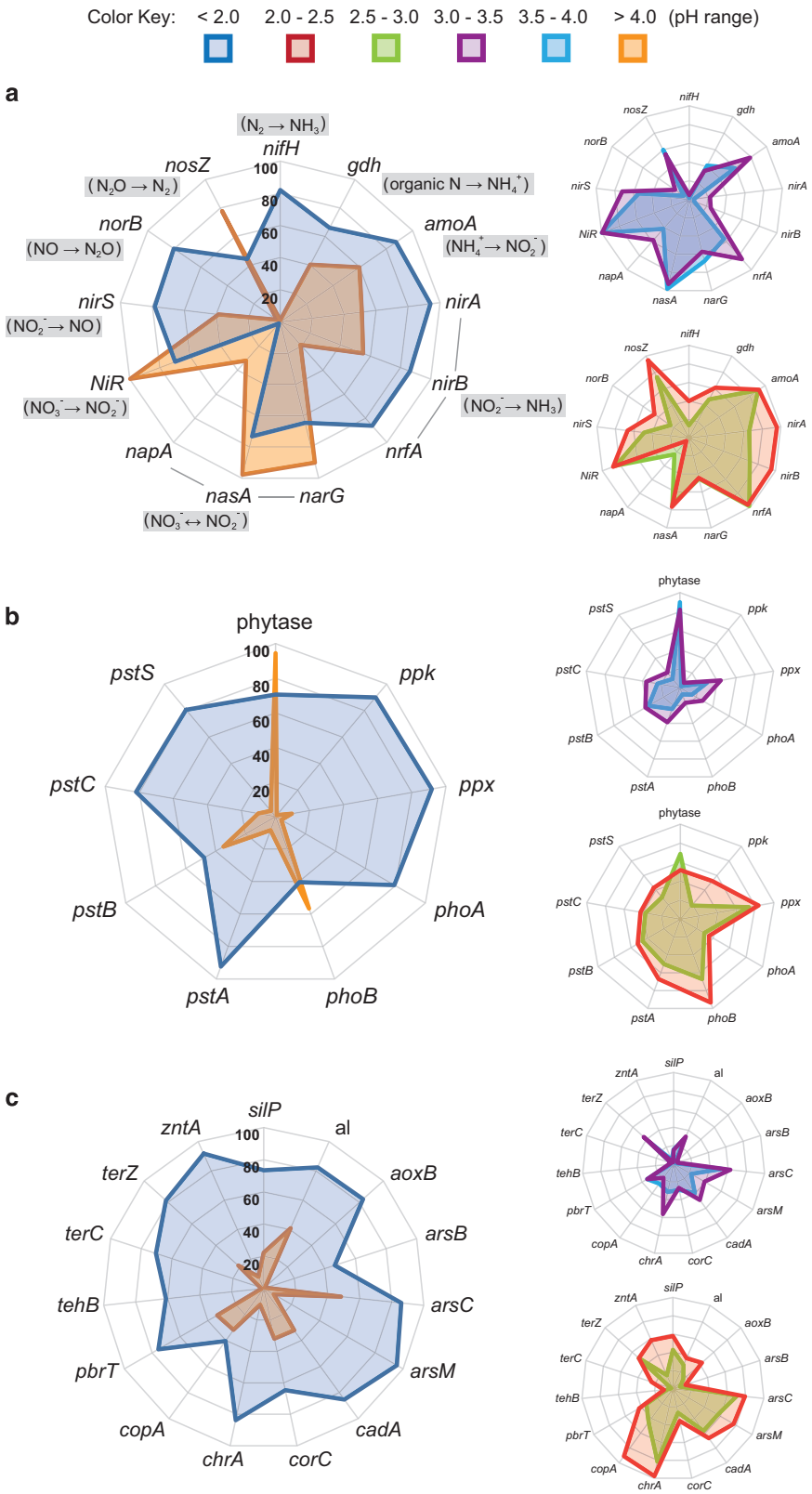

Figure 5 The predictive changes of relative metabolic potential of functional genes in nitrogen cycling (**a**), phosphorus cycling (**b**) and metal resistance (**c**) along the gradient of pH levels. The metabolic potentials were first modeled by original signal intensities with relative abundance information of microbial phyla and then normalized to relative values.

Discussion

Similar to the large-scale taxonomic composition patterns of AMD communites (Kuang *et al*., 2013), the spatial variation of functional community structure resolved by GeoChip was remarkably environment-dependent. Likewise, a previous study has demonstrated significant correlations between proteogenomics and geochemical and physical attributes in shaping communities of AMD biofilm (Mueller *et al*., 2010). These findings highlight the importance of natural selection in this extreme environment. Such severe environmental filtering may lead to a smaller available pool of species/genes that can persist under the harsh conditions, making their structure more niche-assembled (Chase, 2007). These patterns of taxonomic and functional biogeography shaped by the measurable environmental variables rather than geographic distance are highly consistent with the assumption of the modeling method (Larsen *et al*., 2012), enabling a successful application of the predictive strategy in this study.

According to the ANN-based predictive framework, we demonstrated that functional traits were more predictable by environmental variations and provided more useful explanation than taxonomic diversity based on phylogenetic markers in assessing the relationship between microbial communities and ecological processes. Although several recent metagenomic studies have revealed a significant correlation between phylogenetic diversity and functional diversity (Bryant *et al*., 2012; Fierer *et al*., 2012), specific functional traits and microbial species may not always have a the definite relationship, as functional interchange may occur across different taxa (Green *et al*., 2008), resulting in the conspicuous decoupling of ecological attributes from phylogeny (Raes *et al*., 2011; Barberán *et al*., 2014). In supporting this, a recent study has documented that specific functions could be widely detected across a variety of taxa or phylogenetic groups (Burke *et al*., 2011). Importantly, previous research has revealed that lateral gene transfer is prevailing mechanisms for AMD microbes to rapidly acquire and possess new genes involved in survival and habitat-specific functions (for example, heavy metals resistance) (Baker and Banfield, 2003; Tyson *et al*., 2004). Indeed, it was recently suggested that functional traits are valuable ecological markers to understand bacterial community assembly (Barberán *et al*., 2012) and to explain shifts in microbial community composition across environmental gradients (Edwards *et al*., 2013). As such, it is reasonable to obtain more accurate prediction of the metabolic potentials of key functional genes in response to environmental change, as these specific functional capabilities may directly impact how microbial communities interact with their environments.

Application of the predictive models allowed an accurate estimation of the dynamics of taxonomic and functional community structure along a pH range typically reported for AMD environments. As nitrogen resources are very limited in natural AMD systems (Baker and Banfield, 2003), their bioavailability and biogeochemical processes are vital to the acidophilic communities and essential in understanding of how these extraordinary assemblages respond and adapt to the harsh conditions. Diverse genes involved in nitrogen cycling were detected and predicted to show clear patterns of relative metabolic potentials (Figure 5a). In a recent transcriptional analysis of several AMD communities (Chen *et al*., 2015), nitrogen-fixation transcripts such as *nifH* were commonly found and associated with *Leptospirillum ferrooxidans*, *Leptospirillum ferrodiazotrophum*, *Acidithiobacillus ferrivorans* and *Acidithiobacillus sp.* GGI-221. In our predictive models, the relative metabolic potential of *nifH* exhibited a notable increase with the decrease of solution pH, which was possibly attributed to the dominance of *Leptospirillum* spp. and *Acidithiobacillus* spp. under more acidic conditions (Figures 4 and 5a). In addition, an increase of relative metabolic potential of glutamate dehydrogenase (*gdh*) mostly derived from *Thermoplasma* was predicted, indicating an alternative strategy of ammonium acquisition from organic N conducted by this dominant population in low pH conditions (Ruepp *et al*., 2000). Similar patterns of increased predictive potential activities were also found for genes encoding the enzymes for nitrite utilization (for example, *nirA*, *nirB* and *nrfA*). This accumulation of ammonia/ammonium might indicate a high requirement of nitrogen resources for microbial protein synthesis, which further supported by the higher relative metabolic potential of glutamine synthetase (*glnA*) that associated with incorporation of ammonium into glutamine (Leigh and Dodsworth, 2007) (Supplementary Figure S5b). These findings indicated the nitrogen-limited adaptation and the prosperity of these extremely acidophilic populations. Phosphate represents another key nutrient limited in the extreme AMD environment. With the increase of acidity, the high amounts of Fe^{3+} and Al^{3+} ions might favor phosphate precipitation (Moreno-Paz *et al*., 2010), resulting in further phosphate starvation. As predicted in our models, multiple strategies of phosphate uptake and utilization were used by enhancing the relative metabolic potentials of genes involved in polyphosphate metabolism (*ppk* and *ppx*) (Vera *et al*., 2003), phosphate regulon (*Pho*) (Lamarche *et al*., 2008) and specific phosphate ABC transporters (*pstSCAB*) (Parro *et al*., 2007) (Figure 5b), reflecting a positive response to the phosphate deficiency in the AMD systems.

Various protective mechanisms were identified to compensate for the deleterious effects of the extreme acidity. Diverse genes encoding proteins of heavy metal resistance and cation efflux systems were widely detected, and their relative metabolic potentials were predicted to be remarkably higher in lower

pH conditions (Figure 5c). This was possibly attributed to and stimulated by the increased concentrations of dissolved heavy metals. Likewise, the relative metabolic potentials of genes involved in the defense against oxidative and osmotic stress (for example, *oxyR*, *proV* and ABC transporters) were predicted to be highly increased as well (Supplementary Figures S5b and c). The membrane-binding ABC transporters are identified to function as pumps to exclude toxins and drugs from the cell (Higgins, 2001), and these transport systems such as potassium transporters (*kdpBAC*) are suggested to be an effective strategy to maintain pH homeostasis and cellular osmotic pressure (Baker-Austin and Dopson, 2007; Parro *et al*., 2007; Moreno-Paz *et al*., 2010). Collectively, these stress-resistant mechanisms may provide the populated microbes important strategy for surviving and thriving in the extreme environment.

Our predictive models also revealed some clues about microbial interaction in the AMD communities. It was suggested that more extreme conditions are less conducive to microbial growth, making survival capacities more important than the abilities for enhancing microbial competition (Fierer *et al*., 2012). However, the indigenous AMD populations are assumed to be well adapted to the extremely acidic conditions, whereas the decrease of energy sources such as pyrite and ferrous iron in lower pH environments might largely increase the importance of competition between sulfur/iron oxidizers. This assumption is partly supported by the higher relative metabolic potentials of genes involved in antibiotic resistance in our predictive models (Supplementary Figure S5d), as elevated microbial competition would select for increased antibiotic resistance (Fierer *et al*., 2012).

In summary, our analyses of the dynamics of taxonomic and functional community structure in response to the environmental changes by modeling strategy represents a crucial step toward a predictive model-based understanding of the distribution mechanisms of acidophilic microorganisms in the extreme AMD system. Our results showed that the environmentally dependent patterns of taxonomy and traits (functional genes) are readily predictable, whereas the notable enhancement of relative metabolic potentials of a suite of key functional genes under more acidic and metal-rich conditions may reflect an important adaptation strategy of these extraordinary assemblages. More importantly, we demonstrated that natural microbial communities in the AMD model system are better predicted at the functional genes level rather than species, at least by the set of functional genes considered in the current study. It should be pointed out that although the microbial taxonomic composition was resolved by pyrosequencing of universal 16S ribosomal RNA gene, the functional structure of AMD assemblages was profiled by GeoChip, which is a high-throughput microarray-based genomic technology designed for detecting 'known' genes specifically involved in biogeochemical processes and stress toleration and adaptation (Zhou *et al*., 2015). Alternatively, metagenomic sequencing represents another way to study the microbial community and its traits by simultaneously generating information on functional and taxonomic data sets. Such approaches could be adopted to verify our findings for its universality in diverse habitats.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Ping Zhang, Tong Yuan and Caiyun Yang for their experimental assistance with the GeoChip analysis. This work was supported by the National Natural Science Foundation of China (No. U1201233 and U1501232), the Major Science and Technology Project of Ministry of Agriculture of the People's Republic of China (No. 2009ZX08009-002B), the Guangdong Province Key Laboratory of Computational Science and the Guangdong Province Computational Science Innovative Research Team.

References

Austin M. (2007). Species distribution models and ecological theory: a critical assessment and some possible new approaches. *Ecol model* **200**: 1–19.

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

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

Barberán A, Fernández-Guerra A, Bohannan BJ, Casamayor EO. (2012). Exploration of community traits as ecological markers in microbial metagenomes. *Mol Ecol* **21**: 1909–1917.

Barberán A, Ramirez KS, Leff JW, Bradford MA, Wall DH, Fierer N. (2014). Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol Lett* **17**: 794–802.

Benjamini Y, Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B Stat Methodol* **57**: 289–300.

Benjamini Y, Yekutieli D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann Statist* **29**: 1165–1188.

Bokulich NA, Thorngate JH, Richardson PM, Mills DA. (2013). Microbial biogeography of wine grapes is conditioned by cultivar, vintage, and climate. *Proc Natl Acad Sci USA* **111**: E139–E148.

Bryant JA, Stewart FJ, Eppley JM, DeLong EF. (2012). Microbial community phylogenetic and trait diversity declines with depth in a marine oxygen minimum zone. *Ecology* **93**: 1659–1673.

Burke C, Steinberg P, Rusch D, Kjelleberg S, Thomas T. (2011). Bacterial community assembly based on

functional genes rather than species. *Proc Natl Acad Sci USA* **108**: 14288–14293.
Chase JM. (2007). Drought mediates the importance of stochastic community assembly. *Proc Natl Acad Sci USA* **104**: 17430–17434.
Chen LX, Hu M, Huang LN, Hua ZS, Kuang JL, Li SJ *et al.* (2015). Comparative metagenomic and metatranscriptomic analyses of microbial communities in acid mine drainage. *ISME J* **9**(7): 1579–1592.
De'ath G. (2007). Boosted trees for ecological modeling and prediction. *Ecology* **88**: 243–251.
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.
Druschel GK, Baker BJ, Gihring TM, Banfield JF. (2004). Acid mine drainage biogeochemistry at Iron Mountain, California. *Geochem Trans* **5**: 13–32.
Edwards KF, Litchman E, Klausmeier CA. (2013). Functional traits explain phytoplankton responses to environmental gradients across lakes of the United States. *Ecology* **94**: 1626–1635.
Elith J, Leathwick JR. (2009). Species distribution models: ecological explanation and prediction across space and time. *Annu Rev Ecol Evol Syst* **40**: 677–697.
Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.
Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL *et al.* (2012). Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc Natl Acad Sci USA* **109**: 21390–21395.
Fukami T, Bezemer TM, Mortimer SR, van der Putten WH. (2005). Species divergence and trait convergence in experimental plant community assembly. *Ecol Lett* **8**: 1283–1290.
Gonzalez A, King A, Robeson MS 2nd, Song S, Shade A, Metcalf JL *et al.* (2012). Characterizing microbial communities through space and time. *Curr Opin Biotechnol* **23**: 431–436.
Green JL, Bohannan BJM, Whitaker RJ. (2008). Microbial biogeography: from taxonomy to traits. *Science* **320**: 1039–1043.
Guisan A, Lehmann A, Ferrier S, Austin M, Overton J, Aspinall R *et al.* (2006). Making better biogeographical predictions of species' distributions. *J Appl Ecol* **43**: 386–392.
Guisan A, Thuiller W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecol Lett* **8**: 993–1009.
Hanson CA, Fuhrman JA, Horner-Devine MC, Martiny JB. (2012). Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol* **10**: 497–506.
He Z, Deng Y, Zhou J. (2012). Development of functional gene microarrays for microbial community analysis. *Curr Opin Biotechnol* **23**: 49–55.
He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC *et al.* (2007). GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. *ISME J* **1**: 67–77.
Higgins CF. (2001). ABC transporters: physiology, structure and mechanism—an overview. *Res Microbiol* **152**: 205–210.
John R, Dalling JW, Harms KE, Yavitt JB, Stallard RF, Mirabello M *et al.* (2007). Soil nutrients influence spatial distributions of tropical tree species. *Proc Natl Acad Sci USA* **104**: 864–869.
Johnson DB, Hallberg KB. (2003). The microbiology of acidic mine waters. *Res Microbiol* **154**: 466–473.
King AJ, Freeman KR, McCormick KF, Lynch RC, Lozupone C, Knight R *et al.* (2010). Biogeography and habitat modelling of high-alpine bacteria. *Nat Commun* **1**: 53.
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.
Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J *et al.* (2013). Global marine bacterial diversity peaks at high latitudes in winter. *ISME J* **7**: 1669–1677.
Lamarche MG, Wanner BL, Crepin S, Harel J. (2008). The phosphate regulon and bacterial virulence: a regulatory network connecting phosphate homeostasis and pathogenesis. *FEMS Microbiol Rev* **32**: 461–473.
Larsen PE, Dai Y, Collart FR. (2015). Predicting bacterial community assemblages using an artificial neural network approach. *Methods Mol Bio* **1260**: 33–43.
Larsen PE, Field D, Gilbert JA. (2012). Predicting bacterial community assemblages using an artificial neural network approach. *Nat Methods* **9**: 621–625.
Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75**: 5111–5120.
Leigh JA, Dodsworth JA. (2007). Nitrogen regulation in bacteria and archaea. *Annu Rev Microbiol* **61**: 349–377.
Lozupone C, Knight R. (2007). Global patterns in bacterial diversity. *Proc Natl Acad Sci USA* **104**: 11436–11440.
Martiny JB, Eisen JA, Penn K, Allison SD, Horner-Devine MC. (2011). Drivers of bacterial β-diversity depend on spatial scale. *Proc Natl Acad Sci USA* **108**: 7850–7854.
Moreno-Paz M, Gómez MJ, Arcas A, Parro V. (2010). Environmental transcriptome analysis reveals physiological differences between biofilm and planktonic modes of life of the iron oxidizing bacteria Leptospirillum spp. in their natural microbial community. *BMC Genomics* **11**: 404.
Mueller RS, Denef VJ, Kalnejais LH, Suttle KB, Thomas BC, Wilmes P *et al.* (2010). Ecological distribution and population physiology defined by proteomics in a natural microbial community. *Mol Syst Biol* **6**: 374.
Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB *et al.* (2015). Vegan: community ecology packageR package version 2.3-0. Available at: https://github.com/vegandevs/vegan.
Parro V, Moreno-Paz M, González-Toril E. (2007). Analysis of environmental transcriptomes by DNA microarrays. *Environ Microbiol* **9**: 453–464.
Pearman PB, Randin CF, Broennimann O, Vittoz P, van der Knaap WO, Engler R *et al.* (2008). Prediction of plant species distributions across six millennia. *Ecol Lett* **11**: 357–369.
Core Team R. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing: Vienna, Austria. Available at: http://www.R-project.org/.
Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. (2011). Toward molecular trait-based ecology through

integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* **7**: 473.
Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW *et al.* (2000). The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. *Nature* **407**: 508–513.
Schmidt M, Lipson H. (2009). Distilling free-form natural laws from experimental data. *Science* **324**: 81–85.
Smith VA, Yu J, Smulders TV, Hartemink AJ, Jarvis ED. (2006). Computational inference of neural information flow networks. *PLoS Comput Biol* **2**: e161.
Szabo G, Preheim SP, Kauffman KM, David LA, Shapiro J, Alm EJ *et al.* (2013). Reproducibility of *Vibrionaceae* population structure in coastal bacterioplankton. *ISME J* **7**: 509–519.
Tu Q, Yu H, He Z, Deng Y, Wu L, Van Nostrand JD *et al.* (2014). GeoChip 4: a functional gene arrays-based high throughput environmental technology for microbial community analysis. *Mol Ecol Resour* **14**: 914–928.
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.
Vera M, Guiliani N, Jerez CA. (2003). Proteomic and genomic analysis of the phosphate starvation response of *Acidithiobacillus ferrooxidans*. *Hydrometallurgy* **71**: 125–132.
Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T *et al.* (2015). Various R programming tools for plotting data. R package version 2.17.0.
Whitaker RJ, Grogan DW, Taylor JW. (2003). Geographic barriers isolate endemic populations of hyperthermophilic archaea. *Science* **301**: 976–978.
Zhou J, He Z, Yang Y, Deng Y, Tringe SG, Alvarez-Cohen L. (2015). High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *mBio 6* **1**: e02288–14.

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