PEDOSPHERE（土壤圈英文版）

Pedosphere ISSN 1002-0160/CN 32-1315/P doi:10.1016/S1002-0160(21)60016-4

Running Title: DATABASES FOR SOIL METAPROTEOMICS

Protein sequence databases generated from metagenomics and public database produced similar soil metaproteomic results of microbial taxonomic and functional changes

Yi XIONG[1,2], Lu ZHENG[1], Xiangxiang MENG[1,2], Renfang SHEN[1,2], Ping Lan[1,2*]

[1]*State Key Laboratory of Soil and Sustainable Agriculture, Institute of Soil Science, Chinese Academy of Sciences, Nanjing 210008 (China)*

[2]*University of Chinese Academy of Sciences, Beijing 100049 (China)*

ABSTRACT

Soil metaproteomics has excellent potential to elucidate how soil microbial communities could change structurally and functionally in response to environmental alterations. However, soil metaproteomics is hindered by some challenges and gaps. Soil microbial communities possess extremely complex microbial composition, including many uncultured microorganisms without whole-genome sequencing. Thus, how to select a suitable protein sequence database remains challenging in soil metaproteomics. The Public database and Meta database were constructed using protein sequences from public databases and metagenomics, respectively. Here we comprehensively analyzed and compared the soil metaproteomic results using these two kinds of protein sequence databases for protein identification based on a published soil metaproteomic raw data. The results demonstrated that much more proteins, higher sequence coverages, and even more microbial species and functional annotations could be identified using the Meta database compared with those identified using the Public database. These findings indicated that the Meta database was more specific as a protein sequence database. However, the follow-up in-depth metaproteomic analyses exhibited similar main results regardless of the databases used. The microbial community composition at the genus level was similar, especially these species annotations with high Peptide-spectrum match and high abundance. The functional analyses in response to stress, such as the gene ontology enrichment about biological progress and molecular function, and key functional microorganisms, were also similar regardless of databases. Our analysis revealed that the Public database could also meet the demand to explore the functional responses of microbial proteins to some extent. This study provides valuable insights into the choice of protein sequence databases and their impact on the subsequent bioinformatics analysis in soil metaproteomic research for better experimental design for different purposes.

Key Words: bioinformatics, microbial communities, protein sequence database, soil metagenomics, soil metaproteomics

INTRODUCTION

Soil is a dynamic system with complex and heterogeneous physical, chemical, and biological interactions. Soil microorganisms play critical roles in ecosystems and are heavily involved in a large number of biogeochemical processes, including nutrient acquisition, recycling of elements (carbon, nitrogen or phosphorus (P)), and organic matter transformation (van der Heijden *et al.*, 2008; Bastida *et al.*, 2009). Recently, several molecular techniques have been applied to explore soil microbial communities and their functions, mainly metagenomics (Daniel, 2005), metatranscriptomics (Carvalhais *et al.*, 2012), and metaproteomics (Keiblinger *et al.*, 2016). The metagenomics was widely

[*]Corresponding author. E-mail: plan@issas.ac.cn.

used and provided compositions and diversities of soil microbial communities. However, the metagenome information represents only the prediction of community functional potential. It is necessary to measure the actual expression of genes at the mRNA and protein levels. Correspondingly, metatranscriptomics and metaproteomics could display the soil microbial responses to environmental stimuli transcriptionally and post-transcriptionally, respectively. Since proteins carry out most functions in cells, the investigation of metaproteomics, the protein expression in soil microbiota, can expand the understanding of why and how the microbiome changed the community structure to adapt to the environmental stimuli. Furthermore, we can elucidate the roles of soil microbe in the uptake of nutrients for plant and biogeochemical cycles of elements in the soil at the protein level. The metaproteomics could integrate the community structure, functions, and regulation of soil microorganisms (Nannipieri, 2006; Renella *et al.*, 2014; Jansson and Hofmockel, 2018).

Due to the extremely spatial-temporal complexity and heterogeneity of the soil matrix, pH, mineral nutrition, organic compounds, and microorganisms, the development of proteomics in soil lags far behind proteomics in other fields (Maron *et al.*, 2007; Chapman and Bellgard, 2014). Among these challenges, there are two prior problems to be solved. First, the complicated compositions of soil, such as humic compounds and clay, significantly affected the extraction efficiency of soil protein (Chourey *et al.*, 2010; Burns *et al.*, 2013; Tartaglia *et al.*, 2020). The low extraction efficiency and purity of proteins seriously interfered with the downstream protein identification of mass spectrometry. However, several effective methods for soil protein extraction were developed in the past decade, such as the cell lysis step using the alkaline detergent buffer and purification step using trichloroacetic acid (TCA)/acetone solution (Murase *et al.*, 2003; Benndorf *et al.*, 2007; Chourey *et al.*, 2010; Wang *et al.*, 2011; Johnson-Rollings *et al.*, 2014; Xiong *et al.*, 2016; Kunath *et al.*, 2019). The removal of humic substances can conduct by phenol extraction and differential solubility at low pH at the tryptic digestion step (Qian and Hettich, 2017). Besides, protocols like filter-aided sample preparation (FASP) (Wiśniewski *et al.*, 2009) could also help to remove other contaminants (Zampieri *et al.*, 2016). All these processes significantly increased the quality of soil protein extraction, meeting the requirement for analysis in a high-resolution mass spectrometer. Second, protein identification and bioinformatical evaluation are also challenging, particularly construction of databases, grouping of redundant proteins, and taxonomic and functional annotation (Heyer *et al.*, 2017). Different samples contain many soil microbial proteins with high amino acid similarity due to minor strain variations, horizontal gene transfer, or recurring functional domain. The identical peptides belonging to homologous proteins cause redundant protein identification, making it hard to obtain precise functional interpretation. Also, the proteins from the closely related species usually have high amino acid similarity, making it difficult to get accurate taxonomic annotations (Nesvizhskii and Aebersold, 2005; Heyer *et al.*, 2017; Kunath *et al.*, 2019).

Unlike sequencing techniques used in genomics and transcriptomics, mass spectrometry in shotgun proteomics is commonly conducted depending on the use of database search engines nowadays (Verheggen *et al.*, 2020). In this approach, the MS or MS/MS spectra should be matched to theoretical peptides from the in-silico digested protein sequence database for protein identification. It was well known that soil microbial communities contain thousands of different microbial species, and their composition varies considerably among soil samples (Torsvik and Øvreås, 2002; Muth *et al.*, 2015). Thus, the construction of well-suited protein sequence database becomes more challenging in soil metaproteomics, and needs to be first solved (Yadav *et al.*, 2012). In some proteogenomic researches about artificially assembled microbial communities (Tanca *et al.*, 2013; Kleiner *et al.*, 2017) and low-complexity microbial communities, such as human oral (Grassl *et al.*, 2016) and gut (Muth *et al.*, 2015; Brown *et al.*, 2018), the specificity and sensitivity of reference protein sequence databases obtained by 16S-rRNA profiling, curated public databases and metagenomic sequencing have been meticulously validated. Only part of peptides is identified in common by these three kinds of databases (Tanca *et al.*, 2016). The microbial species and abundance obtained by amplicon-based sequencing, metagenomic sequencing, and metaproteomic mass-spectrometry cannot correlate well (Kleiner *et al.*, 2017). Taxonomic and functional results were strongly database-dependent (Tanca *et al.*, 2016). The short-read-length 16S-rRNA profiling limits the accuracy of detecting microbial species at a deep level of reference protein sequence databases built by this technology (Johnson *et al.*, 2019). The binning and assembly in metagenomic workflow and complete microbial genomes in public databases relieve this defect to a certain extend. However, the incompleteness of reference protein sequence databases built by metagenomic sequencing and distantly related taxonomic composition of curated public databases affect the quantification of proteins and organisms in proteomic research about high-complexity communities such as soil. There have not been systematic studies focused on the selection of protein sequence databases in soil metaproteomics yet.

Recently, the protein sequence databases used in soil metaproteomics are mainly classified into two types, the protein sequences from public database (Wang *et al.*, 2011; Bastida *et al.*, 2014; Johnson-Rollings *et al.*, 2014; Bastida *et al.*, 2016) and the protein sequences predicted by the assembly of metagenomic sequencing results (Butterfield *et al.*, 2016; Yao *et al.*, 2018). Public databases were widely used in previous researches because of the high expense of metagenomics. In the metaproteomics analysis of ratoon sugarcane rhizospheric soil, only 143 protein spots were identified via 2-DE and MALDI TOF-TOF MS analyses using a public database from NCBI (Lin *et al.*, 2013). Later, liquid chromatography with tandem mass spectrometry (LC-MS/MS) significantly increased the number of identified proteins. Using a public database derived from UniRef100, about 1048 proteins were identified in three soil samples with different organic matter contents using Chourey's method of protein extraction and LC-MS/MS analyses (Chourey *et al.*, 2010; Bastida *et al.*, 2014). Likewise, using a public database derived from the NCBI, about 3 082 non-redundant (nr) proteins were identified in soils from a dryland region. Recently, proteogenomics (metagenomics and metaproteomics) is being applied in analyses of soil microbial communities and functions. Subsequently, more and more studies used the assembly of metagenomics as the protein sequence database. The soil microbial communities in the sub-root zone were systematically analyzed (Butterfield *et al.*, 2016). The protein sequences from metagenomics data of four soil samples were used as the database for metaproteomics, including 3 408 250 predicted protein sequences. A total of 6 835 proteins were identified based on 28 782 distinct peptides. The methylotrophy proteins were among the most abundant proteins. In researching the systemic impact of P availability in the tropical forest, the P-deficient and P-rich soils that endured a 17-year fertilization experiment were analyzed using proteogenomics (Yao *et al.*, 2018). An average of 7 114 proteins was identified per soil sample, using the protein sequences predicted from the metagenome of this soil sample. Proteins for the degradation of P-containing nucleic acids and phospholipids were significantly enhanced in the P-deficient soils.

Despite the rapid development of high-resolution mass spectrometry in recent years, it is still a lack of efficient and standardized data analysis processes for complex samples such as soil metaproteomics. In analyzing uncharacterized microbial communities, different protein sequence databases indeed significantly affected the results of protein identification (Muth *et al.*, 2015; Tanca *et al.*, 2016; Xiao *et al.*, 2018). In fact, soil microbial communities possess extremely complex microbial composition, which greatly increased the difficulty of protein identification in soil metaproteomics. However, no comprehensive information is available about the influences of protein sequence databases on protein identification in soil metaproteomics so far. Here, we systematically demonstrated and compared the typical strategies of construction of protein sequence databases in soil metaproteomics studies and the downstream bioinformatics analyses. The differences in the number of identified proteins, taxonomic and functional annotations, Gene Ontology (GO) enrichment, and phylogenetic relationships were all displayed. We try to offer a more comprehensive understanding of how protein sequence databases influence soil metaproteomic results, helping the future experimental design of soil metaproteomics.

MATERIALS AND METHODS

Data collection

The soil metaproteomics data were obtained from P-rich and P-deficient soils in a 17-year fertilization experiment in the tropical forest by shotgun proteomics measurements (Yao *et al.*, 2018). The soils of each treatment were collected from two plots with two technical replicates, P-rich soils from plots 1 and 30 and P-deficient soils from plots 6 and 36. The MS/MS raw files and the fasta files of predicted protein sequences were downloaded from ProteomeXchange databases (PXD005910).

Generation of protein sequence databases

For assessing the influence of protein sequence databases on protein identification and further bioinformatics analyses, two typical databases were constructed, named the Meta database and the Public database (Fig. 1). The Meta database was obtained from the metagenomic datasets of the corresponding soil samples described by (Yao *et al.*, 2018) with some modification. The predicted protein sequences in the four fasta files obtained from four samples were merged into one file. In this file, duplicated sequences with the amino acid identity of 100% were removed by SeqKit (version 0.12.1) (Shen *et al.*, 2016), and only one copy of the duplicated sequences was kept. The non-redundant fasta file was considered as the Meta database. Another kind of protein sequence database was derived from the Public database, generated as follows. All related protein sequences were selected from the NCBI protein database with the entries contained the keywords, soil or rhizosphere, and the organisms limited to fungi, protists, bacteria, archaea, or viruses (downloaded on October 30, 2018). The selection was done by R packages 'Rentrez' (version 1.2.2) (Winter, 2017) and 'Parallel' (version 4.0.2). Analysis using R packages in this study is performed in the R software environment (version 4.0.2) in Ubuntu 18.04.5 LTS (Bionic Beaver) developed by R Core Team (2019). According to the range of sequence length in the Public database, the protein sequences with a length shorter than 10 amino acids or longer than 2 700 amino acids were removed by SeqKit (version 0.12.1) (Shen *et al.*, 2016). For the large number of protein sequences with high sequence similarity from NCBI, the redundancy was reduced with 90% similarity by CD-HIT (version 4.8.1) (Fu *et al.*, 2012). The generated fasta files were used for protein identification. The processed data and codes used in this study are all available in the GitHub repository for share (Chiapello *et al.*, 2020), https://github.com/xyz1396/Meta-proteomics-analysis-pipeline-based-on-Proteome-Discovery-output.

Fig. 1 Schematic illustration of the soil metaproteomics analyses using two strategies of protein sequence database construction. In this analysis, we skipped the assembly and prediction steps, and downloaded the predicted protein sequences in the archive.

Protein identification using the Meta database and Public database

The MS/MS raw files were searched using Proteome Discoverer (version 2.2, Thermo Scientific) against the Meta database and Public database, respectively. The contaminated protein sequence database was from cRAP (The Global Proteome Machine, http://www.thegpm.org/crap/). The protein quantification strategy was modified from the Label-free quantification template in Proteome Discoverer. The search engine was Sequest HT (built in Proteome Discoverer), and the confidence of the identified proteins was measured by the decoy search strategy. The mass tolerances of precursor ions and product ions were 10 ppm and 0.02 Da, respectively. Precursor's mass range was from 350 Da to 5 000 Da, and the S/N threshold was 1.5. Trypsin was set as the proteolytic enzyme, and two missed cleavages at most were allowed. Cysteine carbamidomethylation was set as static modification, and methionine oxidation was set as a dynamic modification. Protein abundances were calculated by the intensity of the precursor, and the identified proteins were grouped by the maximum parsimony principle. Contaminated proteins and proteins with a false discovery rate (FDR) greater than 0.01 were removed by R package 'dplyr' (version 1.0.1) (Wickham *et al.*, 2015) to obtain protein groups. Proteome Discoverer selected the representative protein with the largest value in the "Protein Unique Peptides column" and the smallest value in the "Coverage column" (the longest protein) from each protein group. The representative proteins with high confidence were used for downstream analysis.

Bioinformatics analysis

The protein abundance analysis and visualization were accomplished by R packages 'ggplot2', (version 3.3.2) (Wickham, 2011), 'ggforce' (version 0.3.2) (Pedersen, 2016), 'limma' (version 3.44.3) (Smyth, 2005), 'DESeq2' (version 1.28.1) (Love *et al.*, 2014), 'VennDiagram' (version 1.6.20) (Chen, 2016), 'pheatmap' (version 1.0.12) (Kolde, 2015), and 'UpSetR' (version 1.4.0) (Conway *et al.*, 2017). Further, the matched protein sequences were extracted by R package 'Biostrings' (version 0.34.0) (Pagès *et al.*, 2019) and aligned against the NCBI nr database by the locally installed BLAST (version 2.10.1+) (Madden, 2013). The alignment results were parsed by Blast2Go (version 5.2.5) for GO annotations (Conesa *et al.*, 2005). The KEGG annotations were obtained from the ghostKOALA website (Kanehisa *et al.*, 2016). The taxonomic interpretation of the proteins identified using the Public database was obtained from the NCBI by taxize (version 0.9.98) (Chamberlain and Szocs, 2013). Taxonomic interpretations of proteins identified by both databases were also obtained from the ghostKOALA website for more accurate interpretation at the genus level. The visualizations of annotation results were accomplished by R packages 'stringr' (version 1.4.0) (Wickham, 2017), 'tidyr' (version 1.1.1) (Wickham and Henry, 2019), 'clusterProfiler' (version 3.16.0) (Yu *et al.*, 2012), and 'GO.db' (version 3.11.4) (Carlson, 2019).

Protein sequences of phosphatases and phospholipases were extracted to build the phylogenetic trees by MEGA X (version 10.0.5) (Kumar *et al.*, 2018), and the conserved motifs were obtained by Meme (version 5.1.1) (Bailey *et al.*, 2009). The motif sequences were annotated by Web CD-Search Tool (Marchler-Bauer and Bryant, 2004). The visualizations of the phylogenetic tree and motif were accomplished by TBtools (version 1.068) (Chen *et al.*, 2018).

RESULTS AND DISCUSSION

More proteins and higher sequence coverages were obtained using the Meta database

Two typical protein sequence databases, the Meta database and Public database, were generated (Table 1). Taken as a whole, the total numbers of proteins and total lengths of proteins were very similar between the two databases, as well as the protein length distributions. In both databases, most proteins (about 93 -- 95%) had 51 -- 800 amino acids. Moreover, the number of proteins was the highest in the length range of 201 − 400 amino acids, and gradually decreased when the lengths were higher than 400 or lower than 201. The average protein length (263 amino acids) based on the Meta database was shorter than that for the Public database (316 amino acids). Overall, the two databases were nearly the same size, which avoided the retrieval differences in search space.

TABLE I The detailed information of two protein sequence databases.

Protein length (amino acids)	Number of proteins		Proportion (%)	
	Meta database	Public database	Meta database	Public database
0 -- 50	439 980	95 298	4.66	1.08
51 -- 100	1 388 280	927 419	14.70	10.54
101 -- 200	2 530 885	2 087 094	26.80	23.72
201 -- 400	3 346 212	3 458 216	35.43	39.31
401 -- 800	1 537 865	1 872 361	16.28	21.28
801 -- 1 600	195 288	330 747	2.07	3.76
1 601 -- 3 200	5 594	26 573	0.059	0.30
3 201 -- 12 200	334	0	0.004	0.00
Total number of proteins	9 444 438	8 797 708		
Total length of proteins	2 480 437 575	2 777 724 902		
Average protein length	263	316		

The same MS/MS raw files were searched against the two databases, respectively. The different proteins in one protein group shared the same peptides, and the representative protein of each identified protein group was used for further bioinformatic analysis. A total of 170 565 proteins in 18 947 protein groups with high confidences were identified using the Meta database, while only 20 779 proteins in 4 320 protein groups were identified using the Public database (Fig. 2a and Table S1). The numbers of protein groups and proteins using the Meta database were about 4.4-folds and 8.2-folds higher than those identified using the Public database. This result indicated that the Meta database was more specific for protein identification. The protein length distribution of the representative proteins using two databases was displayed in Fig. 2b. Although the numbers of identified proteins in every length were much higher using the Meta database, the proportion in each protein length range was similar between two databases (Table S2). The highest proportion (about 45%) of the identified proteins was in the length range of 201 -- 400 amino acids. The protein coverage distribution of the representative proteins exhibited significant differences using two databases (Fig. 2c). Using the Public database, about 75% of the proteins had sequence coverages of 0-10%, while the percentages of the identified proteins with high sequence coverages (higher than 20%) were extremely low, only about 6% (Table S2). Using the Meta database, the sequence coverages were more evenly distributed from 0 to 100% (Fig. 2c). Notably, the percentages of proteins with sequence coverages higher than 30% were up to 14%, which was only 1% using the Public database. In short, higher proportion of proteins with high sequence coverages obtained using the Meta database indicated more specificity of the Meta database as protein sequence database for protein identification.

Fig. 2 The total number of protein groups and proteins identified using the two protein sequences databases (a). The distribution of protein lengths (b) and sequence coverages (c) of the representative proteins in each identified protein group using two databases.

The identified proteins using the two databases were aligned with each other by BLAST with a cutoff of E-value of 10^{-6}. The distributions of the amino acid identity, alignment length, bit score and -log$_{10}$($^{\text{E-value}}$) of the two identified protein sets showed extremely significant differences by Wilcoxon tests. It was proved that neither set of the identified proteins was sub-sequence of the other one (Fig. S1). About 86% of proteins identified by the Meta database could be matched to the proteins identified by the Public database, and the median value of identity was 64.95%. In comparison, most proteins (98%) identified by the Public database could be matched to the proteins identified by the Meta database, and the median value of identity (75.25%) was higher. It also indicated that the proteins identified by the Meta database were more specific than those identified by the Public database.

The repeatability of the identified proteins for each treatment using the two databases was displayed by UpSet plots and abundance heatmaps. Regardless of the used databases, majority of the identified proteins were shared between the four samples within the same treatment, especially the two technical replicates from one plot (Fig. S2). The abundance heatmaps demonstrated the low within-plot variation and high cross-plot variation regardless of the used database (Fig. S3). The samples from two plots of the P-rich soils even did not cluster together regardless of used databases.

More unique proteins and differentially expressed proteins in response to P deficiency were identified using the Meta database

To fully assess the impact of different reference databases on the final soil metaproteomic results, the identified proteins between P-deficient and P-rich soils were compared (Fig. 3). Using the Meta database, 2 069 and 2 675 unique proteins were identified only in the P-rich soil and P-deficient soil, respectively. Using the Public database, the unique proteins identified only in the P-rich soil and P-deficient soil were 450 and 559, respectively. Undoubtedly, the numbers of unique proteins were much higher using the Meta database. However, the percentages of unique proteins were similar, about 25.04% (4 744 proteins) using the Meta database and 23.36% (1 009 proteins) using the Public database in all identified proteins.

Fig. 3 Different identified proteins between the P-deficient and P-rich soils using the two protein sequence databases, the Meta database (a) and the Public database (b).

The differentially expressed proteins (DEPs) in response to P deficiency were classified based on criteria that fold changes of protein abundances were above 2 or under 1/2 between the two kinds of soils (P-deficient/P-rich), and the adjusted p values were under 0.05 (Table S3). Among the proteins identified using the Meta database without missing value of fold change and adjusted P-value, 1 319 and 1 539 proteins were up-regulated and down-regulated under P-deficient treatment, respectively (Fig. 4a). While for the proteins identified using the Public database, only 426 and 488 proteins were up-regulated and down-regulated, respectively (Fig. 4b). The DEPs identified by the Meta database were about 3.13-folds more than those identified by the Public database. It is noted that the proportion of DEPs identified using the Public database (25.00%) was higher than that using the Meta database (18.51%). Thus, more unique proteins and DEPs could be identified using the Meta database.

Fig. 4 Volcano plots of proteins identified in the P-deficient and P-rich soils using the Meta database (a) and Public database (b). The X-axis is the binary logarithms of the fold changes for the abundance of the same protein between the P-deficient and P-rich soils, and the Y-axis is the negative base 10 logarithms of adjusted p-value corresponding to the fold change. The vertical dotted line on the right indicates the fold change is 2, and the vertical dotted line on the left indicates that the fold change is 1/2. The horizontal dotted line indicates that the adjusted p-value is 0.05. Proteins with a fold change above 2 and adjusted *p* - value under 0.05 are proteins with up-regulated abundances. Proteins with fold change under 1/2 and adjusted *p* -value under 0.05 are proteins with down-regulated abundances.

More microbial species and functional annotations were obtained using the Meta database

Among the proteins identified using the Meta database, 18 386 proteins (98.39%) were from bacteria, while only 219 and 82 proteins were from archaea and fungi, respectively. Similarly, 4 191 proteins (97.85%) identified using the Public database were from bacteria, and only 40 and 52 proteins were from archaea and fungi, respectively (Table S4). Thus, regardless of the database used, majority of the identified proteins were from bacteria. A few proteins identified were from fungi and archaea, probably due to the low percentages of proteins from fungi and archaea in two databases. Protein sequences from bacteria (99.76% in the Meta database and 92.57% in the Public database) are the dominant component in both databases.

The differences of microbial taxonomies for identified proteins and their Peptide-Spectrum Match (PSMs) obtained using the two databases were displayed at the genus level (Fig. 5 and Table S4). The number of genera identified using the Meta database was 854 genera in total using the annotation from ghostKOALA, and was almost 1.5-folds more than identified using the Public database (579 genera). A total of 327 genera could only be identified using the Meta database (Fig. 5a). Majority of the genera (91%) identified by the Public database could also be identified by the Meta database. Similarly, much more genera could be identified using the Meta database for each kind of soil. Thus, the Meta database could identify more microbial species than the Public database, which was beneficial to study the biodiversity of the microbial community. The PSMs for the genera identified by the Meta database were much higher than those identified by the Public database. Notably, the common genera identified by the Meta database had a much higher median (220) than others. It should be noted that the distributions of PSMs were significantly different between the common genera and unique genera identified by the two databases (Fig. 5c, d). The medians of PSMs for unique genera were much lower. The median was 12 for the PSMs identified by the Meta database, and 8 for those identified by the Public database. It was suggested that credible shared genera with high PSMs could be identified by both databases, and more specific genera with low PSMs could be identified by the Meta database.

Fig. 5 Differences of microbial taxonomies of identified proteins interpreted by ghostKOALA website at the genus level, and their corresponding PSMs using the two different databases. Venn diagram of the microbial taxonomies in total (a) and P-rich and P-deficient soils using the Meta database (M) and Public database (P) (b). The distribution of PSMs of genera identified by the Meta database (c) and Public database (d). Common genera mean the genera could be identified by both databases, and unique genera mean the genera could only be identified by the Meta database or Public database. The asterisks on the top mean the extremely significant difference ($p < 0.0001$) between the two combinations obtained by Wilcoxon tests.

The abundances of the microbial taxonomies for identified proteins at the genus level using the Meta database were higher than those identified using the Public database. However, the microbial taxonomies and their proportions of the most abundant microbial taxonomies using the two databases were similar (Fig. 6a and Table S5). Among the top ten most abundant microbial taxonomies, nine of them were the same. Rhodoplanes and Bradyrhizobium were the most abundant microorganisms identified in the P-rich and P-deficient soils, regardless of the used database and the treatment of P fertilizer in soils. Besides, we analyzed the correlations of microbial abundance at the genus level identified by the two databases in P-rich and P-deficient soils (Fig. 6b, c). They both exhibited significantly high positive correlations (r = 0.92 and r = 0.89), indicating both databases could efficiently interpret the microbial taxonomies for identified proteins. Thus, the used protein sequence database did not change the results of the elemental composition of the microbial community at the genus level.

Fig. 6 The proportions (a) of the most abundant ten species of proteins identified in the P-rich and P-deficient soils using the two databases at the genus level. The correlation of the microbial abundance at the genus level identified using the two databases in P-rich soils (b) and P-deficient soils (c).

Similar results of GO enrichment about biological progress and molecular function obtained using the two databases

The identified proteins were matched to sequences in the NCBI nr database by BLAST alignment for GO and KEGG annotations (Fig. 7 and Table S6). Using the Meta database, 18 706 proteins could be matched to sequences in the NCBI nr database by BLAST, 13 023, and 7 813 proteins could obtain GO annotations and KEGG annotations, respectively. Meanwhile, all proteins could be matched to sequences in the NCBI nr database by BLAST alignment, 3 184, and 2 438 proteins can obtain GO annotations and KEGG annotations using Public database. These annotations were only 23%, 41%, and 31% of the annotations obtained using the Meta database. However, compared with the total identified protein groups, a higher proportion of proteins could get GO and KEGG annotations using the Public database, probably due to the more detailed annotation of proteins from the Public database. However, the distributions of functions of the identified proteins on a comparably deep level (level 3 of KEGG) using the two databases were similar (Fig. S4 a). Correlations of KEGG annotations at the level 2 and level 3 identified by the two databases are also extremely high in P-rich soils (Fig. S4 b and Fig. S5 a) and P-deficient soils (Fig. S4 c and Fig. S5 b). This indicates that the two databases produced similar protein function analyses.

Fig. 7 The number of the blast alignments, GO and KEGG annotations of the proteins identified using the two databases.

The GO enrichment analysis about biological progress (BP) and molecular function (MF) for the up-regulated proteins in response to P deficiency was demonstrated using proteins identified by the two databases (Fig. 8 and Table S7). Despite the considerable variances in the amounts of proteins identified using the two databases, the GO enrichment results in biological progress and molecular function of the up-regulated proteins were remarkably similar. Furthermore, the proteins using the two databases were both mainly involved in various transport processes, including the phosphate ion transport, inorganic anion transport, ion transmembrane transport, phosphate ion transmembrane transport (Fig. 8a). These results were also similar to those results from the original paper (Yao *et al.*, 2018).

Fig. 8 The GO enrichment about the biological progress (a) and molecular function (b) based on the up-regulated proteins in response to P deficiency identified using the two databases. The adjusted *P*-values of GO terms in the plot are under 0.05, and the size of the points indicates the number of proteins. The ratio means the proportion of the up-regulated proteins in all identified proteins with the same GO term.

Nevertheless, there were still some subtle differences. The processes related to ion transport, anion transport, and transmembrane transport were only enriched using the Public database. Nonetheless, these did not affect the main results that many transporting activities related to P acquisition were greatly enhanced under P-deficiency stress. However, the much deeper measurement using the Meta database provided higher confidence than the Public database in the membership population of the GO enrichments.

GO enrichment in MF also got similar results using the two databases (Fig. 8b). The enriched up-regulated proteins' MF was mainly related to phosphate ion binding, phosphoric ester hydrolase activity, phospholipase activity, and the acid phosphatase activity. These proteins were involved in P acquisition and cycling, an adaptation to P stress. Also, some differences existed using the two databases. Four more MF annotations were enriched using the Public database, including lipase activity, hydrolase activity, phosphatidylcholine phospholipase C activity, and anion binding. Thus, more BP and MF annotations were enriched using the Public database, which was not enriched using the Meta database. However, the main significant changes in the biological progress and molecular function between the two kinds of soil could be obtained using either database.

Similar functional microorganisms were obtained using the two databases

Phosphatases and phospholipases are necessary functional enzymes in response to P deficiency in soils (Yao *et al.*, 2018). The identified phosphatases and phospholipases obtained by the two databases were used for constructing the phylogenetic trees, respectively (Fig. 9), to assess the influences of databases on exploring the corresponding key functional microorganisms. A total of 28 phosphatases and 30 phospholipases were identified using the two databases. About half of enzymes (14 phosphatases and 17 phospholipases) were from the proteins identified using the Public database, and the others were from the Meta database. Although more proteins were matched using the Meta database, the numbers of these two critical functional enzymes identified using the two databases were almost the same. The phosphatases identified could be divided into two main branches, representing the alkaline phosphatase (AL) and acid phosphatase (AC), respectively (Fig. 9). It showed that the majority of subbranches contained proteins identified using the two databases. The proteins clustered together shared the highly conserved motifs. The amino acid similarities of the identified motifs using the two databases were extremely high. Among 17 ALs, ten ALs were identified using the Meta database. The ALs were mainly from *Alphaproteobacteria* bacterium (5) and *Candidatus rokubacteria* (2). Moreover, these ALs from these bacteria could also be identified using the Public database, because *Pseudolabrys taiwanensis* and *Methylobacterium* sp. belong to *Alphaproteobacteria*. *Cyanobacteria* bacterium could only be identified using the Meta database, while AL from *Frankia* sp. could only be identified using the Public database. For the ACs, most of the microorganisms identified using the two databases belongs to the Betaproteobacteria, although the microorganisms identified at the family or the species levels had differences. The phylogenetic tree of phospholipases could also be divided into two main branches, representing the phospholipase (Branch I) and acid phospholipase (Branch II), respectively (Fig. S6). Most of the microorganisms identified using the two databases belonged to the Betaproteobacteria and *Actinobacteria*.

Fig. 9 Phylogenetic trees and the sequence alignment of the representative conserved motifs of alkaline phosphatase (AL) acid phosphatase (AC) identified using Meta database (M, red) and Public database (P, blue).

The protein identification number in the Meta database and accession number from the Public database were showed in brackets. The same conserved motifs were showed in the same color boxes. The multiple sequence alignments are motif 1 and motif 4 from branches I and II of the phylogenetic tree, respectively. The amino acid similarity higher than 70% showed below the sequences.

Nevertheless, for acid phospholipases, the microorganisms identified using the Public database were much more specific than those identified using the Meta database. However, it should be noticed that the cross-species identifications may cause more specific taxonomic interpretation by the Public database for the high amino acid similarity between closely related species, or horizontal gene transfer between species. In summary, the main functional microorganisms identified using the two databases were similar, which was not affected by the different databases used.

DISCUSSION

Soil metaproteomics has been applied increasingly, analyzing the soil microbial functions with the development of soil protein extraction methods and mass spectrometry technology in recent ten years. However, bioinformatic analyses for complex and unknown microbial communities are still confusing and poorly studied. This study thoroughly and systematically demonstrated the soil metaproteomic workflow and results using the two protein sequence databases, the Meta database and the Public database. It was evident that more proteins and microbial taxonomies could be identified using the Meta database in soil metaproteomics. However, the primary metaproteomic results could also be obtained using the Public database, getting a rough overview of function and taxonomy quickly. This study provides a reliable basis for database selection in future soil metaproteomics.

The total number of proteins in the two protein sequence databases used in this study were similar, and the total lengths of proteins were even longer in the Public database, but the amount of the identified proteins had vast differences. The Meta database usage acquired much more and better results with more identified proteins, higher coverages of protein sequences, and more microorganisms at the genus level compared with those obtained using the Public database. Thus, soil metagenomics provides a more special and detailed overview of uncultured soil microbes' genomes and customized protein sequence databases for soil metaproteomics (Nesme *et al.*, 2016). The Meta database derives from the soil microorganisms in situ, and the specificity and coverage are higher than the commonly used Public databases, which may be critical to the more identified proteins. It was verified that more proteins identified by the Meta database could not be aligned to the proteins identified by the Public database, which also reflected the complexity and specificity of soil microbial communities.

Furthermore, we can get more functional annotations, more specific proteins from distinctive microorganisms living in soil environments corresponding to current research by the Meta database. It is more conducive to subsequent data mining and phenomenon interpretation. In recent years, integrated-omics studies have been widely used in studies of the soil microbial communities and functions in diverse ecosystems (Johnson-Rollings *et al.*, 2014; Hultman *et al.*, 2015; Butterfield *et al.*, 2016; Yao *et al.*, 2018). Soil proteogenomics could display an accurate and comprehensive picture of soil microbial communities and functions at the gene and protein level.

The GO enrichment analysis was necessary for exploring the key processes or functions induced by environmental factors. It is worth noting that the use of protein sequences related to soil microbes from the Public database lead to similar soil metaproteomic results about the cluster analysis of the different samples, microbial community composition at the genus level, and the enrichment analysis of GO annotations of proteins and functional microorganisms. However, it saved the expenses of metagenomics. Even though the cost of genomic sequencing is getting lower, researchers should pay high efforts on the soil metagenomes, including extraction, amplification, sequencing, binning, and annotation. Nevertheless, in soil metagenomics analysis, all soil microbial microorganisms' complete coverage is almost impossible for soil's complicated composition. Inevitably, some microorganisms are missed in the process of experiments. Sometimes, the measurements even dramatically differed from the truth, for these experiments are biased toward detecting some taxa over others (McLaren and Callahan, 2018; McLaren *et al.*, 2019). These resulted in unidentified spectra, even some high-quality or high-abundant ones. So, it is still challenging to assemble and annotate soil metagenomes accurately. With the cost reduction of sequencing and the going-deep researches, more and more microorganism genomes are well annotated. The datasets can be easily accessed and are well-curated. Thus, for the limit of the fund or laboratory conditions or the aim of analyzing the microbial responses in differently treated soils, the Public database could be a high-efficient and low-cost alternative.

In metaproteomic researches about previously uncharacterized and complex environments such as the marine microbial community, different database construction approaches lead to divergent taxonomic and functional interpretations (Timmins-Schiffman *et al.*, 2017). Some researchers prefer databases built by assembling metagenome than public databases because the metaproteome-specific database excludes non-specific sequences and yields a greater variety of taxa and functional annotations. While in most cases, the soil metagenomes are in quite poor quality and highly fragmented, and the taxa assignments are often dismal. Our results point out the similar microbial composition and functional change trend can be obtained by these two kinds of databases. The selection of databases and downstream data interpretation should be executed with care according to sample types and research purposes.

Expect the above two kinds of databases, integration of two databases purposefully is probably a good strategy. Using the data from previous gut microbiota metaproteomics (Tanca *et al.*, 2016), a merged database was deployed, including taxonomy-guided reference protein sequences from public databases and proteins from metagenome assembly (Xiao *et al.*, 2018). Compared with the metagenomic database, about two-folds peptides could be identified using the merged database with high sensitivity and peptide identification precision. More proteins from poorly characterized species within the gut microbiota could be detected. In human intestinal metaproteomics, the protein sequence database from a wide range of expected protein in fecal samples and several above databases' subsets was tested (Muth *et al.*, 2015). The integration of orthogonal information from other research fields such as metagenomics and 16S rRNA sequencing increased the metaproteomic results' confidence. In soil metaproteomics, this strategy was used to analyze microbial communities from the permafrost, active layer, and thermokarst bog soil by multi-omics technology (Hultman *et al.*, 2015). The matched metagenomic assembly, microbial genomes of 180 environmental isolates, and 13 recently sequenced isolates from cold environments were also used as the protein sequence database. Approximately 7 000 proteins were identified in this study. The environmental isolates increased the specificity of the database, especially the isolates from cold environments. Therefore, the combination of protein sequences from metagenomic assembly and genomes of environmental isolates from public databases may increase the integrity and specificity of the reference protein database in soil proteomics.

CONCLUSIONS

In this study, we used two strategies to construct protein sequence databases with comparable distribution in their protein lengths in soil metaproteomics, and demonstrated similarities and differences of their downstream bioinformatic analysis results using two kinds of databases. Using the Meta database, more proteins, higher sequences coverages, and even more microbial taxonomies could be identified. The Meta database showed some superiority over the Public database in soil metaproteomics. However, regardless of databases used, the enrichment analysis of GO terms with differential abundance identified exhibited high similarity, independent from the number of identified proteins, which means the Public database could also meet the demand to explore the functional responses of microbial proteins between soils with different treatments. This study provides useful insights into choosing the protein sequence database and how to perform bioinformatic analyses in soil metaproteomic analyses.

ACKNOWLEDGEMENT(S)

The first two authors, XIONG Yi and ZHENG Lu, contribute equally to this work. This work was supported by the National Key Research and Development Program of China (No. 2016YFD0200308), the National Key Basic Research Program of China (No. 2015CB150501), and the Project of Priority and Key Areas, Institute of Soil Science, Chinese Academy of Sciences (No. ISSASIP1605 and ISSASIP1640)

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version.

REFERENCES

Bailey T L, Boden M, Buske F A, Frith M, Grant C E, Clementi L, Ren J, Li W W, Noble W S. 2009. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**: W202-W208.

Bastida F, Hernandez T, Garcia C. 2014. Metaproteomics of soils from semiarid environment: Functional and phylogenetic information obtained with different protein extraction methods. *J Proteomics.* **101**: 31-42.

Bastida F, Jehmlich N, Lima K, Morris B E L, Richnow H H, Hernandez T, von Bergen M, Garcia C. 2016. The ecological and physiological responses of the microbial community from a semiarid soil to hydrocarbon contamination and its bioremediation using compost amendment. *J Proteomics.* **135**: 162-169.

Bastida F, Moreno J L, Nicolás C, Hernández T, García C. 2009. Soil metaproteomics: a review of an emerging environmental science. Significance, methodology and perspectives. *Eur J Soil Boil.* **60**: 845-859.

Benndorf D, Balcke G U, Harms H, von Bergen M. 2007. Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J.* **1**: 224-234.

Brown C T, Xiong W L, Olm M R, Thomas B C, Baker R, Firek B, Morowitz M J, Hettich R L, Banfield J F. 2018. Hospitalized premature infants are colonized by related bacterial strains with distinct proteomic profiles. *mBio.* **9**: e00441-00418.

Burns R G, DeForest J L, Marxsen J, Sinsabaugh R L, Stromberger M E, Wallenstein M D, Weintraub M N, Zoppini A. 2013. Soil enzymes in a changing environment: Current knowledge and future directions. *Soil Biol Biochem.* **58**: 216-234.

Butterfield C N, Li Z, Andeer P F, Spaulding S, Thomas B C, Singh A, Hettich R L, Suttle K B, Probst A J, Tringe S G, Northen T, Pan C, Banfield J F. 2016. Proteogenomic analyses indicate bacterial methylotrophy and archaeal heterotrophy are prevalent below the grass root zone. *PeerJ.* **4**: e2687.

Carlson M. 2019. GO. db: A set of annotation maps describing the entire Gene Ontology. *R package* version 3.8.2. from https://doi.org/doi:10.18129/B9.bioc.GO.db.

Carvalhais L C, Dennis P G, Tyson G W, Schenk P M. 2012. Application of metatranscriptomics to soil environments. *J Microbiol Methods.* **91**: 246-251.

Chamberlain S A, Szocs E. 2013. taxize: taxonomic search and retrieval in R. *F1000Res.* **2**: 191.

Chapman B, Bellgard M. 2014. High-throughput parallel proteogenomics: a bacterial case study. *Proteomics.* **14**: 2780-2789.

Chen C, Xia R, Chen H, He Y. 2018. TBtools, a Toolkit for Biologists integrating various biological data handling tools with a user-friendly interface. *BioRxiv.* 289660.

Chen H. 2016. VennDiagram: generate high-resolution venn and euler plots. *R package* version 1.6.17. from https://CRAN.R-project.org/package=VennDiagram.

Chiapello M, Zampieri E, Mello A. 2020. A small effort for researchers, a big gain for soil metaproteomics. *Front Microbiol.* **11**: 88.

Chourey K, Jansson J, VerBerkmoes N, Shah M, Chavarria K L, Tom L M, Brodie E L, Hettich R L. 2010. Direct cellular lysis/protein extraction protocol for soil metaproteomics. *J Proteome Res.* **9**: 6615-6622.

Conesa A, Götz S, García-Gómez J M, Terol J, Talón M, Robles M. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* **21**: 3674-3676.

Conway J R, Lex A, Gehlenborg N. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics.* **33**: 2938-2940.

Daniel R. 2005. The metagenomics of soil. *Nat Rev Microbiol.* **3**: 470-478.

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

Grassl N, Kulak N A, Pichler G, Geyer P E, Jung J, Schubert S, Sinitcyn P, Cox J, Mann M. 2016. Ultra-deep and quantitative saliva proteome reveals dynamics of the oral microbiome. *Genome Med.* **8**: 44.

Heyer R, Schallert K, Zoun R, Becher B, Saake G, Benndorf D. 2017. Challenges and perspectives of metaproteomic data analysis. *J Biotechnol.* **261**: 24-36.

Hultman J, Waldrop M P, Mackelprang R, David M M, McFarland J, Blazewicz S J, Harden J, Turetsky M R, McGuire A D, Shah M B, VerBerkmoes N C, Lee L H, Mavrommatis K, Jansson J K. 2015. Multi-omics of permafrost, active layer and thermokarst bog soil microbiomes. *Nature.* **521**: 208-212.

Jansson J K, Hofmockel K S. 2018. The soil microbiome-from metagenomics to metaphenomics. *Curr Opin Microbiol.* **43**: 162-168.

Johnson-Rollings A S, Wright H, Masciandaro G, Macci C, Doni S, Calvo-Bado L A, Slade S E, Plou C V, Wellington E M H. 2014. Exploring the functional soil-microbe interface and exoenzymes through soil metaexoproteomics. *ISME J.* **8**: 2148-2150.

Johnson J S, Spakowicz D J, Hong B Y, Petersen L M, Demkowicz P, Chen L, Leopold S R, Hanson B M, Agresta H O, Gerstein M, Sodergren E, Weinstock G M. 2019. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. *Nat Commun.* **10**: 5029.

Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. *J Mol Biol.* **428**: 726-731.

Keiblinger K M, Fuchs S, Zechmeister-Boltenstern S, Riedel K. 2016. Soil and leaf litter metaproteomics-a brief guideline from sampling to understanding. *FEMS Microbiol Ecol.* **92**: fiw180.

Kleiner M, Thorson E, Sharp C E, Dong X, Liu D, Li C, Strous M. 2017. Assessing species biomass contributions in microbial communities via metaproteomics. *Nat Commun.* **8**: 1558.

Kolde R. 2015. Pretty Heatmaps. *R package* version 1.0.12. from https://CRAN.R-project.org/package=pheatmap.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol.* **35**: 1547-1549.

Kunath B J, Minniti G, Skaugen M, Hagen L H, Vaaje-Kolstad G, Eijsink V G H, Pope P B, Arntzen M Ø. 2019. Metaproteomics: sample preparation and methodological considerations. *In* Jos é-Luis Capelo-Mart ńez (ed.) Emerging sample treatments in proteomics.

Lin W, Wu L, Lin S, Zhang A, Zhou M, Lin R, Wang H, Chen J, Zhang Z, Lin R. 2013. Metaproteomic analysis of ratoon sugarcane rhizospheric soil. *BMC Microbiol.* **13**: 135.

Love M I, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology.* **15**: 550.

Madden T. 2013. The BLAST sequence analysis tool. *In* McEntyre J and Ostell J (ed.) The NCBI Handbook. *In* Bethesda (MD): National Center for Biotechnology Information, US.

Marchler-Bauer A, Bryant S H. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* **32**: W327-W331.

Maron P-A, Ranjard L, Mougel C, Lemanceau P. 2007. Metaproteomics: a new approach for studying functional microbial ecology. *Microb Ecol.* **53**: 486-493.

McLaren M R, Callahan B J. 2018. In nature, there is only fiversity. *mBio.* **9**: e02149-02117.

McLaren M R, Willis A D, Callahan B J. 2019. Consistent and correctable bias in metagenomic sequencing experiments. *eLife.* **8**: 31.

Murase A, Yoneda M, Ueno R, Yonebayashi K. 2003. Isolation of extracellular protein from greenhouse soil. *Soil Biol Biochem.* **35**: 733-736.

Muth T, Kolmeder C A, Salojarvi J, Keskitalo S, Varjosalo M, Verdam F J, Rensen S S, Reichl U, de Vos W M, Rapp E, Martens L. 2015. Navigating through metaproteomics data: a logbook of database searching. *Proteomics.* **15**: 3439-3453.

Nannipieri P. 2006. Role of stabilised enzymes in microbial ecology and enzyme eextraction from soil with potential applications in soil proteomics. *In* Nannipieri P and Smalla K (ed.) Nucleic Acids and Proteins in Soil. Springer Berlin Heidelberg, New York.

Nesme J, Achouak W, Agathos S N, Bailey M, Baldrian P, Brunel D, Frostegard A, Heulin T, Jansson J K, Jurkevitch E, Kruus K L, Kowalchuk G A, Lagares A, Lappin-Scott H M, Lemanceau P, Le Paslier D, Mandic-Mulec I, Murrell J C, Myrold D D, Nalin R, Nannipieri P, Neufeld J D, O'Gara F, Parnell J J, Puhler A, Pylro V, Ramos J L, Roesch L F, Schloter M, Schleper C, Sczyrba A, Sessitsch A, Sjoling S, Sorensen J, Sorensen S J, Tebbe C C, Topp E, Tsiamis G, van Elsas J D, van Keulen G, Widmer F, Wagner M, Zhang T, Zhang X, Zhao L, Zhu Y G, Vogel T M, Simonet P. 2016. Back to the future of soil metagenomics. *Front Microbiol.* **7**: 73.

Nesvizhskii A I, Aebersold R. 2005. Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics.* **4**: 1419-1440.

Pagès H, Aboyoun P, Gentleman R, DebRoy S. 2019. Biostrings: Efficient manipulation of biological strings. *R package* version 2.52.0. from https://doi.org/doi:10.18129/B9.bioc.Biostrings.

Pedersen T L. 2016. ggforce: Accelerating 'ggplot2'. *R package* version 0.2.2. from https://CRAN.R-project.org/package=ggforce.

Qian C, Hettich R L. 2017. Optimized extraction method to remove humic acid interferences from soil samples prior to microbial proteome measurements. *J Proteome Res.* **16**: 2537-2546.

R_Core_Team. 2019. R: A language and environment for statistical computing. *R for Statistical Computing* Vienna, Austria. from https://www.R-project.org/.

Renella G, Ogunseitan O, Giagnoni L, Arenella M. 2014. Environmental proteomics: A long march in the pedosphere. *Soil Biol Biochem.* **69**: 34-37.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one.* **11**: e0163962.

Smyth G K. 2005. Bioinformatics and computational biology solutions using R and Bioconductor. Springer.

Tanca A, Palomba A, Deligios M, Cubeddu T, Fraumene C, Biosa G, Pagnozzi D, Addis M F, Uzzau S. 2013. Evaluating the impact of different sequence databases on metaproteome analysis: insights from a lab-assembled microbial mixture. *PloS one.* **8**: 14.

Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, Muth T, Rapp E, Martens L, Addis M F, Uzzau S. 2016. The impact of sequence database choice on metaproteomic results in gut microbiota studies. *Microbiome.* **4**: 51.

Tartaglia M, Bastida F, Sciarrillo R, Guarino C. 2020. Soil metaproteomics for the study of the relationships between microorganisms and plants: A review of extraction protocols and ecological insights. *Int J Mol Sci.* **21**: 8455.

Timmins-Schiffman E, May D H, Mikan M, Riffle M, Frazar C, Harvey H R, Noble W S, Nunn B L. 2017. Critical decisions in metaproteomics: achieving high confidence protein annotations in a sea of unknowns. *ISME J.* **11**: 309-314.

Torsvik V, Øvreås L. 2002. Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol.* **5**: 240-245.

van der Heijden M G, Bardgett R D, van Straalen N M. 2008. The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett.* **11**: 296-310.

Verheggen K, Raeder H, Berven F S, Martens L, Barsnes H, Vaudel M. 2020. Anatomy and evolution of database search engines-a central component of mass spectrometry based proteomic workflows. *Mass Spectrom Rev.* **39**: 292-306.

Wang H B, Zhang Z X, Li H, He H B, Fang C X, Zhang A J, Li Q S, Chen R S, Guo X K, Lin H F, Wu L K, Lin S, Chen T, Lin R Y, Peng X X, Lin W X. 2011. Characterization of metaproteomics in crop rhizospheric soil. *J Proteome Res.* **10**: 932-940.

Wickham H. 2011. ggplot2. *Wiley Interdiscip Rev: Comput Stat.* **3**: 180-185.

Wickham H. 2017. stringr: Simple, consistent wrappers for common string operations. *R package* version 1.4.0. Retrieved 0, 1, from https://CRAN.R-project.org/package=stringr.

Wickham H, Francois R, Henry L, Müller K. 2015. dplyr: A grammar of data manipulation. *R package* version 0.4.3. from http://CRAN.R-project.org/package=dplyr.

Wickham H, Henry L. 2019. tidyr: Easily tidy data with 'spread()' and 'gather()' functions. *R package* version 0.8.3. from https://CRAN.R-project.org/package=tidyr.

Winter D J. 2017. rentrez: An R package for the NCBI eUtils API. *R Journal.* **9**: 520-526.

Wiśniewski J R, Zougman A, Nagaraj N, Mann M. 2009. Universal sample preparation method for proteome analysis. *Nat Methods.* **6**: 359-362.

Xiao J, Tanca A, Jia B, Yang R, Wang B, Zhang Y, Li J. 2018. Metagenomic taxonomy-guided database-searching strategy for improving metaproteomic analysis. *J Proteome Res.* **17**: 1596-1605.

Xiong Y, Lin X, Lan P. 2016. The development of soil protein extraction methods in soil metaproteomics. *Soils (in Chinese).* **48**: 855-843.

Yadav A K, Kumar D, Dash D. 2012. Learning from decoys to improve the sensitivity and specificity of proteomics database search results. *PloS one.* **7**: e50651.

Yao Q, Li Z, Song Y, Wright S J, Guo X, Tringe S G, Tfaily M M, Pasa-Tolic L, Hazen T C, Turner B L, Mayes M A, Pan C. 2018. Community proteogenomics reveals the systemic impact of phosphorus availability on microbial functions in tropical soil. *Nat Ecol Evol.* **2**: 499-509.

Yu G, Wang L G, Han Y, He Q Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics.* **16**: 284-287.

Zampieri E, Chiapello M, Daghino S, Bonfante P, Mello A. 2016. Soil metaproteomics reveals an inter-kingdom stress response to the presence of black truffles. *Sci Rep.* **6**: 25773.

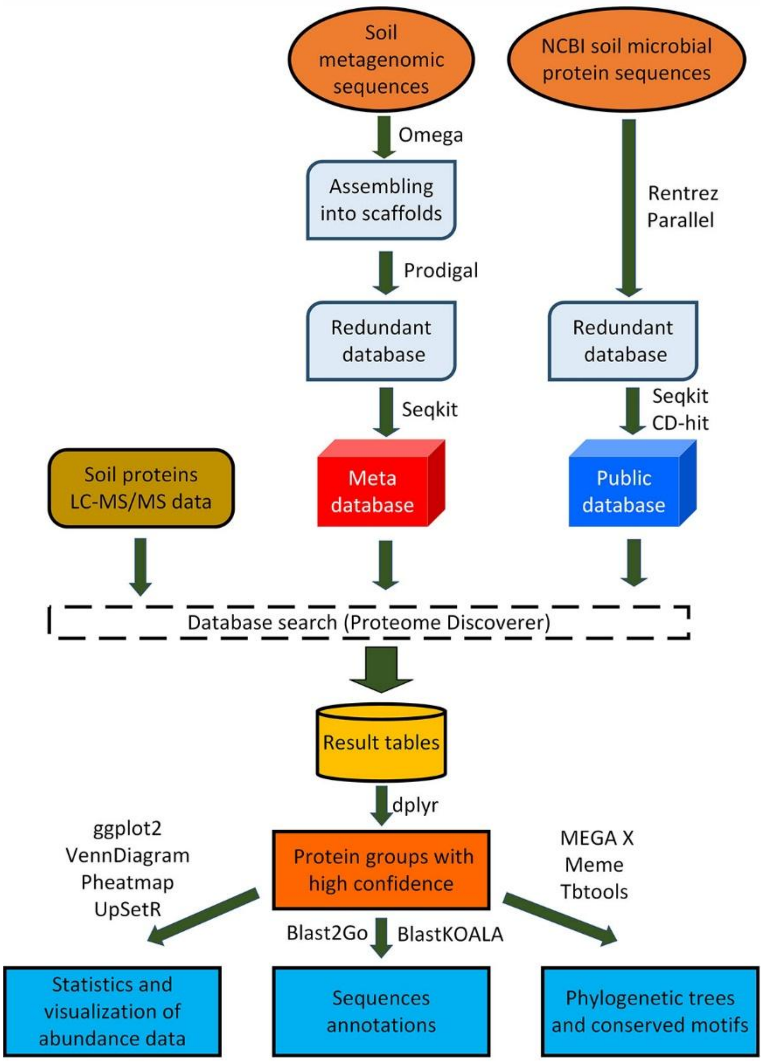
Soil metagenomic sequences
NCBI soil microbial protein sequences
Omega
Assembling into scaffolds
Rentrez
Parallel
Prodigal
Redundant database
Redundant database
Seqkit
Seqkit
CD-hit
Soil proteins LC-MS/MS data
Meta database
Public database
Database search (Proteome Discoverer)
Result tables
dplyr
ggplot2
VennDiagram
Pheatmap
UpSetR
Protein groups with high confidence
MEGA X
Meme
Tbtools
Blast2Go
BlastKOALA
Statistics and visualization of abundance data
Sequences annotations
Phylogenetic trees and conserved motifs

Fig. 1

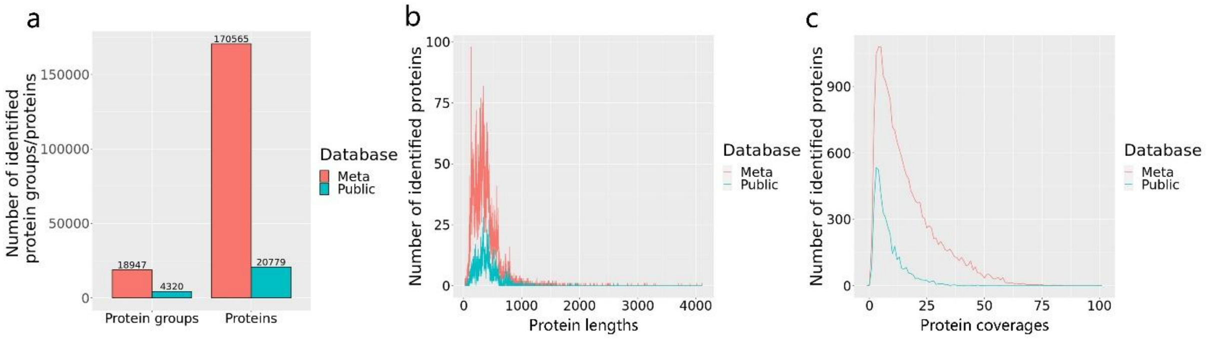
a
Number of identified protein groups/proteins
150000
100000
50000
0
18947
4320
170565
20779
Protein groups
Proteins
Database
Meta
Public
b
Number of identified proteins
100
75
50
25
0
0
1000
2000
3000
4000
Protein lengths
Database
Meta
Public
c
Number of identified proteins
900
600
300
0
0
25
50
75
100
Protein coverages
Database
Meta
Public

Fig. 2

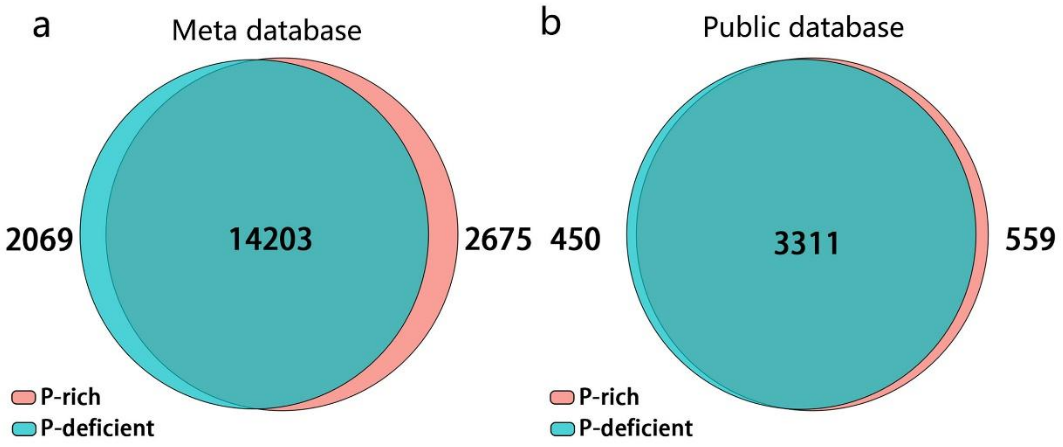

Fig. 3

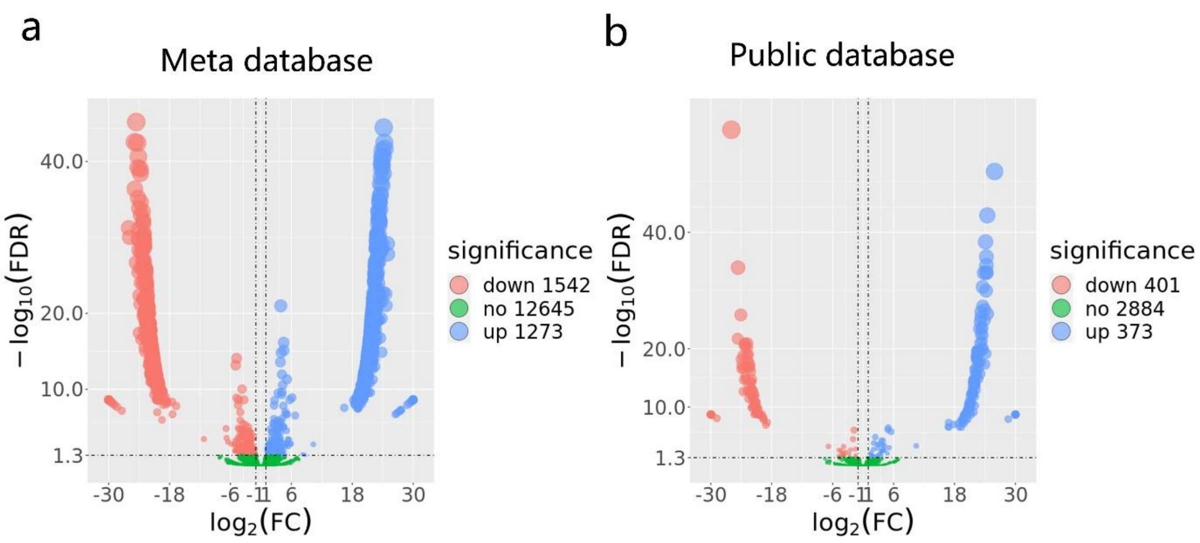

Fig. 4

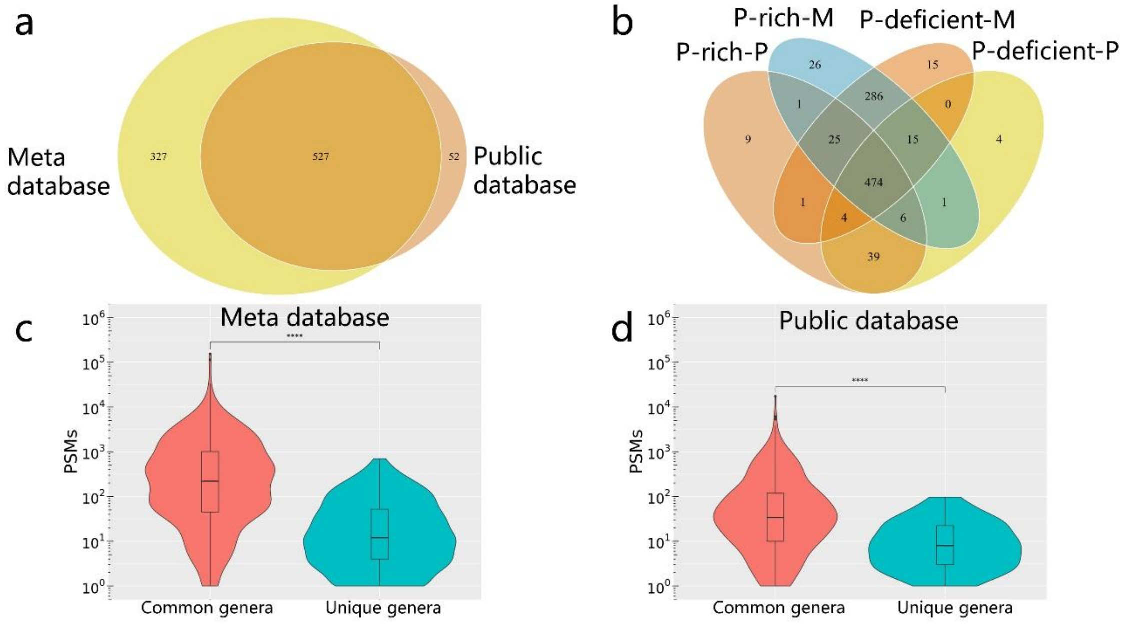

Fig. 5

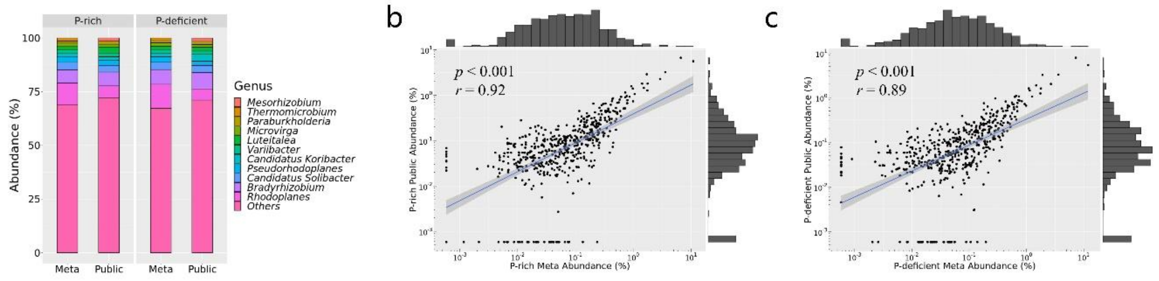

Fig. 6

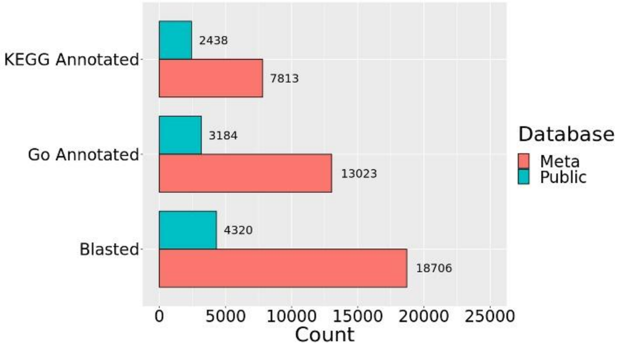

Fig. 7

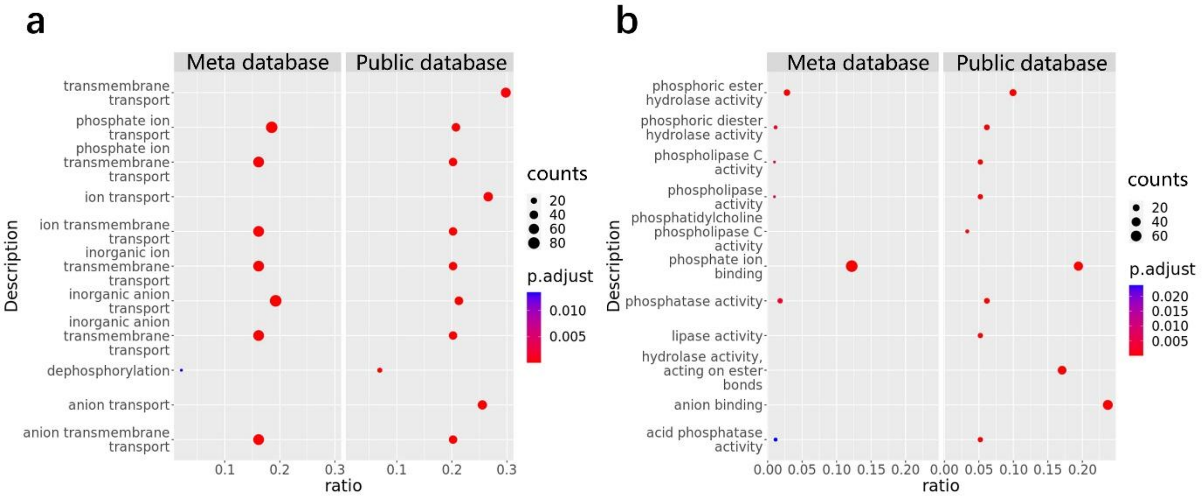

Fig. 8

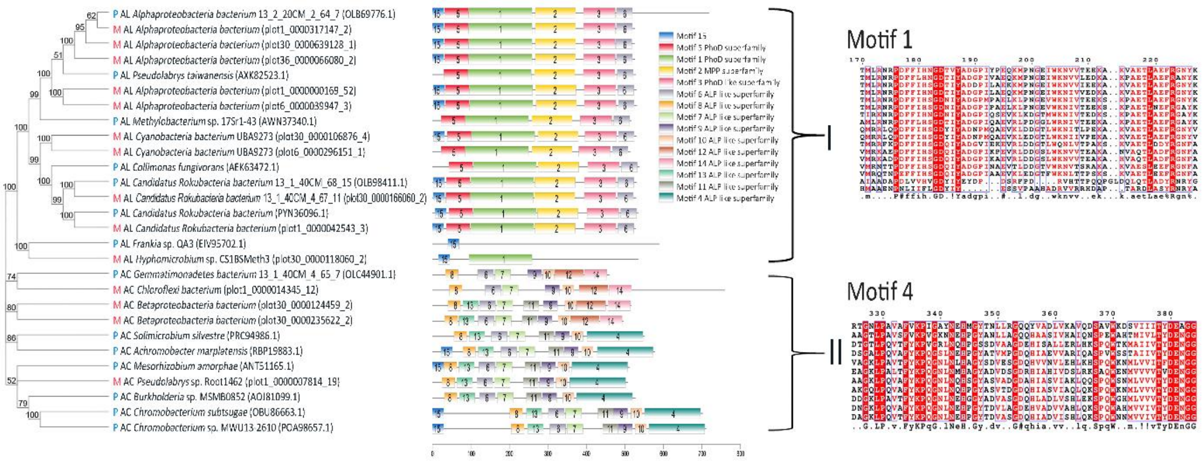
P AL Alphaproteobacteria bacterium 13_2_20CM_2_64_7 (OLB69776.1)
M AL Alphaproteobacteria bacterium (plot1_0000317147_2)
M AL Alphaproteobacteria bacterium (plot30_0000639128_1)
M AL Alphaproteobacteria bacterium (plot36_0000066080_2)
P AL Pseudolabrys taiwanensis (AXK82523.1)
M AL Alphaproteobacteria bacterium (plot1_0000000169_52)
M AL Alphaproteobacteria bacterium (plot6_0000039947_3)
P AL Methylobacterium sp. 17Sr1-43 (AWN37340.1)
M AL Cyanobacteria bacterium UBA9273 (plot30_0000106876_4)
M AL Cyanobacteria bacterium UBA9273 (plot6_0000296151_1)
P AL Collimonas fungivorans (AEK63472.1)
P AL Candidatus Rokubacteria bacterium 13_1_40CM_68_15 (OLB98411.1)
M AL Candidatus Rokubacteria bacterium 13_1_40CM_4_67_11 (plot30_0000166060_2)
P AL Candidatus Rokubacteria bacterium (PYN36096.1)
M AL Candidatus Rokubacteria bacterium (plot1_0000042543_3)
P AL Frankia sp. QA3 (EIV95702.1)
M AL Hyphomicrobium sp. CS1BSMeth3 (plot30_0000118060_2)
P AC Gemmatimonadetes bacterium 13_1_40CM_4_65_7 (OLC44901.1)
M AC Chloroflexi bacterium (plot1_0000014345_12)
M AC Betaproteobacteria bacterium (plot30_0000124459_2)
M AC Betaproteobacteria bacterium (plot30_0000235622_2)
P AC Solimicrobium silvestre (PRC94986.1)
P AC Achromobacter marplatensis (RBP19883.1)
P AC Mesorhizobium amorphae (ANT51165.1)
M AC Pseudolabrys sp. Root1462 (plot1_0000007814_19)
P AC Burkholderia sp. MSMB0852 (AOJ81099.1)
P AC Chromobacterium subtsugae (OBU86663.1)
P AC Chromobacterium sp. MWU13-2610 (POA98657.1)
62
95
100
51
100
99
100
100
100
99
100
100
99
100
100
100
74
80
86
52
79
100
Motif 15
Motif 5 PhoD superfamily
Motif 1 PhoD superfamily
Motif 2 MPP superfamily
Motif 3 PhoD like superfamily
Motif 6 ALP like superfamily
Motif 8 ALP like superfamily
Motif 7 ALP like superfamily
Motif 9 ALP like superfamily
Motif 10 ALP like superfamily
Motif 12 ALP like superfamily
Motif 14 ALP like superfamily
Motif 13 ALP like superfamily
Motif 11 ALP like superfamily
Motif 4 ALP like superfamily
0
100
200
300
400
500
600
700
800
I
II
Motif 1
Motif 4

Fig. 9