

# UC San Diego

## UC San Diego Previously Published Works

### Title

The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales

### Permalink

<https://escholarship.org/uc/item/1jd260dn>

### Journal

Microbiome, 6(1)

### ISSN

2049-2618

### Authors

Li, Hong-Yi

Wang, Hang

Wang, Hai-Tao

et al.

### Publication Date

2018-12-01

### DOI

10.1186/s40168-018-0561-x

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

RESEARCH

Open Access



# The chemodiversity of paddy soil dissolved organic matter correlates with microbial community at continental scales

Hong-Yi Li<sup>1†</sup>, Hang Wang<sup>2†</sup>, Hai-Tiao Wang<sup>3,4</sup>, Pei-Yong Xin<sup>5</sup>, Xin-Hua Xu<sup>1</sup>, Yun Ma<sup>6</sup>, Wei-Ping Liu<sup>1</sup>, Chang-Yun Teng<sup>1,7</sup>, Cheng-Liang Jiang<sup>1,7</sup>, Li-Ping Lou<sup>1</sup>, Wyatt Arnold<sup>3,4</sup>, Lauren Cralle<sup>3,4</sup>, Yong-Guan Zhu<sup>8</sup>, Jin-Fang Chu<sup>5</sup>, Jack A Gilbert<sup>3,4\*</sup> and Zhi-Jian Zhang<sup>1,7,9\*</sup>

## Abstract

**Background:** Paddy soil dissolved organic matter (DOM) represents a major hotspot for soil biogeochemistry, yet we know little about its chemodiversity let alone the microbial community that shapes it. Here, we leveraged ultrahigh-resolution mass spectrometry, amplicon, and metagenomic sequencing to characterize the molecular distribution of DOM and the taxonomic and functional microbial diversity in paddy soils across China. We hypothesized that variances in microbial community significantly associate with changes in soil DOM molecular composition.

**Results:** We report that both microbial and DOM profiles revealed geographic patterns that were associated with variation in mean monthly precipitation, mean annual temperature, and pH. DOM molecular diversity was significantly correlated with microbial taxonomic diversity. An increase in DOM molecules categorized as peptides, carbohydrates, and unsaturated aliphatics, and a decrease in those belonging to polyphenolics and polycyclic aromatics, significantly correlated with proportional changes in some of the microbial taxa, such as *Syntrophobacterales*, *Thermoleophilia*, *Geobacter*, *Spirochaeta*, *Gaiella*, and *Defluviococcus*. DOM composition was also associated with the relative abundances of the microbial metabolic pathways, such as anaerobic carbon fixation, glycolysis, lignolysis, fermentation, and methanogenesis.

**Conclusions:** Our study demonstrates the continental-scale distribution of DOM is significantly correlated with the taxonomic profile and metabolic potential of the rice paddy microbiome. Abiotic factors that have a distinct effect on community structure can also influence the chemodiversity of DOM and vice versa. Deciphering these associations and the underlying mechanisms can precipitate understanding of the complex ecology of paddy soils, as well as help assess the effects of human activities on biogeochemistry and greenhouse gas emissions in paddy soils.

**Keywords:** Dissolved organic matter, Paddy soil, Chemodiversity, Microbial diversity, FT-ICR-MS

## Background

Paddy fields, 90% of which are in Asia, feed more than half of the world's population [1]. The continuous flooding in banded fields of cultivated rice (*Oryza sativa*) utilizes 24–30% of the world's developed freshwater resources and represents one of the major sources of

inland aquatic dissolved organic matter (DOM) [1, 2]. High concentrations and fluxes of DOM from plant debris during flooding seasons trigger microbial activity, while anaerobic conditions stabilize DOM against microbial decay via interactions with clay minerals and iron oxides [1, 3]. DOM plays a central role in biogeochemical processes in both flooded and unflooded paddy soils, as well as an active role in the global carbon cycle [1, 4].

Recently, the evidence-based soil continuum model questioned the secondary synthesis of “humic substances,” or the “humification,” and interpreted organic debris as a unique source of soil organic matter (SOM) and DOM [5]. This theory emphasized the inherent

\* Correspondence: gilbertjack@uchicago.edu; gilbertjack@gmail.com; zhangzhijian@zju.edu.cn

<sup>†</sup>Hong-Yi Li and Hang Wang contributed equally to this work.

<sup>3</sup>The Microbiome Center, Biosciences Division, Argonne National Laboratory, Lemont, IL 60439, USA

<sup>1</sup>College of Environment and Natural Resource Sciences, Zhejiang University, 866 Yuhangtang Ave, Hangzhou 310058, China

Full list of author information is available at the end of the article



association between soil microbial metabolism and DOM heterogeneity. Microbes play an important role in carbon and nitrogen cycling as well as methane production and consumption in soils [1, 6, 7], which influences carbon balance, greenhouse gas production, crop productivity, and water eutrophication [8]. Therefore, a growing body of research has focused on the biogeography of microbial communities [9–11]. However, despite efforts to characterize the drivers of DOM concentration dynamics [8, 12–15], no attempts have been made to combine these data so as to understand the associated properties of each and the environmental factors that drive them. Ultrahigh-resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR-MS) enables detailed characterization of DOM molecular distribution [13, 16, 17]. This approach has been applied to marine [18, 19] and inland water [4, 20, 21], and a handful of comparative experiments have examined soil DOM at the molecular level [3, 14, 15, 22]; however, the microbial taxonomic and metabolic structures that influence the molecular distribution of soil DOM remain unknown.

To our knowledge, no comprehensive study has yet been performed to elucidate the natural relationship between microbial metabolisms and DOM molecular distribution in paddy soils on continental scales. We apply FT-ICR-MS plus amplicon and metagenomic sequencing to characterize the association between microbial community structure and function with DOM molecular composition in flooded paddy soils. We hypothesized that taxonomic and functional composition of soil microbial communities is significantly associated with DOM molecular composition in paddy soils and, moreover, that geographic and edaphic factors significantly affect this interdependence.

## Results

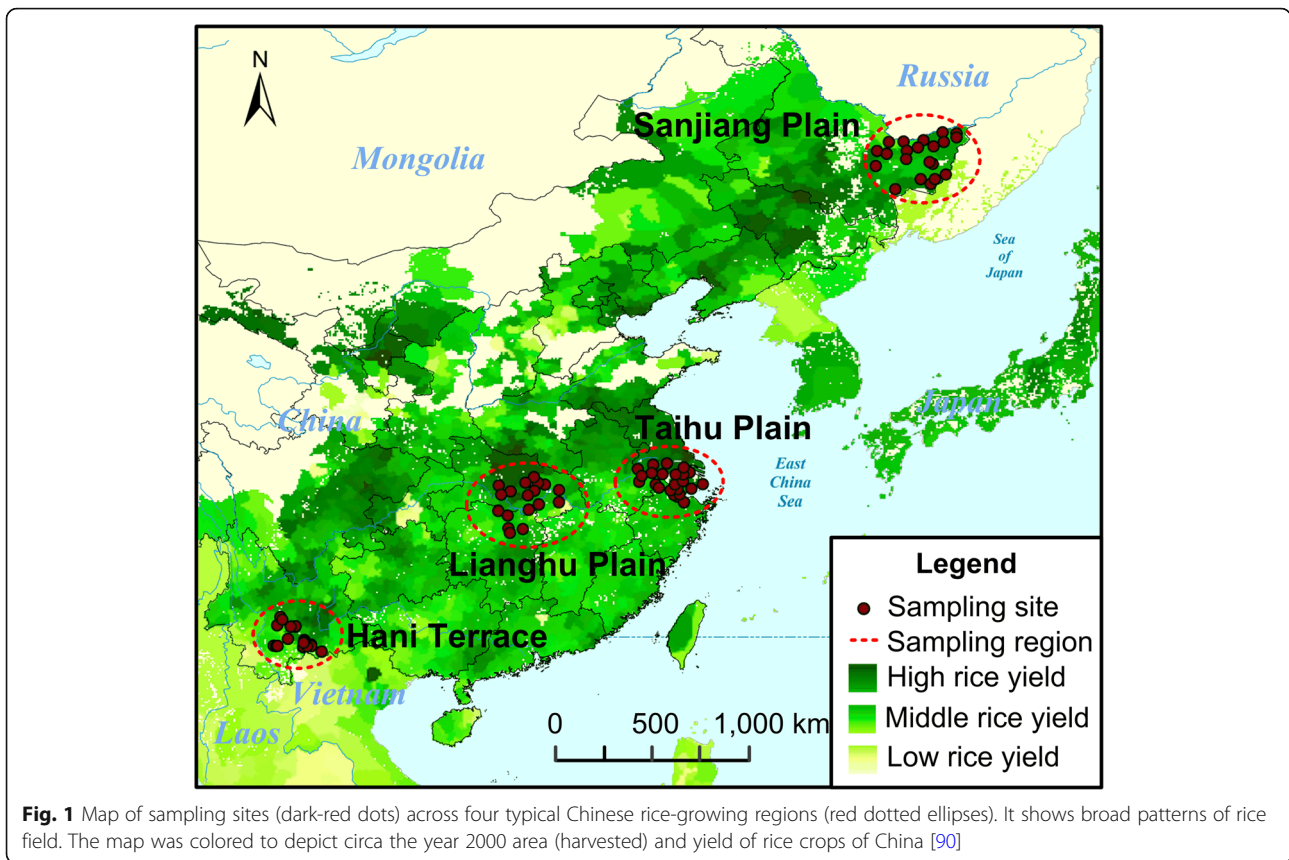
### Microbial and DOM biogeography in paddy soils

Across four rice-growing regions in China, we collected soil samples from 88 flooded paddy fields, wherein most rice plants were at the tillering phase (Fig. 1 and Table 1). Based on 16S rRNA gene sequencing, *Anaeromyxobacter* (1.9% ± 0.9%), *Geobacter* (1.0% ± 0.5%), *Anaerolinea* (0.9% ± 0.6%), and *Haliangium* (0.8% ± 0.3%) were the most abundant genera. Bacterial richness (Chao1 and observed species) and diversity (Shannon and PD whole tree) were significantly different between regions (Additional file 1: Figure S1A), with the lowest diversity and richness observed in samples from Sanjiang Plain ( $P < 0.05$ , Dunn's test). Microbial  $\beta$ -diversity (variance adjusted weighted UniFrac; VAW-UniFrac) was significantly correlated to the distance between sites (Mantel  $r = 0.52$ ,  $P < 0.001$ ; Fig. 2a). The proportions of the dominant taxa at each phylogenetic level differed significantly by region

(Additional file 1: Figure S2). Among the 20 most abundant genera, only *Gemmatimonas* and *Pseudolabrys* were stably abundant across regions, while the other genera were significantly different between regions (Fig. 2b).

DOM analysis revealed 81,759 compounds in total with an average of  $8262 \pm 1187$  compounds at each site (Additional file 2: Table S1). A core group of 18,538 molecules was observed in at least 10 sites; 12,791 of these compounds could be assigned putative molecular formulae. Chao1 of DOM molecules was significantly different between the four regions (Additional file 1: Figure S1B), while Observed species and Shannon diversity were not.  $\beta$ -diversity (Bray-Curtis) of DOM was still significantly associated with distance between sites (Mantel  $r = 0.24$ ,  $P < 0.001$ ; Fig. 2c). Weighted density plots of DOM components in van Krevelen diagrams (Fig. 2d) visualized the significant influence of geographic region on the abundance of DOM molecular groups (permutational multivariate analysis of variance, PERMANOVA  $r^2 = 0.17$ ,  $P < 0.001$ ). For example, unsaturated/aromatic hydrocarbons were enriched in Hani Terrace and Lianghu Plain samples (area  $\alpha$ ), phenolics were enriched in the Hani Terrace samples (area  $\beta$ ), polycyclic aromatics were enriched in Taihu Plain samples (area  $\gamma$ ), and polyphenols were enriched in Sanjiang Plain and Taihu Plain samples (area  $\delta$ ). Unsaturated aliphatics and peptides were enriched in samples of Hani Terrace (area  $\epsilon$ ).

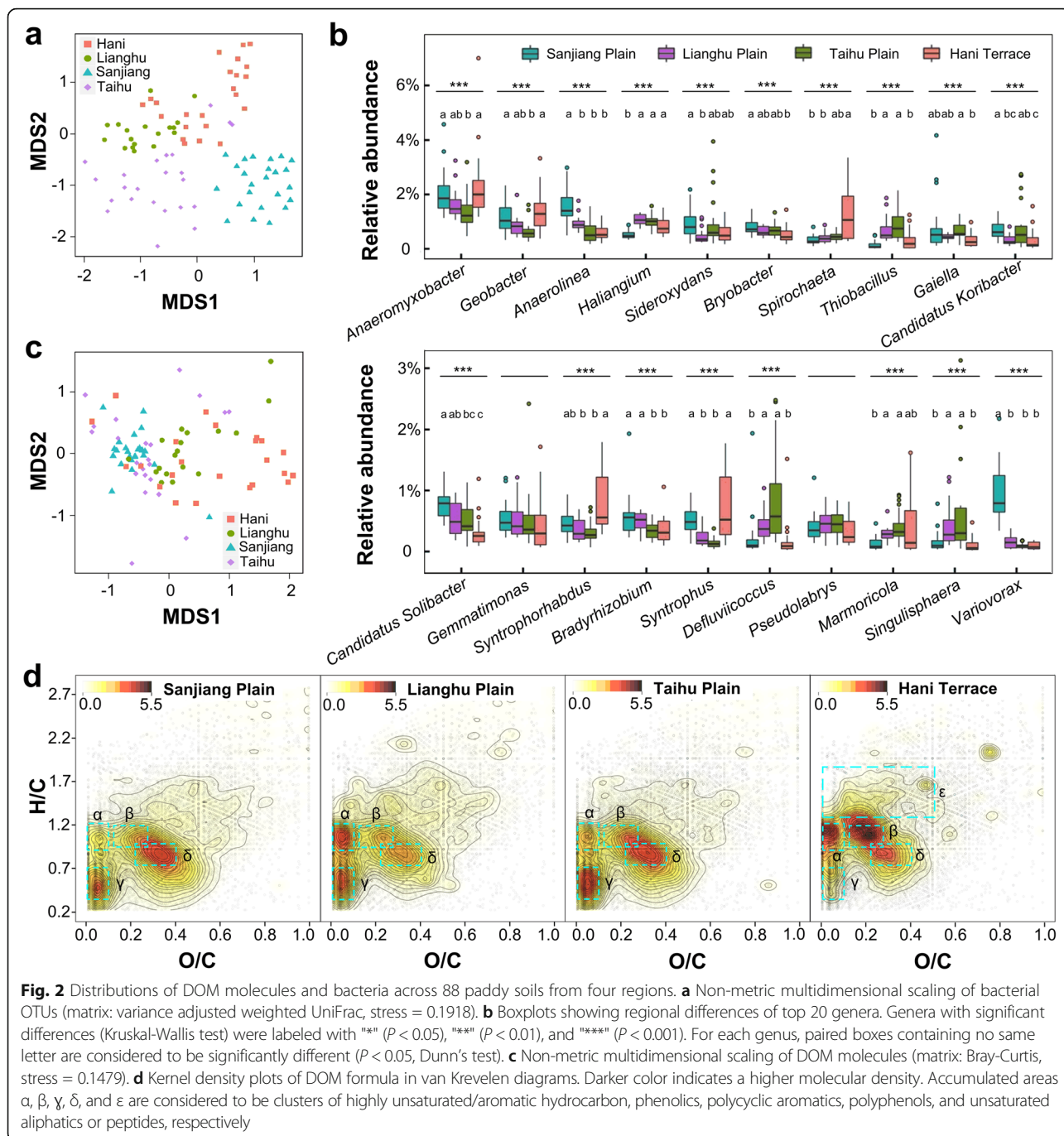
Canonical correspondence analysis (CCA) and partial CCA were performed to estimate the contribution of environmental factors to the variance in microbial and DOM diversity across sites. The variance in microbial community was best explained by the variance in DOM composition (27%; Fig. 3a and Additional file 3: Table S2), with the converse being true as well (26.8%; Fig. 3b and Additional file 4: Table S3). This was confirmed using Procrustes analysis, which demonstrated that the dissimilarity of DOM and microbial communities between samples was significantly and strongly correlated ( $mI2^2 = 0.66$ ,  $P_{\text{Monte Carlo}} < 0.001$ ). More specifically, the second principal coordinate (PCo2) of the microbial Bray-Curtis (5.5%) and VAW-UniFrac distances (5.9%) as well as functional potentials (3.0%–4.5%) were the main contributors of DOM variance; PCo1 of DOM Bray-Curtis distance (4.3%) was the main contributor of microbial variance, followed by DOM alpha diversity indices and PCo1–2 of phenolics, peptides, polyphenols, and polycyclic aromatic (in the column of V.E. CCA and V.E. pCCA, Additional file 3: Table S2 and Additional file 4: Table S3). Individual edaphic factors of pH (3.5%), conductivity (3.7%), real-time air temperature (RAT; 3.5%), real-time soil temperature (RST; 4.6%), tiller number (4.3%), and all geographic factors (4.4–6.9%) also described microbial variance (Additional file 3: Table S2). It



**Table 1** Various characteristics of the four typical Chinese rice-growing regions

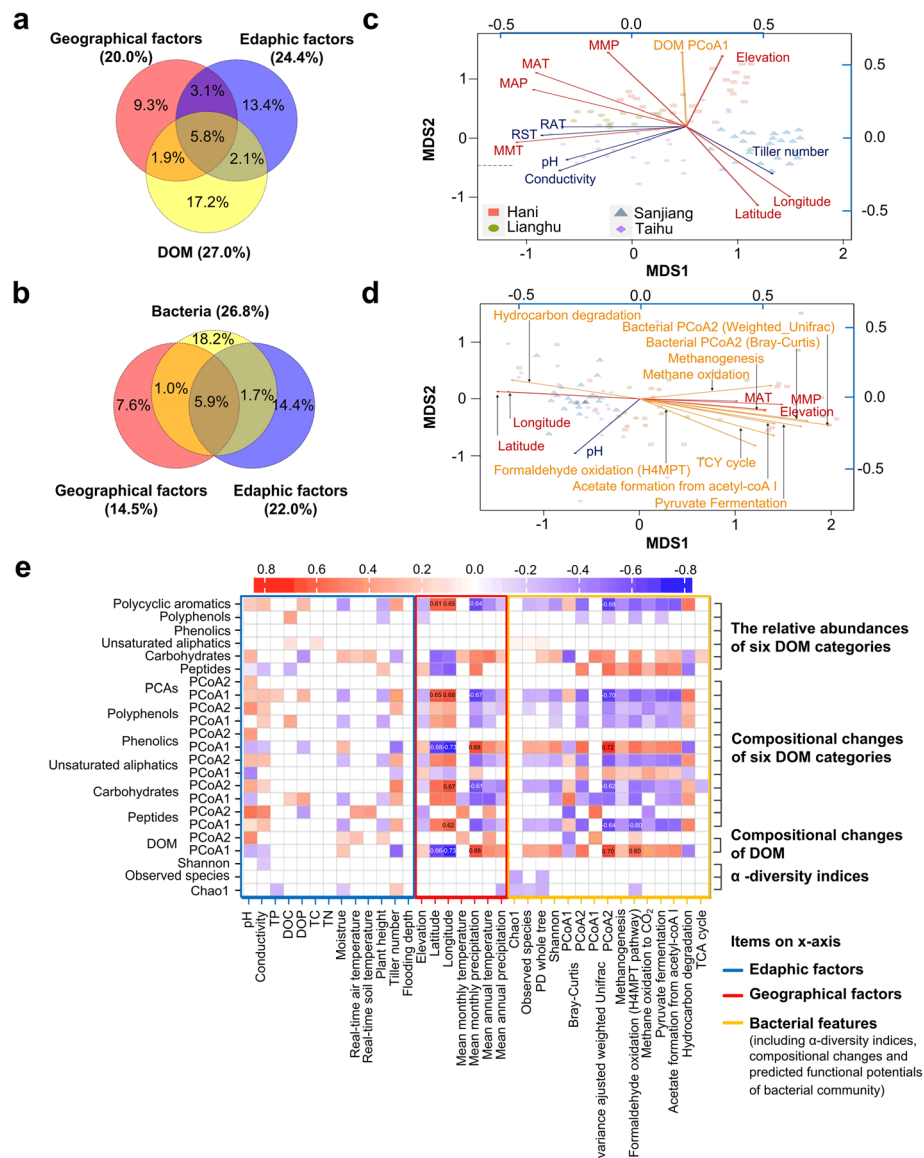
	Hani Terrace	Taihu Plain	Sanjiang Plain	Lianghu Plain
Administrative region	Yunnan Province.	Jiangsu, Zhejiang, Anhui Provinces, and Shanghai Municipality	Heilongjiang Province	Hunan and Hubei Provinces
Longitude range	100°55'27"–103°14'57"	118°38'12"–121°47'48"	130°12'45"–134°8'45"	111°50'51"–114°48'12"
Latitude range	22°46'44"–24°28'13"	30°3'19"–31°56'15"	45°16'53"–48°2'40"	28°34'7"–31°16'21"
Climate	3-D climate, (sub)tropics	subtropics	temperate	subtropics
Elevation (m)	350–3000	0–45	50–160	10–100
Rice cultivation history	~ 1200 years	~ 5000 years	~ 60 years	~ 4000 years
Rice cultivars ( <i>Oryza sativa</i> L.) [89]	Most are <i>japonica</i>	Half are <i>indica</i> and half are <i>japonica</i>	All are <i>japonica</i>	Most are <i>indica</i>
Cultivation method	traditional	modern and small-scale	modern and large-scale	modern and medium-scale
Mean pH	6.07 ± 0.60	6.60 ± 0.70	6.32 ± 0.32	6.80 ± 0.68
Mean annual temperature (°C)	16.21 ± 1.77	15.66 ± 0.41	2.85 ± 0.78	16.68 ± 0.35
Mean annual precipitation (mm/d)	3.20 ± 0.19	3.74 ± 0.19	1.86 ± 0.13	3.45 ± 0.29
Dissolved organic carbon (mg/ml)	23.51 ± 6.84	31.68 ± 11.41	29.28 ± 8.39	31.61 ± 14.40





should be noted that conductivity was used to quantify water-soluble ions [23]; tiller number and plant height were used to roughly indicate the rice growth stage and the size of plants, and they were classified into edaphic factors for simplicity. Meanwhile, edaphic factors of pH (3.4%), dissolved organic carbon (DOC; 3.3%), and geographic factors of elevation (4.4%), latitude (4.3%), longitude (4.1%), and mean monthly precipitation (MMP; 4.3%) and mean annual temperature (MAT; 3.2%) also described the variance in DOM (Additional file 4: Table S3).

To confirm these findings, we fitted these factors to unconstrained non-metric multidimensional scaling (NMDS) ordination (Fig. 3c, d, Additional file 3: Table S2 and Additional file 4: Table S3). Significant correlations ( $P \leq 0.001$ ) were observed between the variances of microbial/DOM composition and MMP ( $r^2 = 0.637/0.355$ ), MAT ( $r^2 = 0.742/0.200$ ), pH ( $r^2 = 0.407/0.409$ ), and elevation ( $r^2 = 0.342/0.347$ ), as well as latitude ( $r^2 = 0.613/0.358$ ) and longitude ( $r^2 = 0.553/0.305$ ; Additional file 3: Table S2 and Additional file 4: Table S3). The associations



**Fig. 3** Associations between DOM composition, bacterial community, and the environmental drivers. **a** The influences of geographical factors, edaphic factors, and DOM composition features on the bacterial community structure estimated via canonical correspondence analyses (CCA). **b** The influences of geographical factors, edaphic factors, and bacteria community features on the DOM composition estimated via CCA. The percentages represent the variance explained. **c, d** Multivariate analysis of microbial or molecular data and drivers using non-metric multidimensional scaling (NMDS). Ordinations are based on Bray-Curtis (**c**, stress = 0.1656; **d**, stress = 0.1479). Geographical factors, edaphic factors, and DOM composition (**c**) or bacteria (**d**) community were fit to the ordination using *envfit* function, respectively. Only factors with significance level  $\leq 0.001$  were shown. **e** Spearman's rank correlations between DOM diversity features (y axis) and edaphic, geographical, and bacterial factors (x axis), with color coded in blue, red, and yellow, respectively. All factors imported and their influences are listed in Additional file 3: Table S2 and Additional file 4: Table S3. MAP, mean annual precipitation; MAT, mean annual temperature; MMP, mean monthly precipitation; MMT, mean annual temperature; RST, real-time soil temperature; RAT, real-time air temperature

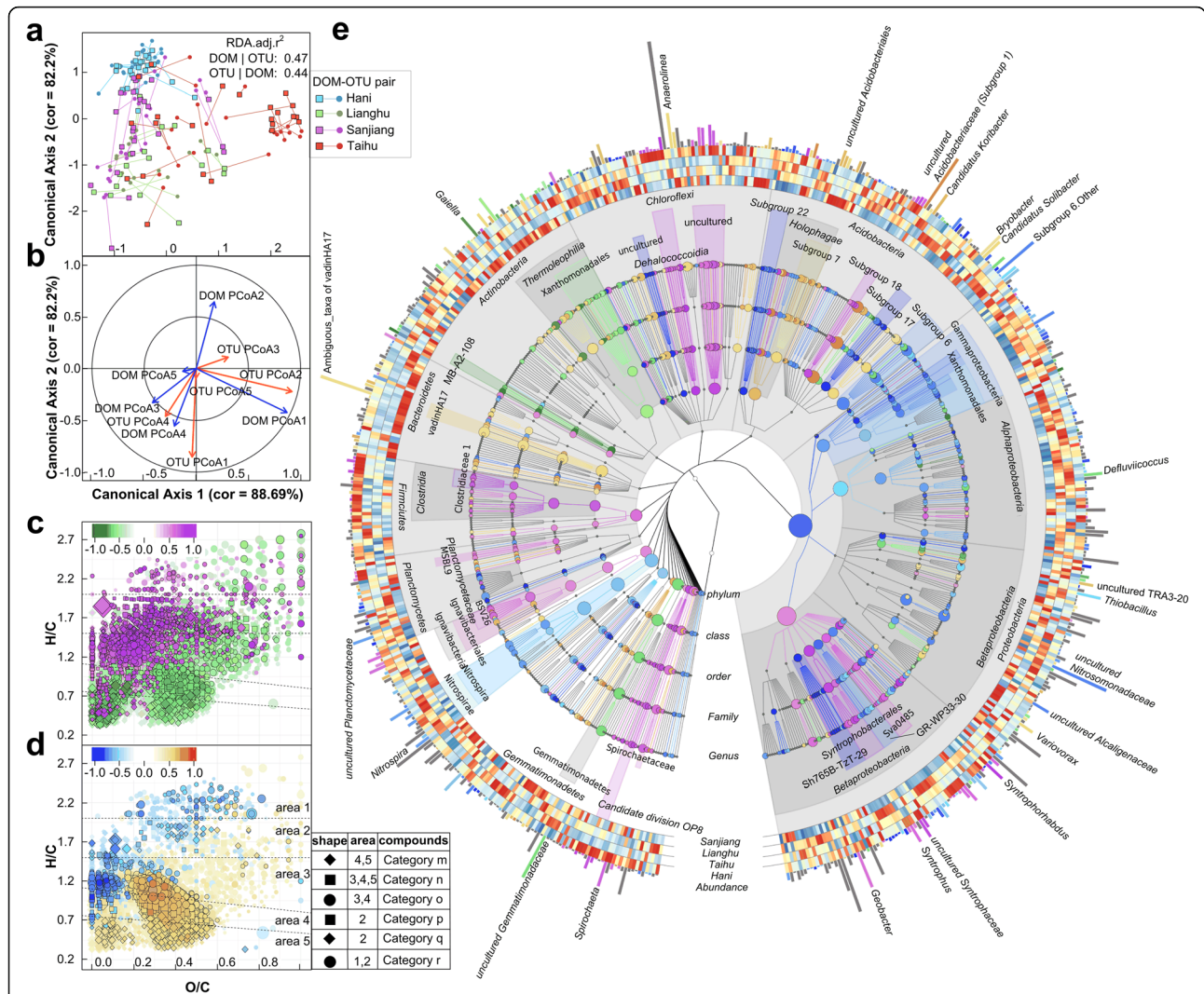
between edaphic, geographical, bacterial (*x*-axis), and DOM (*y*-axis) factors were calculated using Spearman's rank correlation (Fig. 3e). The PCo2 of the microbial Bray-Curtis and VAW-Unifrac distances, latitude, longitude, MMP, MAT, pH, and predicted functional potential were all strongly correlated with the PCo1 of DOM ( $P < 0.001$ ; Fig. 3e and Additional file 4: Table S3). MMP, as well as the second coordinates (PCo2) of microbial

Bray-Curtis and VAW-Unifrac distances, described the geographic variance in DOM (PCo1) and DOM features, while pH, and the PCo1 of microbial distances, described another pattern of DOM variance (PCo2) (Fig. 3e). Precipitation was positively correlated with the relative abundance of carbohydrates and peptides while negatively correlated with the relative abundance of polycyclic aromatic compounds (Fig. 3e and Additional file 1: Figure S3).

### Characterizing the association between DOM molecular and bacterial taxonomic composition

We performed canonical correlation analysis (CCorA) to characterize the association between bacterial and DOM composition (Fig. 4a, b). Using the PCo1–5 axes of DOM and microbial Bray-Curtis matrices, we observed four out

of five canonical axes with significant correlation coefficients ( $P < 0.01$ , chi-square test), which confirmed the multi-dimensional associations between bacterial community and DOM composition ( $P = 0.0001$ ; Fig. 4a). Each pair of ordinations along canonical axis represents two correlated dimensions of the multivariate matrices of microbial



**Fig. 4** Associations between DOM composition and bacterial community. Canonical correlation analysis was conducted using first five principal coordinates (PCo1–5). **a** Combined score plots of the corresponding ordinations pairs of DOM composition and microbial community along the first and second canonical axes. The lengths of connecting lines represent the dissimilarities of DOM composition and microbial community. **b** Plot showing loading coefficients of the five pairs of principal coordinates imported on the corresponding first two pairs of canonical axes. **c, d** van Krevelen plots of DOM molecules showing positive and negative Spearman's rank correlations with the DOM ordinations along the first (c) and second (d) canonical axes, indicating their association with microbial community dynamics. **e** Taxonomic cladogram showing positive and negative correlations with microbial ordinations along the first two canonical axes, indicating their association with DOM compositional changes. Rings of the cladogram provide a heatmap of the genera (> 0.01%) among the sampling regions with red and blue meaning more and less accumulated, respectively. Relative abundances of the taxa or genera are shown as the clade marker size or level 5 ring height. The color gradient bars in c and d, "purple to green" (for the first canonical axis) and "orange red to blue" (for the second canonical axis), indicate the values of coefficients. These color scales are also applied to the associations between taxa and microbial ordinations along the first and second canonical axes (e). Compounds category m: polycyclic aromatics; n: polyphenols and polycyclic aromatics with aliphatic chains; o: phenolic and highly unsaturated compounds; p: unsaturated aliphatics and aromatics with aliphatic chains; q: saturated fatty, sulfonic acids, and carbohydrates; r: N-containing compounds, i.e., peptides



community and DOM composition (e.g., Spearman's rank correlation  $\rho = 0.89$ ,  $P = 2.2 \times 10^{-16}$  for the first pair; Fig. 4b).

DOM molecules and microbial taxa that were responsible for the first two pairs of correlated ordinations of DOM and microbial community were figured out using Spearman's rank correlation test (Fig. 4c–e). Bacterial taxa showing significant correlation in the first canonical axis were enriched in Hani Terrace samples and attenuated in Sanjiang Plain samples. These included organisms (e.g., *Geobacter*, *Syntrophorhabdus*, and *Spirochaeta*) (Fig. 4e) that were positively correlated with highly unsaturated/aromatic hydrocarbon (Fig. 4c; area  $\alpha$  in Fig. 2d) and phenolics (area  $\beta$ ) and negatively correlated with polycyclic aromatics (area  $\gamma$ ) and polyphenols (area  $\delta$ ). However, some taxa (e.g., *Gaiella* and *Defluviococcus*) demonstrated an opposite trend (Fig. 4e). Also, other taxa and DOM compounds were correlated in the second canonical axis (Fig. 4d, e).

#### Factors of the covariation between microbial taxa and DOM molecules

To determine the factors of covariation between the microbiome and DOM, we correlated MMP, MAT, and pH against the DOM and microbial community ordinations along the first two canonical axes using Spearman's rank correlation test. For covariation along the first axis (Fig. 4b), MMP and pH showed strong and significant correlations ( $\rho = -0.68/-0.65$ ,  $P = 3.85 \times 10^{-13}/9.03 \times 10^{-12}$  for pH with DOM/microbial ordination;  $\rho = 0.52/0.46$ ,  $P = 2.17 \times 10^{-7}/6.48 \times 10^{-6}$  for MMP with DOM/microbial ordination), while MAT was not correlated ( $P > 0.1$ ). For covariation along the second axis (Fig. 4b), MAT, MMP, and pH showed strong and significant correlations ( $\rho = -0.40/-0.59$ ,  $P = 1.01 \times 10^{-4}/1.33 \times 10^{-9}$  for pH with DOM/microbial ordination;  $\rho = -0.51/-0.50$ ,  $P = 3.93 \times 10^{-7}/8.01 \times 10^{-7}$  for MMP with DOM/microbial ordination;  $\rho = -0.71/-0.72$ ,  $P = 9.50 \times 10^{-15}/2.30 \times 10^{-15}$  for MAT with DOM/microbial ordination). In our study, since MMP and MAT changed in opposite trends along the increased elevation in Hani Terrace sites, MMP and MAT, as well as pH, were not correlated with each other ( $P > 0.05$ ).

#### DOM composition correlates with microbial functional potential

To characterize the microbial functional potential, four samples from each of the four regions were selected for deep shotgun metagenomic sequencing. The most abundant genera were *Streptomyces* (3.8%), *Anaeromyxobacter* (3.2%), *Bradyrhizobium* (3.0%), *Mycobacterium* (1.7%), *Solibacter* (1.3%), and *Geobacter* (1.2%). The abundance of this particular collection of taxa roughly paralleled what was previously observed in the 16S

rRNA analysis (for details see Additional file 5 and Additional file 1: Figure S4).

The relationship between DOM composition and the relative abundance of metagenomic functional genes was determined by CCorA using PCo1–5 axes of DOM and PCo1–2 axes of FOAM (Functional Ontology Assignments for Metagenomes, a functional gene database) orthologs (explained 70% and 65% variances, respectively). We observed two canonical axes with significant correlation coefficients ( $P < 0.01$ , chi-square test), which confirmed the association between microbial function potential and DOM composition ( $P = 0.012$ ,  $r_1 = 0.89$ ,  $r_2 = 0.76$ ). The association between DOM and FOAM orthologs (Fig. 5a) followed a similar trend to the association between DOM and microbial taxonomy (Fig. 4c, d). Redundancy analysis (built-in function of CCorA) indicated that the functional potential distribution could be predicted by the DOM distribution with an adjusted  $R^2$  of 0.70, while the distribution of DOM was correlated to functional potentials with an adjusted  $R^2$  of 0.35. Consistently, 109 out of 129 observed FOAM level 2 functional pathways were significantly ( $|r| > 0.5$ ,  $P < 0.05$ ) correlated with the variances of DOM in the first two canonical axes (Additional file 6: Table S4).

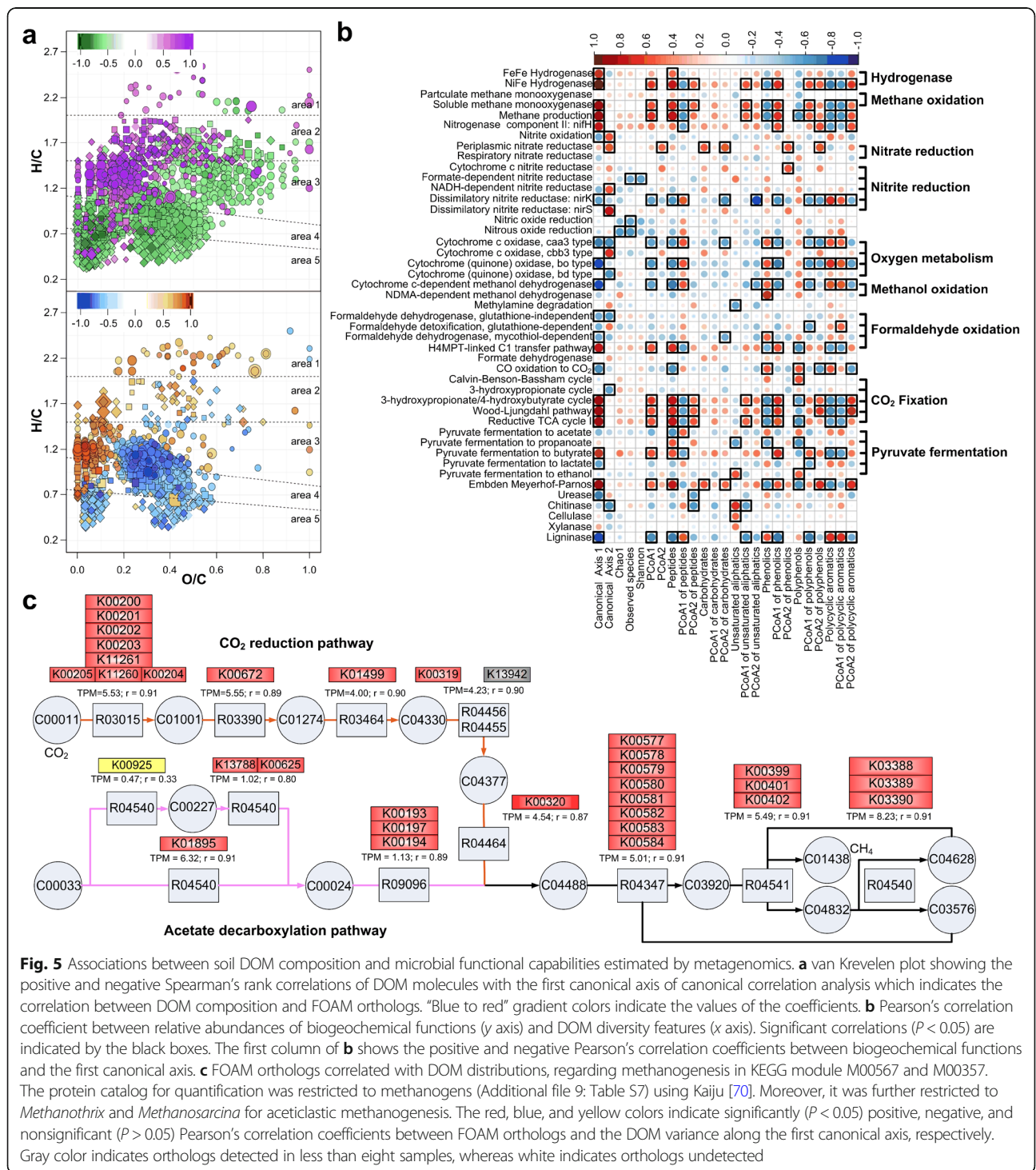
To avoid risk of false correlations, we further used marker genes of certain biogeochemical functions to qualitatively evaluate the associations (Fig. 5b). Consistently, functional genes involved in glycolysis (Embden-Meyerhof-Parnos, EMP), anaerobic C-fixation (light-independent), pyruvate fermentation (to butyrate), methanogenesis, methane oxidation (soluble methane monooxygenase), and hydrogen metabolism, as well as H4MPT-linked C1 transfer pathway (hydrogenotrophic methanogenesis associated), showed positive correlation with compositional changes in DOM along the first canonical axis, whereas those related to pyruvate fermentation (to lactate), carbon monoxide oxidation, urea degradation, nitrite oxidation, dissimilatory nitrite reduction (*nirK*), and ligninase, as well as the *caa3*-type and *bo*-type cytochrome oxidases, were negatively correlated (Fig. 5b). Compositional changes of DOM along the first canonical axis were characterized by increases in highly unsaturated/aromatic hydrocarbon and phenolics, peptides, and unsaturated aliphatics ( $O/C < 0.5$ ), as well as decreases in polycyclic aromatics and polyphenols (Fig. 5a, upper panel). FOAM function categories associated with the tricarboxylic acid (TCA) cycle and homoacetogenesis were positively correlated with compositional change of DOM along the first canonical axis, whereas hydrocarbon degradation, cellular response to oxidative stress, and fatty acid oxidation pathways were negatively correlated (Additional file 6: Table S4).

To confirm the findings from this limited metagenomic dataset, Tax4fun [24] was used to predict the abundance of functional genes based on the 16S rRNA

amplicon data. The correlation between predicted functional potential and DOM was mostly consistent with the correlation observed for the real metagenomic data (Additional file 5 and Additional file 7: Table S5). The metagenomic taxa that correlated with DOM along the first canonical axis of Fig. 4c also showed significant correlation with DOM along the first canonical axis of

Fig. 5a. These include *Geobacter* (Pearson's correlation coefficient,  $r = 0.83$ ,  $P = 6.32 \times 10^{-5}$ ), *Syntrophobacterales* ( $r = 0.94$ ,  $P = 5.98 \times 10^{-8}$ ), *Spirochaeta* ( $r = 0.92$ ,  $P = 3.23 \times 10^{-7}$ ), *Thermoleophilia* ( $r = -0.62$ ,  $P = 1.01 \times 10^{-2}$ ), and *Gaiella* ( $r = -0.72$ ,  $P = 1.53 \times 10^{-3}$ ).

The relative abundance of methanogens (0.43%) as predicted from 16S rRNA analysis had the same correlation



with the DOM variance along the first canonical axis (Pearson's correlation coefficient  $r = 0.93$ ,  $P = 2.16 \times 10^{-7}$ ), as the genes encoding for methanogenesis (Fig. 5c). Despite their distinctly different substrate range, we found that all methanogenic genera correlated with DOM variance along the first canonical axis (Additional file 8: Table S6). Consistently, both functional ortholog groups (FOAM orthologs) of hydrogenotrophic and acetoclastic methanogenesis showed significant correlation with DOM variance along the first canonical axis (Fig. 5c and Additional file 6: Table S4). Strikingly, methanogenesis via  $\text{CO}_2$  reduction (TMP = 4.54, K00320) had a greater functional potential than acetoclastic methanogenesis (TMP = 1.93, K00193, K00194, K00197) in all paddy sites. Methanogenesis can be performed by syntrophic methanogenic consortia via direct interspecies electron transfer (DIET), mediated by electrically conductive pili (e-pili) or biochar and other conductive materials [25, 26]. So we calculated the Pearson's correlation coefficient between Fe(III)-reducing bacteria with/without e-pilin encoding genes and the first canonical axis of DOM (Additional file 9: Table S7). Seventeen out of 19 bacteria encoding e-pilin were significantly correlated with DOM variance along the first canonical axis, while 35 out of 70 bacteria without e-pili were also significantly correlated. Consistently, electrically conductive pilin (e-pilin, 46.6%) was significantly positively correlated to DOM compositional change, while long type IVa pilin (53.4%) was not ( $r = 0.61$ ,  $P = 0.013$  for e-pilin;  $P > 0.05$  for long type IVa pilin).

## Discussion

In this study, we demonstrated that DOM molecular distribution correlates with microbial community structure, taxonomy, and functional potential in paddy soils from sites representing gradients of temperature, precipitation, pH, and human activity. Both molecular distribution of DOM and microbial community structure exhibited significant biogeographic patterns. We considered how biotic and abiotic factors, like geographic distance, MMP, pH, MAT, elevation, and the interactions between microbial communities and DOM molecules, might drive the biogeography of microbial communities and the chemogeography of DOM molecules. These results expand our knowledge of how microbes and DOM molecules are distributed throughout the rice paddy ecosystem.

While variations in the type and abundance of paddy soil DOM molecules have typically been chalked-up to temperature, moisture, pH, and mineralogy [15, 27–29], here, we demonstrated that the microbiome describes DOM variance to a greater extent than any other edaphic or geographic factor investigated. This result emphasizes the overriding potential impact that biotic function factors can have on constructing the DOM

heterogeneity when compared to other abiotic functions in paddy soil. The heterogeneous characteristic of DOM is partly attributed to the chemistry of plant-derived compounds and their decomposition byproducts [15, 30, 31]. For instance, genes associated with ligninase or hydrocarbon (mostly aromatics) degradation were positively correlated with polyphenolic and phenolic compounds and polycyclic aromatics in DOM of flooded paddy soils. This indicated biodegradation processes from large biopolymers (e.g., lignin) towards small molecules with concurrent increases in polar and ionizable groups and thus with increase in solubility [5]. Meanwhile, these compounds (known as terrestrial DOM) characterized by high cyclization were selectively coprecipitated with iron at the redox interface during their upward diffusion together with the ferrous ion [3]. Dissimilatory Fe(III)-reducing microbes, e.g., *Geobacter*, were responsible for the production of ferrous ion and thus the coprecipitation of terrestrial DOM with Fe(III) in paddy soil [32, 33]. Many of the DOM compounds may also be derived from microbial residues, e.g., cellular materials and extracellular secretions [34, 35]. Recently, SOM chemistry study in model soils also demonstrated that the accumulation of chemically diverse SOM was driven by distinct microbial communities, per se, rather than the substrates they utilized [36].

Conversely, we also realized that the abundance and distribution of DOM explained variations in the paddy soil microbiome better than any other examined edaphic or geographic factors. The chemical nature of SOM was reported to affect the structure and functioning of paddy soil microbiota [6, 37]. In our study, more biodegradable substances, e.g., peptides and carbohydrates, favored microbes (e.g., *Geobacter*) utilizing mainly simple C forms and functions regarding glycolysis (EMP) and TCA cycle in the flooded paddy. This heightened concentration of biodegradable substances also facilitated pyruvate fermentation, and associated metabolic processes, e.g. methanogenesis and homoacetogenesis [38]. Microbial consortia that cooperatively exchange electrons were pivotal in the anaerobic processing of SOM [25, 26]. Therefore, the increase of Fe(III)-reducing bacteria encoding e-pilin may promote the propagation of hydrogenotrophic methanogens. Consistently, we found that gene encoding e-pilin, as a potential indicator of the DIET [39, 40], was significantly correlated with DOM variance. Moreover, a recent study proved that the acetoclastic methanogen *Methanotherix* spp. receives electrons for the reduction of  $\text{CO}_2$  via DIET from microorganisms expressing e-pilin genes, e.g., *Geobacter* [39]. Decreased abundance of DOM compounds with quinone moieties, which were likely derived from polyphenolic and phenolic compounds and polycyclic aromatics, reduced the possibility that electrons are being transferred to these



DOM and used in Fe(III) reduction [41, 42], thus increasing the electrons available for methanogenesis.

Of the abiotic factors, precipitation (MMP), temperature (MAT), and pH explained the majority of the main variance in DOM composition across the continental scale paddy fields. pH and temperature have also been reported to drive soil microbial community composition at continental scales [10, 11]. Here, we have found that temperature (MAT) and pH were significantly correlated with the covariations between DOM and microbial community. Therefore, the influences of pH and temperature on DOM diversity may be mediated by microbial community. Consistently, pH was not strongly correlated with the majority of individual DOM molecules (Additional file 1: Figure S3), but it could explain the DOM component variance, implying an indirect effect on DOM molecular distribution. Elevated temperature may stimulate the biodegradation of plant residues, but the consequences for microbial-derived residues are less clear [11, 36].

Precipitation shapes DOM chemodiversity as it enhances upward movement of ferrous ions and DOM molecules from deep soils, and influences the influx of selective DOM from the surface soil. Precipitation events dilute not only the ion and DOM concentrations of the standing water per se, but also that of the irrigation water sources (i.e., the nearby bodies of water), enhancing the upward movement of Fe(II) and DOM from deep soils [43]. Meanwhile, terrestrial DOM would be selectively trapped at the Ap horizon (oxic and partly oxic) via coagulation with the precipitating Fe(III) during the upward diffusion [1, 3, 44], which then accelerates the upward diffusion of these compounds. Moreover, terrestrial DOM is relatively harder to be regenerated when compared to carbohydrate and peptide [45]. Besides, abundant rainfall preserves the gradient in reduction potential across depth, which increases not only the mineral reduction but also the opportunity of DOM reduction in paddy fields [1, 6, 43]. As evidence, significant correlations were found between precipitation (MMP) and genes encoding *caa3*-type and *bo*-type cytochrome oxidases (Fig. 5b) and cellular response to oxidative stress (Additional file 7: Table S5). The *caa3*-type and *bo*-type cytochrome oxidases are mostly found in aerobes, while *cbb3*-type and *bd*-type cytochrome oxidases have been reported to be utilized by anaerobes or microaerophiles in microaerobic energy metabolism [46, 47]. In this case, precipitation may also shape microbial community.

Factors like parental material, redox state, fertilization level, pesticide application, mineralogy, rice cultivar, and growth stage may also influence the geography of the microbial community and DOM [29, 48–53]. In submerged paddy soil, rice aerenchyma enables the transport of atmospheric O<sub>2</sub> to the roots [54], influencing soil redox states [55];

moreover, rice straw and stubble are assumed to provide substrates for microbial activity in the early growth stage, while exudates become more important during late tillering and ripening [56, 57]. Therefore, the effect of the rice plant on the soil microbial community largely depends on the plants growth stage. In this study, the tiller number of the rice (indicating the growth stage) was significantly correlated with DOM variance (PCo1) and DOM features (Fig. 3e). Recently, a relatively comprehensive study on the DOM chemodiversity of paddy soil and factors, including mineral elements of Fe, Mn, Al, Mg, Ca, and Si, demonstrated that the iron complexing index (Fe<sub>p</sub>/Fe<sub>R</sub>), together with pH and C/N ratio, were key factors controlling DOM profiles [29]. Another study on agricultural, meadow, and forest soils revealed that pH and nitrate significantly affect the chemical composition of DOM molecules [14]. These factors were also separately found to significantly correlate with the microbiome in paddy soils [48, 58, 59]. Although rice cultivation management dominated the microbial community assembly in paddy soils [60], the soil parent material was also influential [58], and hence, the influence on DOM distribution should be further investigated. Although these previous studies and the current one presented here principally focus on the spatial distribution of the microbiome and DOM molecules, how these patterns change over time is also important [61]. Redox potential is one of the key temporal factors and is controlled by soil ventilation [1]. Temporal changes of irrigation management, precipitation, and even light intensity may quickly change the redox condition in soils. Researchers have revealed that redox potential could significantly shift microbial community composition [6, 53]. Retention of certain DOM molecules by soil minerals and their subsequent stabilization against microbial decay were also largely dependent on the redox state [3, 5]. However, it remains unknown whether the temporal dynamics of microbial communities correlates shifts in DOM composition. Therefore, there is a continued need for new, well-controlled studies to further elucidate DOM chemogeography and microbial biogeography.

## Conclusions

Understanding the relationship between soil DOM and microbial community structure and function remains a research goal for biogeochemists, especially at the molecular level [8, 52]. Here, we integrated mass spectra and genomics data to characterize the association between DOM molecular distribution and microbial diversity and applied gene-centric analysis to elucidate the microbial metabolic potential that responds and shapes DOM heterogeneity. DOM chemodiversity was significantly and broadly correlated with the taxonomy and functional potential of the microbial community in paddy soil. Besides pH and temperature, precipitation was also found to be a potential factor of microbial

community and DOM chemical distribution. These findings are foundational, but could be of great importance for environmental and agricultural management in paddy soils.

## Methods

### Site selection and soil sampling

Soil samples were collected from 88 flooded paddy sites across four typical Chinese rice-growing regions in 2014 and 2015 (Fig. 1). Most of the soils were sampled during the tillering phase of rice plants in the paddy fields. Among 88 sampling sites, there were 23 from Hani Terrace, 24 from Sanjiang Plain, 18 from Lianghu Plain, and 23 from Taihu Plain (Table 1). At each site, soil cores (2.5 cm diameter by 15 cm depth) were sieved (2 mm) and homogenized, and plant materials were removed immediately before sealing and transportation. For more details about soil sampling and characteristics measurements (Additional file 10: Table S8), please see supplementary methods in Additional file 11.

### FT-ICR-MS data analysis

FT-ICR-MS samples were prepared and measured according to Kellerman et al. [4] with some modifications (for details see supplementary methods in Additional file 11). Detected peaks were considered if the signal-to-noise ratio was greater than five. After calibration, different spectral peaks were clustered into operational units within a mass tolerance with  $m/z$  difference ratios less than  $1 \times 10^{-6}$ . The detailed methods for calibration and clustering are described in Additional file 11. Clusters with fewer than ten peaks were not considered for further annotation.

Based on the two mandatory and two optional steps for peak annotation by Koch et al. [16], we introduced a carbon isotope ratio-based molecular annotation approach (Additional file 1: Figure S5), in which molecular formulae are assigned to peaks according to stringent criteria with elemental combinations of  $C_{1-100}H_{1-150}O_{0-50}N_{0-4}P_{0-1}S_{0-1}$ . The isotope-based approach first tries to find the carbon isotope peak of a certain large peak according to mass differences and then calculates the potential C number in the molecular formula based on the relatively stable ratio of naturally occurring  $^{13}C$ -isotope to  $^{12}C$  (i.e., 1.07%) and the intensity ratio of the two peaks. At the same time, several alternative formulas are calculated according to an a priori definition of elements and unequivocal exclusion criteria. Then, the carbon numbers of these formulas are subtracted by the potential C number for carbon number differences ( $C_{dev}$ ), and the molecular formula with the smallest  $C_{dev}$  is chosen. In this study,  $C_{dev}$  was defined as acceptable when the  $C_{dev}$  was (-3, 1) [16]. Annotated formulas were then used as a scaffold, and a “chemical building block” approach was adopted to annotate the rest of the peaks. Meanwhile, the relevant  $^{13}C$ ,  $^{15}N$ ,  $^{34}S$ ,  $^{33}S$ ,  $^{18}O$ ,

$^{17}O$ ,  $^2H$ ,  $^{13}C_2$ ,  $^{13}C_3$ , and  $^{13}C^{34}S$  isotope molecules were all determined, if detected.

The annotated molecules were assigned to compound categories based on the stoichiometry of their molecular formulas [45]: polycyclic aromatics (aromaticity index,  $AI > 0.66$ ) [62], polyphenols and polycyclic aromatics with aliphatic chains ( $0.66 \geq AI > 0.50$ ), phenolic and highly unsaturated/aromatic compounds ( $AI \leq 0.50$  and  $H/C \leq 1.5$ ), unsaturated aliphatics and aromatics with aliphatic chains ( $N = 0$ ,  $AI \leq 0.50$ ,  $2.0 \geq H/C \geq 1.5$ ), saturated fatty, sulfonic acids and carbohydrates ( $H/C \geq 2.0$  or  $O/C \geq 0.9$ ), and N-containing compounds, i.e., peptides ( $N \geq 1$ ,  $AI \leq 0.50$ ,  $2.0 \geq H/C \geq 1.5$ ).

### 16S rRNA sequencing and analysis

DNA was extracted using the MoBio PowerSoil DNA extraction kit following the manufacturer’s protocol. The concentration and qualification of the total DNA were examined by electrophoresis on 1% agarose gels, which was diluted 1 ng/ $\mu$ L using sterile water before downstream processing. PCR procedure was carried out as described previously [63]. Briefly, the V4-V5 region of the 16S rRNA gene from each soil sample was amplified using the F515/R907 primer set with a unique 6-nt barcode at the 5’ of the forward primer [64]. All PCR reactions were carried out with Phusion® High-Fidelity PCR Master Mix (New England Biolabs). After electrophoresis on 2% agarose gel, PCR products with bright main strip between 400 and 450 bp were mixed in equidensity ratios and purified with Qiagen Gel Extraction Kit (Qiagen, Germany). Sequencing libraries were generated using the TruSeq® DNA PCR-Free Sample Preparation Kit (Illumina, USA) following the manufacturer’s recommendations, and index codes were added. After quality assessing, the libraries were sequenced on an Illumina HiSeq2500 (Novogene, China), and 250 bp paired-end reads were generated.

The raw sequence data were processed using the QIIME v1.9.1 pipeline [65]. Firstly, the forward and reverse Illumina reads were joined using the default setting. Then, the multi-lane fastq data were demultiplexed and quality filtered ( $Q30 \geq 75\%$  and  $Q20 = 100\%$ ). Chimeras were identified using “identify\_chimeric\_seqs.py” with “-m usearch61” and then removed. A total of 15,339,665 reads were kept with a number over 43,000 for each sample. Filtered sequences were clustered into operational taxonomic units (OTUs) using the function “pick\_open\_reference\_otu.py” against the SILVA 119 database [66], based on a 97% consensus threshold. Then, the singletons were removed and taxonomy was assigned using the RDP classifier against the SILVA database. R package Tax4fun [24] with UProC long read mode was used to predict the functional capabilities of microbial community in each sample based on assigning OTUs of

16S rRNA gene to the reference sequences in the SILVA (version 119, 97 set) database via SortMeRNA [67]. HUMAnN2 [68] was used to map the resulted orthologs to functional ontology of FOAM database [69].

#### Shotgun sequencing, metagenome assembly, and annotation

A sum of 16 samples (4 samples from each region) was randomly selected for shotgun metagenome sequencing. Total DNA was extracted using the same method as that of 16S rRNA sequencing. DNA concentration was measured using Qubit® dsDNA Assay Kit in Qubit® 2.0 Fluorometer (Life Technologies, USA). Using 1 µg DNA per sample as input material, sequencing libraries were generated using NEB Next® Ultra™ DNA Library Prep Kit (NEB, USA), and index codes were added to attribute sequences to each sample. Briefly, the DNA was broken into 350 bp fragments using sonication, polished and extracted using the AMPure XP system. Libraries were prepared on a cBot Cluster Generation System following the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina HiSeq platform (Novogene, China) and 150 bp paired-end reads were generated.

Paired-end reads were quality controlled using Readfq v8 (<https://github.com/cjfields/readfq>): sequences with more than 40 bases, with quality score lower than 38, or with N bases more than 10 were filtered. The adapter was also removed from the sequences. Metagenome sequencing yielded about 12.8 G clean bases per sample (Additional file 12: Table S9). Taxonomical classification of the sequencing reads of each sample was performed using Kaiju [70] with greedy-5 mode against an nr-derived database including proteins from archaea, bacteria, viruses, fungi, and microbial eukaryotes. Based on the classification result, relative abundances of Fe(III)-reducing bacteria with or without e-pili were calculated (Additional file 9: Table S7). The 93 known Fe(III)-reducing microorganisms for which genomes are available were obtained from a previous study [39]. Metagenome assemblies were conducted for each sample using MEGAHIT v1.1.1 [71] with "--presets meta-large." Metagenome assemblies yielded a total of 41.8 M contigs over 200 bases length and 10.1 M contigs over 500 bases length (average length was 458 bases; Additional file 13: Table S10). A total of 53.5 M nucleotide sequences or protein translations of genes were predicted from the contigs (≥ 200 bp) of each sample using prodigal [72] with "-p meta." These genes or proteins were then pooled together to form a gene catalog and a protein catalog, respectively. The protein catalog was annotated using the FOAM database [69] and hmmer3.1 [73] to obtain function orthologs as defined in KEGG Orthology [74]. We chose profile's trusted cutoffs to set

all thresholding. The resulted orthologs were then mapped to an associated functional ontology of FOAM database [69] to describe the functional groups and hierarchy. Metagenomic contigs were annotated to gene and enzyme using prokka pipeline [75]. The marker genes used in the analyses of biogeochemical functions were selected from a hidden markov model database [76] with a few modifications. Pfam [77] and TIGRFam [78] protein families were assigned using hmmer3.1 [79]. Profile's trusted cutoffs were used as thresholds. The presence of type IV pilA genes was estimated by assigning the gene catalog to nucleotide sequence database with 33 e-pilin genes and 27 long pilin genes [40] using Diamond [80] with parameters set as following: --more-sensitive, -e 0.00001, -l 20. For each gene, the best blast hit (one HSP > 60 bits) result was selected for downstream analyses. To quantify the annotated genes in each sample, we mapped the paired-end reads back to the assemblies according to the pipeline described here: <http://metagenomics-workshop.readthedocs.org/>, together with the other tools, i.e., bowtie2 [81], samtools [82], and htseq [83]. As suggested by the pipeline, we used the TPM (Transcripts Per Kilobase Million) method [84] to normalize abundance values in metagenomes. The gene encoding for acetyl-CoA synthetase (K01895) is normally multi-copy, so was not chosen for the comparison.

#### DOM and bacterial diversity calculations and multivariate analysis

Accumulation and rank abundance curves of DOM were calculated with the sum-normalized intensity of non-singleton data using R package *Biodiversity R* [85] and *vegan* [86]. Bray-Curtis dissimilarity was used to compute the sparse matrices of DOM molecules and bacterial community. The molecular and bacterial alpha diversities and beta diversities were calculated using QIIME v1.9.1 [65]. VAW-UniFrac dissimilarity was calculated using R package *GUniFrac* [87]. Modified R function *kde2d* weighted from *kde2d* in *MASS* package [88] was used to perform two-dimensional kernel density estimation with an axis-aligned bivariate normal kernel in van Krevelen diagram for the density of DOM molecules in each sampling region. Median values of molecular abundances were defined as the weights parameter. Mantel test was conducted to determine whether two distance matrixes were significantly correlated. PERMANOVA test was conducted to determine whether DOM molecular Bray-Curtis dissimilarity was significantly different between regions. CCA and NMDS were performed using rounded intensities rarefied at the depth of 43,000 for both DOM and bacterial communities (or their dissimilarity matrixes). Partial CCA was used to calculate the independent influences of different categories or parameters on DOM or microbial variance.

Principal coordinates analysis (PCoA) was used to calculate the gradient in compositional changes of microbial community (based on Bray-Curtis or VAW-UniFrac), DOM (based on Bray-Curtis), and different DOM categories (based on Bray-Curtis). Procrustes rotation and Monte Carlo permutation test (permutation = 9999) were performed using the two coordinate matrices (output of PCoA) based on their Bray-Curtis dissimilarities. We performed the test using the first ten axes of DOM and bacterial coordinate matrices (explained 67% and 65% variations, respectively). CCorA was conducted using PCo1–5 axes of DOM composition and PCo1–5 axes of microbial community composition or PCo1–2 axes of functional orthologs distance matrices (Bray-Curtis dissimilarities) as imports following the procedure described by Osterholz et al. [19]. These analyses were all conducted using *vegan* package [86].

## Additional files

**Additional file 1:** Alpha-diversities of bacterial OTUs and DOM molecules in the tested soil samples across four typical paddy fields. **A:** Boxplots of alpha-diversities of bacterial OTUs using four indices: Chao1, observed species, Shannon and PD whole tree. **B:** Boxplots of alpha-diversities of DOM molecules using three indices: Chao1, observed species, and Shannon. We randomly subsampled 43,000 sequences and 9000 per sample ten times to correct for differences in sequencing depth for bacterial OTUs and DOM molecules. In the boxplots, the symbols indicate the following: box, lower and upper quartiles; horizontal line, median value; whiskers, lower and upper inner fence. The circle above or below the box plots indicates outliers. Differences among the four regions were tested using non-parametric Kruskal–Wallis test ( $P < 0.05$ ). Paired boxes containing no same letter are considered to be significantly different (Dunn's test,  $P < 0.05$ ). Figure S2 Boxplots of regional differences of top taxa classified using 16S rRNA gene. Differences among the four areas were tested using non-parametric Kruskal–Wallis test followed by Dunn's test for pairwise multiple comparisons. **A:** Regional differences of the top 10 phyla. **B:** Regional differences of the top 10 classes. **C:** Regional differences of the top 15 orders. **D:** Regional differences of the top 20 families. In the boxplots, the symbols indicate the following: boxes, the interquartile range (IQR) between first and third quartiles; horizontal line, median value; whiskers, the ranges of lower and higher values within  $1.5 \times \text{IQR}$  from the first and the third quartiles, respectively; circles, outliers beyond the whiskers; \*,  $P < 0.05$ ; \*\*,  $P < 0.01$ ; \*\*\*,  $P < 0.001$ . Figure S3 Van Krevelen plots of DOM molecules, showing Spearman's rank correlations with the factors used in Fig. 3e. Only DOM molecules with "BH" FDR-adjusted  $P \leq 0.05$  and  $|\rho| \geq 0.3$  are shown here. Strong correlations ( $|\rho| \geq 0.5$ ) were indicated by black perimeter. Category A: saturated fatty and sulfonic acids, carbohydrates; category B: N-containing compounds, i.e., peptides; category C: unsaturated aliphatic compounds, aromatic hydrocarbon; category D: phenolic and highly unsaturated compounds; category E: polyphenols and polycyclic aromatics (PCAs) with aliphatic chains; category F: combustion-derived PCAs. Figure S4 Comparisons analysis of the relative abundances of dominant genera estimated by 16S rRNA and metagenomics. Fifteen dominant genera are shown here with each plot showing the comparison for a specific genus. The classified taxa of fungi, viruses, and microbial eukaryotes of metagenomic data are not considered here. Samples from Sanjiang Plain: H02, H18, H22, H47; samples from Lianghu Plain: L01, L07, L20, L28; samples from Taihu Plain: T09, T17, T31, T48; samples from Hani Terrace: Y04, Y24, Y30, Y43. Pearson's correlation coefficients ( $\rho$ ) and statistical significances ( $P$ ) are inscribed in each plot. Figure S5 Flowchart of in-house software for the annotation of DOM molecules. Unequivocal exclusion criteria (elements should follow these rules):  $C > 0$ ;  $N \geq 0$ ;  $H > 0$ ;  $O \geq 0$ ;  $1 \geq P \geq 0$ ;  $1 \geq S \geq 0$ ;  $H \leq C/3$ ;  $H \leq 2C + N + P + 2$ ;  $2(N + H + P) = 0$ ;  $N \leq C$ ;  $O + S \leq C + 2N + 3P$ ;  $O + S \geq P$ . Functional group relationships or "chemical building block" used in elemental formula assignment:  $\text{CH}_4$  - O (0.036385 Da),  $\text{C}_2\text{H}_2$  (26.015650 Da),  $\text{C}_2\text{H}_4$  (28.031300 Da),  $\text{CH}_2$

(14.015650 Da),  $\text{H}_2$  (2.015650 Da),  $\text{H}_2\text{O}$  (18.010565 Da), O (15.994915 Da),  $\text{CO}_2$  (43.989830 Da), NH (15.010899 Da), S (31.972071 Da). (DOCX 1840 kb)

**Additional file 2: Table S1.** Statistics of DOM extraction and annotation of soil samples collected in this study. Peak number a = number of peaks after filtering blank peaks and singletons. Peak number b = number of peaks after filtering blank peaks and those present in less than 10 samples, which is further introduced to annotation procedure. Peak number c = number of peaks assigned to putative formulas. (XLSX 18 kb)

**Additional file 3: Table S2.** Multivariate analysis of bacterial data and drivers using canonical correspondence analysis (CCA) and non-metric multidimensional scaling (NMDS). Constrained (CCA) and unconstrained methods (NMDS and envfit) were used to compare and interpret effects of edaphic, geographical, and DOM factors on the microbial diversity. CCA1, CCA2, MDS1, and MDS2 stand for the angle cosines of variables and the axes. Pvals.1, empirical  $P$  values of fit statistic of environmental variables in CCA; Pvals.2, empirical  $P$  values of fit statistic using envfit to present environmental variables in NMDS;  $r^2$ , goodness of fit statistic. Variables with Pvals.2  $\leq 0.001$  were ticked on the "Fig. 3c" column and shown on Fig. 3c. (XLSX 16 kb)

**Additional file 4: Table S3.** Multivariate analysis of DOM data and drivers using canonical correspondence analysis (CCA) and non-metric multidimensional scaling (NMDS). Constrained (CCA) and unconstrained methods (NMDS and envfit) were used to compare and interpret effects of edaphic, geographical, and bacterial factors on the DOM chemodiversity. CCA1, CCA2, MDS1, and MDS2 stand for the angle cosines of variables and the axes. Pvals.1, empirical  $P$  values of fit statistic of environmental variables in CCA; Pvals.2, empirical  $P$  values of fit statistic using envfit to present environmental variables in NMDS;  $r^2$ , goodness of fit statistic. Variables with Pvals.2  $\leq 0.001$  were ticked on the "Fig. 3d" column and shown on Fig. 3d. Variables used in Fig. 3e were ticked on the "Fig. 3e" column. (XLSX 16 kb)

**Additional file 5:** Supplementary results about consistence between metagenomic data and 16S rRNA data. (DOCX 29 kb)

**Additional file 6: Table S4.** Significant correlations between predicted functions and the compositional changes of DOM. Pearson's correlation coefficient tests were performed between predicted FOAM functions (level 1 and 2) and the compositional changes of DOM along first and second canonical axes obtained from canonical correlation analysis (CCorA). CCorA was performed using the first five principal coordinate (PCo1–5) axes of DOM composition and PCo1–2 axes of FOAM orthologs. It resulted in two canonical axes ( $P < 0.01$ , chi-square test) along which significant correlations were observed between functional potentials and DOM composition ( $P < 0.013$ ). Only functions with significant correlations ( $P < 0.05$ ) were shown here. (XLSX 21 kb)

**Additional file 7: Table S5.** FOAM orthologous groups used for comparing metagenomes with Tax4fun-predicted functions. Pearson's correlation coefficient was used for the correlation estimation. Only orthologous groups significantly correlated with DOM variations were tested here. (XLSX 11 kb)

**Additional file 8: Table S6.** Significant correlations between major methanogen genera in the metagenomic samples and the compositional change of DOM. Pearson's correlation coefficient tests were performed between major taxonomic groups of methanogens in the metagenomic samples and the compositional change of DOM along first canonical axis obtained from canonical correlation analysis (CCorA). CCorA was performed using the first five principal coordinate (PCo1–5) axes of DOM composition and PCo1–2 axes of FOAM orthologs. It resulted in two canonical axes ( $P < 0.01$ , chi-square test) along which significant correlations were observed between functional potentials and DOM composition ( $P < 0.013$ ). Only taxa with significant correlations ( $P < 0.05$ ) were shown here.  $\text{MeNH}_2$  is methylamine. Parentheses means utilized by some, but not all species or strains. (XLSX 12 kb)

**Additional file 9: Table S7.** Pearson's correlation coefficients between Fe(III)-reducing microorganisms in the metagenomic samples and the compositional change of DOM. Information of the type IVa pilA genes was given for their potential extracellular electron transport mechanisms. Pearson's correlation coefficient and  $P$  value were correlated using the abundance of each species and the first canonical axis of DOM variation. (XLSX 16 kb)



**Additional file 10: Table S8.** Soil and site characteristics associated with each of the soil samples collected in this study. MAT = mean annual temperature. MAP = mean annual precipitation, MMT = mean monthly temperature, MMP = mean monthly precipitation (the data we used is the month data we sampled at that site or the mean value of 2-month data if the sampling date was at the beginning or the end of a month), RST = real-time soil temperature, RAT = real-time air temperature, DOC = dissolved organic carbon, Olsen-P = rapid available phosphorus, Ph = rice plant height, FD = flooding depth. MAT, MAP, MMT, and MMP were 22-years observation data obtained from <http://neo.sci.gsfc.nasa.gov>. The total carbon (TC), total nitrogen (TN), and total phosphorus (TP) were on dry basis while the RAP was on wet basis. Field moisture was measured immediately after sample collection. (XLSX 25 kb)

**Additional file 11:** Supplementary methods. (DOCX 32 kb)

**Additional file 12: Table S9.** Basic information of the metagenomic sequencing. (XLSX 11 kb)

**Additional file 13: Table S10.** Statistics information of metagenomic assembly for each metagenomic sample. (XLSX 10 kb)

### Abbreviations

AI: Aromaticity index; CCA: Canonical correspondence analysis; CCorA: Canonical correlation analysis; DIET: Direct interspecies electron transfer; DOM: Dissolved organic matter; EMP: Embden-Meyerhof-Parnos; e-pili: Electrically conductive pili; e-pilin: Electrically conductive pilin; FOAM: Functional Ontology Assignments for Metagenomes; FT-ICR-MS: Fourier transform ion cyclotron resonance mass spectrometry; MAT: Mean annual temperature; MMP: Mean monthly precipitation; NMDS: Non-metric multidimensional scaling; OTU: Operational taxonomic unit; PCo: Principal coordinate; PCoA: Principal coordinate analysis; PERMANOVA: Permutational multivariate analysis of variance; SOM: Soil organic matter; TCA: Tricarboxylic acid; TPM: Transcripts Per Kilobase Million; VAW-UniFrac: Variance adjusted weighted UniFrac

### Acknowledgements

We thank L.J. Zhang, F. Ping, X. Z. Sun, and S. Y. Feng for their assistance in sampling and physiological data analysis.

### Funding

This work was supported by the National Natural Science Foundation of China (41673081, 41373074, 41877346 and 31500409) and Zhejiang Science and Technology Innovation Program (2013C33001). This work was supported in part by the U.S. Department of Energy under Contract DE-AC02-06CH11357.

### Availability of data and materials

The raw Illumine sequence data of 16S rRNA have been deposited in the sequence read archive (SRA accession: SRP5500567–654) at NCBI under Bioproject accession #PRJNA385062. The raw Illumine sequence data of metagenomic data are also available (SRA accession SRX2786101–16) in the NCBI under Bioproject accession #PRJNA385547.

### Authors' contributions

ZZ, JAG, YZ, and HL conceived and designed the project. HL, Hang W, CT, CJ, and ZZ collected and processed all field samples. XP and JC conducted FT-ICR-MS. HL analyzed the data and generated the first draft of the manuscript before JAG, Haitao W, ZZ, Hang W, WA, LC and other co-authors contributed by commenting on and revising it. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>College of Environment and Natural Resource Sciences, Zhejiang University, 866 Yuhangtang Ave, Hangzhou 310058, China. <sup>2</sup>National Plateau Wetlands Research Center, Southwest Forestry University, 300 Bailongsi, Kunming 650224, China. <sup>3</sup>The Microbiome Center, Biosciences Division, Argonne National Laboratory, Lemont, IL 60439, USA. <sup>4</sup>Department of Surgery, University of Chicago, 5640 South Ellis Avenue, Chicago, IL 60637, USA. <sup>5</sup>National Center of Plant Gene Research (Beijing), Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, West Beichen Road, Chaoyang District, Beijing 100101, China. <sup>6</sup>College of Biological and Environmental Engineering, Zhejiang University of Technology, 18 Chaowang Ave, Hangzhou 310014, China. <sup>7</sup>Hangzhou Gusheng Agricultural Technology Company Limited, Chongxian Innovation Industrial Park, Chongxian Ave, Hangzhou 311108, China. <sup>8</sup>Key Lab of Urban Environment and Health, Institute of Urban Environment, Chinese Academy of Sciences, 1799 Jimei Ave, Xiamen 361021, China. <sup>9</sup>China Academy of West Region Development, Zhejiang University, 866 Yuhangtang Ave, Hangzhou 310058, China.

Received: 28 March 2018 Accepted: 20 September 2018

Published online: 19 October 2018

### References

- Kögel-Knabner I, Amelung W, Cao Z, Fiedler S, Frenzel P, Jahn R, et al. Biogeochemistry of paddy soils. *Geoderma*. 2010;157:1–14.
- Bouman M, Lampayan RM, Tuong TP. Water management in irrigated rice: coping with water scarcity. Los Baños: International Rice Research Institute; 2013. p. 3–4.
- Riedel T, Zak D, Biester H, Dittmar T. Iron traps terrestrially derived dissolved organic matter at redox interfaces. *Proc Natl Acad Sci U S A*. 2013;110:10101–5.
- Kellerman AM, Dittmar T, Kothawala DN, Tranvik LJ. Chemodiversity of dissolved organic matter in lakes driven by climate and hydrology. *Nat Commun*. 2014;5:3804.
- Lehmann J, Kleber M. The contentious nature of soil organic matter. *Nature*. 2015;528:60.
- Liesack W, Schnell S, Revsbech NP. Microbiology of flooded rice paddies. *FEMS Microbiol Rev*. 2000;24:625.
- Žifčáková L, Větrovský T, Lombard V, Henrissat B, Howe A, Baldrian P. Feed in summer, rest in winter: microbial carbon utilization in forest topsoil. *Microbiome*. 2017;5:122.
- Kalbitz K, Solinger S, Park JH, Michalzik B, Matzner E. Controls on the dynamics of dissolved organic matter in soils: a review. *Soil Sci*. 2000;165:277–304.
- Hanson CA, Fuhrman JA, Hornerdevine MC, Martiny JB. Beyond biogeographic patterns: processes shaping the microbial landscape. *Nat Rev Microbiol*. 2012;10:497–506.
- Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci U S A*. 2006;103:626–31.
- Zhou J, Ye D, Shen L, Wen C, Yan Q, Ning D, et al. Temperature mediates continental-scale diversity of microbes in forest soils. *Nat Commun*. 2016;7:12083.
- Frank S, Tiemeyer B, Bechtold M, Lücke A, Bol R. Effect of past peat cultivation practices on present dynamics of dissolved organic carbon. *Sci Total Environ*. 2016;574:1243–53.
- Nebbioso A, Piccolo A. Molecular characterization of dissolved organic matter (DOM): a critical review. *Anal Bioanal Chem*. 2013;405:109–24.
- Seifert AG, Roth VN, Dittmar T, Gleixner G, Breuer L, Houska T, et al. Comparing molecular composition of dissolved organic matter in soil and stream water: influence of land use and chemical characteristics. *Sci Total Environ*. 2016;571:142–52.
- Roth VN, Dittmar T, Gaupp R, Gleixner G. The molecular composition of dissolved organic matter in forest soils as a function of pH and temperature. *PLoS One*. 2015;10:e0119188.
- Koch BP, Dittmar T, Witt M, Kattner G. Fundamentals of molecular formula assignment to ultrahigh resolution mass data of natural organic matter. *Anal Chem*. 2007;79:1758–63.
- Kujawinski EB, Vecchio RD, Blough NV, Klein GC, Marshall AG. Probing molecular-level transformations of dissolved organic matter: insights on photochemical degradation and protozoan modification of DOM from

- electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mar Chem.* 2004;92:23–37.
18. Kujawinski EB, Longnecker K, Barott KL, Weber RJM, Soule MCK. Microbial community structure affects marine dissolved organic matter composition. *Front Mar Sci.* 2016;3:45.
  19. Osterholz H, Singer G, Wemheuer B, Daniel R, Simon M, Niggemann J, et al. Deciphering associations between dissolved organic molecules and bacterial communities in a pelagic marine system. *ISME J.* 2016;10:1717.
  20. Gonsior M, Valle J, Schmitt-Kopplin P, Hertkorn N, Bastviken D, Luek J, et al. Chemodiversity of dissolved organic matter in the amazon basin. *Biogeosci Discuss.* 2016;13:1–21.
  21. Stegen JC, Fredrickson JK, Wilkins MJ, Konopka AE, Nelson WC, Arntzen EV, et al. Groundwater-surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. *Nat Commun.* 2016;7:11237.
  22. Ohno T, Sleighter RL, Hatcher PG. Comparative study of organic matter chemical characterization using negative and positive mode electrospray ionization ultrahigh-resolution mass spectrometry. *Anal Bioanal Chem.* 2016;408:2497.
  23. Bao SD. *Agro-chemical analysis of soil.* 3rd ed. Beijing: China Agricultural Press; 2000. p. 71–87.
  24. ABhauer KP, Wemheuer B, Daniel R, Meinicke P. Tax4Fun: predicting functional profiles from metagenomic 16S rRNA data. *Bioinformatics.* 2015;31:2882–4.
  25. Reguera G, McCarthy KD, Mehta T, Nicoll JS, Tuominen MT, Lovley DR. Extracellular electron transfer via microbial nanowires. *Nature.* 2005;71:1098–101.
  26. Summers ZM, Fogarty HE, Leang C, Franks AE, Malvankar NS, Lovley DR. Direct exchange of electrons within aggregates of an evolved syntrophic coculture of anaerobic bacteria. *Science.* 2010;330:1413–5.
  27. Wang H, Holden J, Zhang ZJ, Li M, Li X. Concentration dynamics and biodegradability of dissolved organic matter in wetland soils subjected to experimental warming. *Sci Total Environ.* 2014;470–471:907–16.
  28. Schiff S, Aravena R, Mewhinney E, Elgood R, Warner B, Dillon P, et al. Precambrian shield wetlands: hydrologic control of the sources and export of dissolved organic matter. *Clim Chang.* 1998;40:167–88.
  29. Li X, Sun GX, Chen S, Fang Z, Yuan HY, Shi Q, et al. Molecular chemodiversity of dissolved organic matter in paddy soils. *Environ Sci Technol.* 2018;52:963–71.
  30. Monreal CM, Schulten HR, Kodama H. Age, turnover and molecular diversity of soil organic matter in aggregates of a Gleysol. *Can J Soil Sci.* 1997;77:379–88.
  31. Schmidt SK, Reed SC, Nemergut DR, Grandy AS, Cleveland CC, Weintraub MN, et al. The earliest stages of ecosystem succession in high-elevation (5000 metres above sea level), recently deglaciated soils. *Proc Biol Sci.* 2008;275:2793–802.
  32. Hori T, Müller A, Igarashi Y, Conrad R, Friedrich MW. Identification of iron-reducing microorganisms in anoxic rice paddy soil by <sup>13</sup>C-acetate probing. *ISME J.* 2010;4:267–78.
  33. Lovley DR, Ueki T, Zhang T, Malvankar NS, Shrestha PM, Flanagan KA. *Geobacter*: the microbe electric's physiology, ecology, and practical applications. *Adv Microb Physiol.* 2011;59:1–100.
  34. Kogel-Knabner I. The macromolecular organic composition of plant and microbial residues as inputs to soil organic matter. *Soil Biol Biochem.* 2002;34:139–62.
  35. Golchin A, Clarke P, Oades JM. The heterogeneous nature of microbial products as shown by solid-state <sup>13</sup>C CP/MAS NMR spectroscopy. *Biogeochemistry.* 1996;34:71–97.
  36. Kallenbach CM, Frey SD, Grandy AS. Direct evidence for microbial-derived soil organic matter formation and its ecophysiological controls. *Nat Commun.* 2016;7:13630.
  37. Ng EL, Patti AF, Rose MT, Schefe CR, Wilkinson K, Smernik RJ, et al. Does the chemical nature of soil carbon drive the structure and functioning of soil microbial communities? *Soil Biol Biochem.* 2014;70:54–61.
  38. Conrad R, Klose M. Dynamics of the methanogenic archaeal community in anoxic rice soil upon addition of straw. *Eur J Soil Sci.* 2006;57:476–84.
  39. Holmes DE, Dang Y, Walker DJ, Lovley DR. The electrically conductive pili of *Geobacter* species are a recently evolved feature for extracellular electron transfer. *Microb Genom.* 2016;2:e000072.
  40. Holmes DE, Shrestha PM, Walker DJ, Dang Y, Nevin KP, Woodard TL, et al. Metatranscriptomic evidence for direct interspecies electron transfer between *Geobacter* and *Methanotrix* species in methanogenic rice paddy soils. *Appl Environ Microbiol.* 2017;83:AEM.00223–17.
  41. Lovley D, Woodard JC, Philips EJP, Bluntharris EL, Coates JD. Humic substances as electron acceptors for microbial respiration. *Nature.* 1996;382:445–8.
  42. Tan Y, Adhikari RY, Malvankar NS, Ward JE, Nevin KP, Woodard TL, et al. The low conductivity of *geobacter uraniireducens* pili suggests a diversity of extracellular electron transfer mechanisms in the genus *Geobacter*. *Front Microbiol.* 2016;7:1.
  43. Ratering S, Schnell S. Localization of iron-reducing activity in paddy soil by profile studies. *Biogeochemistry.* 1998;3:341–57.
  44. Jahn R, Blume HP, Asio VB, Spaargaren O, Schad P. Guidelines for soil description. 4th ed. Rome: Food and Agriculture Organization of the United Nations; 2006. p. 67–77.
  45. Šantlémek T, Kai F, Dittmar T, Hansen BM, Thyrhaug R, Nielsen NW, et al. Hailstones: a window into the microbial and chemical inventory of a storm cloud. *PLoS One.* 2013;8:e53550.
  46. Baughn AD, Malamy MH. The strict anaerobe *Bacteroides fragilis* grows in and benefits from nanomolar concentrations of oxygen. *Nature.* 2004;427:441–4.
  47. Pitcher RS, Brittain T, Watmough NJ. Cytochrome cbb(3) oxidase and bacterial microaerobic metabolism. *Biochem Soc Trans.* 2002;30:653–8.
  48. Dong WY, Zhang XY, Dai XQ, Fu XL, Yang FT, Liu XY, et al. Changes in soil microbial community composition in response to fertilization of paddy soils in subtropical China. *Appl Soil Ecol.* 2014;84:140–7.
  49. Johnsen K, Jacobsen CS, Torsvik V, Sorenson AJ. Pesticide effects of bacterial diversity in agricultural soils—a review. *Biol Fert Soils.* 2001;33:443–53.
  50. Breidenbach B, Conrad R. Seasonal dynamics of bacterial and archaeal methanogenic communities in flooded rice fields and effect of drainage. *Front Microbiol.* 2014;5:752.
  51. Sun Y, Huang S, Yu X, Zhang W. Differences in fertilization impacts on organic carbon content and stability in a paddy and an upland soil in subtropical China. *Plant Soil.* 2015;397:1–12.
  52. Said-Pullicino D, Miniotti EF, Sodano M, Bertora C, Lerda C, Chiaradia EA, et al. Linking dissolved organic carbon cycling to organic carbon fluxes in rice paddies under different water management practices. *Plant Soil.* 2016;401:273–90.
  53. Noll M, Matthies D, Frenzel P, Derakshani M, Liesack W. Succession of bacterial community structure and diversity in a paddy soil oxygen gradient. *Environ Microbiol.* 2005;7:382–95.
  54. Begg CBM, Kirk GJD, Mackenzie AF, Neue H. Root-induced iron oxidation and pH changes in the lowland rice rhizosphere. *New Phytol.* 2010;128:469–77.
  55. Tyagi L, Verma A, Singh SN. Investigation on temporal variation in methane emission from different rice cultivars under the influence of weeds. *Environ Monit Assess.* 2004;93:91–101.
  56. Lynch JM, Whipps JM. Substrate flow in the rhizosphere. *Plant Soil.* 1990;129:1–10.
  57. Neue HU, Wassmann R, Lantin RS, Ma CRA, Aduna JB, Javellana AM. Factors affecting methane emission from rice fields. *Atmos Environ.* 1996;30:1751–4.
  58. Sheng R, Qin H, O'Donnell AG, Huang S, Wu J, Wei W. Bacterial succession in paddy soils derived from different parent materials. *J Soils Sediments.* 2015;15:982–92.
  59. Sun W, Xiao E, Pu Z, Krumins V, Dong Y, Li B, et al. Paddy soil microbial communities driven by environment- and microbe-microbe interactions: a case study of elevation-resolved microbial communities in a rice terrace. *Sci Total Environ.* 2018;612:884–93.
  60. Rokunuzzaman M, Ueda Y, Chen L, Tanaka S, Ohnishi K. Effects of land use changes from paddy fields on soil bacterial communities in a hilly and mountainous area. *Microbes Environ.* 2016;31:160–4.
  61. Fierer N. Microbial biogeography: patterns in microbial diversity across space and time. In: Zengler K, editor. *Accessing uncultivated microorganisms: from the environment to organisms and genomes and Back.* Washington DC: ASM Press; 2008. p. 95–115.
  62. Koch BP, Dittmar T. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Commun Mass Spectrom.* 2010;20:926–32.
  63. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci U S A.* 2011;108:4516–22.
  64. Jing X, Sanders NJ, Shi Y, Chu H, Classen AT, Zhao K, et al. The links between ecosystem multifunctionality and above- and belowground biodiversity are mediated by climate. *Nat Commun.* 2015;6:8159.
  65. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods.* 2010;7:335–6.
  66. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41:590–6.



67. Kopylova E. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*. 2012;28:3211–7.
68. Guio L. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol*. 2012;8:e1002358.
69. Prestat E, David MM, Hultman J, Taş N, Lamendella R, Dvornik J, et al. FOAM (Functional Ontology Assignments for Metagenomes): a hidden Markov model (HMM) database with environmental focus. *Nucleic Acids Res*. 2014;42:e145.
70. Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat Commun*. 2016;7:11257.
71. Li D, Luo R, Liu CM, Leung CM, Ting HF, Sadakane K, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods*. 2016;102:3–11.
72. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. 2010;11:119.
73. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*. 2016;44:D279.
74. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*. 2000;27:29–34.
75. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014;30:2068.
76. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat Commun*. 2016;7:13219.
77. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42:222–30.
78. Haft DH, Selengut JD, Richter RA, Harkins D, Basu MK, Beck E. TIGRFAMs and genome properties in 2013. *Nucleic Acids Res*. 2013;41:387–95.
79. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res*. 2013;41:e121.
80. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2014;12:59.
81. Langmead B. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 2012;9:357.
82. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*. 2009;25:2078–9.
83. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31:166.
84. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*. 2010;26:493–500.
85. Kindt R, Coe R. Tree diversity analysis. A manual and software for common statistical methods of ecological and biodiversity studies. *J Am Vet Med Assoc*. 2005;235:68–374.
86. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, et al. Vegan: community ecology package. R package version 2.3-0; 2015.
87. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, et al. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics*. 2012;28:2106–13.
88. Venables W, Ripley B. Modern applied statistics with S-PLUS. *Stat Comput*. 2002;52:704–5.
89. Cheng SH, Li J. Modern Chinese rice. 3rd ed. Beijing: JinDun Press; 2007. p. 54–7.
90. Monfreda C, Ramankutty N, Foley JA. Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Glob Biogeochem Cycles*. 2008;22:GB1022.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

