Metagenomic analysis of heavy metal-contaminated soils reveals distinct clades with adaptive features

B. Thakur, R K Yadav, Roland Marmeisse, S. Prashanth, M. Krishnamohan, Laurence Fraissinet-Tachet, M S Reddy

International Journal of Environmental Science and Technology 2022
https://doi.org/10.1007/s13762-022-04635-5

Metagenomic analysis of heavy metal-contaminated soils reveals distinct clades with adaptive features

B. Thakur[1] ·R. K. Yadav[2] R. Marmeisse[3] ·S. Prashanth[4] M. Krishnamohan[5] L. F. Tachet[3] M. S. Reddy[1]

1 Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab 147004, India

2 Department of Botany, University of Allahabad, Prayagraj, Uttar Pradesh 211002, India

3 Ecologie Microbienne, UMR CNRS, UMR INRA, Université Claude Bernard Lyon1, Université de Lyon, 69622 Villeurbanne, France

4 Amrita School of Biotechnology, Amrita Vishwavidyapeetham, Clappana, Kerala 690525, India

5 Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur, Rajasthan 302001, India

Abstract
Heavy metal pollution poses a serious threat to soil, water, and the atmospheric environment. Many microbes sustain and respond to the metal stress, and the mechanisms involved in these processes remain unclear. The metagenomics and metatranscriptomics approaches were applied to study the structure and function of eukaryotic microbial communities in heavy metal-contaminated soils of two geographic locations situated in different climatic regions. To achieve this, amplicons of the hypervariable V4 region of 18S rDNA and cDNA synthesized from 18S rRNA extracted from these soils were generated and sequenced through paired-end sequencing chemistry on the Illumina-MiSeq platform. The NGS dataset processed by the Mothur pipeline and analyzed by Parallel-Meta 3 pipeline illustrated the presence of all the major eukaryotic phyla. Taxa diversity and community structure of micro-eukaryotes within and between the samples from two locations were compared. Clustering, heatmap, and PCA analysis supported the variation in taxonomic diversity and community structure in the datasets of these two sites. Analysis of taxa abundance in both sites identified marker organisms for the further characterization of such types of environments. A functional metatranscriptomics study revealed the identification of various expressed eukaryotic genes, which are involved in metal tolerance. These metal-tolerant gene families were phylogenetically related to the different eukaryotic lineages reported in the metagenomic analysis. Hence, this approach could be widely applied in microbial ecology to understand the role of active microbes in such specific environmental conditions.

Keywords Eukaryotic microbes Metatranscriptomics Microbial ecology· Molecular taxonomy

Introduction

Soil is a complex system of various known and unknown phenomena playing vital ecological functions in environmental cycles. It also contains numerous microorganisms, which are unknown to us, bound with biodiversity and phylogeny. It has been reported that one gram of soil contains about ten billion microorganisms of possibly thousands of prokaryotic and eukaryotic species interwoven in numerous micro-niches performing ecologically significant functions (Gans et al. 2005). However, the individual taxa's precise functional and ecological significance remains uncertain because most microbes cannot be cultured under laboratory conditions (Fierer et al. 2007). Hence, the most basic and essential questions from an environmental biology standpoint of view remain unanswered. With the advancement in technologies such as next-generation sequencing (NGS), metagenomics, and

metatranscriptomics, it has become possible to understand microbial ecology and its functions in detail (Leimena et al. 2013). The metagenomics approach structures and their functions in various ecosystems like soil (Rondon et al., 2000), water (Rodriguez-Brito et al. 2010), and human gut (Qin et al. 2010). Metatranscriptomics has been applied to understand the functional diversity of various ecosystems (Lehembre et al. 2013). Eukaryotic microbes are the indispensable biotic component of many ecosystems. It is arduous to perform both approaches to infer the taxonomic and functional diversity of eukaryotic communities flourishing in adverse environments including heavy metalcontaminated soils.

Heavy metal contaminants are introduced into the soil through various anthropogenic activities like mining, smelting, industrialization, and fossil fuel consumption (Zhang et al. 2018). These processes are long, irreversible, concealed, and uncontrollable and pose a severe environmental threat (Mico et al. 2006). Therefore, efforts to overcome the prevalent issue of heavy metal contamination by characterizing the organisms have been attempted before. Metagenomic and metatranscriptomic approaches played an important role in understanding the structure and functional diversity of organisms in contaminated soil (Lehembre et al. 2013; Marmeisse et al. 2017). Metagenomic and metatranscriptomic approaches were used to analyze the soil microbial communities and their function in this investigation. 18S rDNA/rRNA-based NGS analyses were applied to know the insight into the taxonomic structure of the eukaryotic microbes in the heavy metal-contaminated soils from two locations situated in different climatic regions. An attempt has also been made through functional metatranscriptomics to study the functional diversity of eukaryotic microbes by screening active eukaryotic genes expressed in metalcontaminated soils.

Materials and methods

Site description, soil sampling, and processing

Soil samples were collected from two geographic locations; an agroforestry land at Pierrelaye (PL) in the North-West of Paris, France, and Zawar mines (ZM), Udaipur, India, were used in this study. The PL site is situated in the semi-oceanic temperate region, while the ZM site is located in the tropical monsoon continental climate. The location at Pierrelaye was earlier a maize-growing field, but prolonged irrigation with wastewater had contaminated this site with heavy metals. Subsequently, this field was converted into agroforestry land where poplar trees are being cultivated. The location at Zawar mines is a Zn mining and processing area; therefore, the soil is highly contaminated with heavy metals. To represent the biological replicates, soil from each location was collected form three sites following the standard sampling each site and mixed in equal parts to form the composite sample after serially sieving them with mesh of different sizes to remove unwanted components. Three composite soil samples from Pierrelaye (PL), namely S1A, S1B and S2, and three composite soil samples from Zawar mines (ZM), namely Zn1, Zn2, and ZnN, were collected. All samples were frozen and stored at −70 ℃. The soil sample was analyzed for its physicochemical properties, i.e., pH, organic carbon (Walkley 1947), total P (Kitson and Mellon 1944), available P (Olsen 1954), and total nitrogen (Piper 1966). The metal contents of the soil samples were determined by ICP-AES (ARCOS, Simultaneous ICP Spectrometer, SPECTRO Analytical Instruments GmbH, Germany).

Table 1 Physicochemical properties of the soil samples collected from two geographic locations

Soil characteristics	Pierrelaye, France (PL)	Zawar mines, India (ZM)
Soil type	Sandy luvisol soil	Sandy ustochrepts soil
pH	7.06	7.56
Organic carbon (%)	1.6	2.27
Total phosphorus (mg/kg)	291	252.4
Available phosphorus (mg/kg)	14.2	1.92

Total nitrogen (%)	0.12	0.18
Cadmium (mg/kg)	2.5	3.0
Zinc (mg/kg)	385	447.3
Copper (mg/kg)	64	0.94
Lead (mg/kg)	2.02	175.5

Nucleic acid extraction and cDNA synthesis

Total genomic DNA (gDNA) was extracted from frozen soil samples according to the method described by Bailly et al. (2007). Total RNA was extracted from the frozen soil samples using RNA PowerSoil® Total RNA Isolation Kit (MoBio Laboratories, Carlsbad, CA, USA) according to the manufacturer's instructions. Each of the extracted DNA and RNA samples was treated with RNase A and DNase I, respectively, to remove DNA and RNA impurities. Quality of nucleic acids was analyzed by agarose gel electrophoresis, while quantity and purity were determined by UV spectrophotometry (SAFAS UVmc2, SAFAS Monaco). First-strand complementary DNA (cDNA) from total RNA was synthesized by using random hexamers and SuperScript reverse transcriptase (Invitrogen). Total twelve nucleic acid samples: six genomic DNA (gDNA) and six cDNA, representing six soil samples, were further used for targeted amplification through PCR.

Targeted amplification of V4 region of 18S rDNA/ rRNA

The targeted V4 region of 18S rDNA/rRNA sequence from gDNA as well as cDNAs was amplified by using eukaryote specific degenerate primer sets (Tables 2 and 3). These primers were specific to ~ 380 bp long V4 region of the 18S rDNA gene and its transcribed rRNA (Hadziavdic et al. 2014). Each primer consisted of four random nucleotide sequences as linker sequence followed by an eight-nucleotide samplespecific index sequence and 18S rDNA/rRNA V4 region specific forward primer sequence of twenty nucleotides or reverse primer sequence of eighteen nucleotides (Manoharan et al. 2017). The PCR was performed in two stages: Stages 1: 94 ℃ for 3 min initial denaturation; ten cycles of 94 ℃: 10 s, 53 ℃: 30 s, 72 ℃: 40 s; Stage 2: 23 cycles of 94 ℃: 10 s, 48 ℃: 30 s, 72 .C: 40 s with a final extension at 72 ℃ for 2 min. PCR was performed in 100 µl reaction volume with 30 ng of template DNA. Amplified products (cDNA and gDNA amplicons) were purified by Qiaquick PCR purification kit (Qiagen, Netherlands) and mixed in equimolar concentrations before sequencing on Illumina MiSeq platform (IGA Technology Services, Italy). For each nucleic acid sample, two amplification reactions were performed as two technical replicates using different combination sets of forward and reverse primers (Table 2) producing total 24 amplicons representing twelve gDNA and twelve cDNA samples corresponding to six soil samples.

Table 2 Primer sets used for the amplification of V4 region of 18S DNA and 18S cDNA from different soil samples

S. No	Sample	SET1	SET2
1	gS1A	F1, R1	F10, R9
2	gS1B	F4, R3	F2, R2
3	gS2	F7, R6	F5, R4
4	cS1A	F6, R5	F8, R7
5	cS1B	F3, R3	F9, R9
6	cS2	F11, R6	F10, R6
7	gZn1	F3, R5	F5, R3
8	gZn2	F3, R6	F4, R1
9	gZnN	F1, R6	F10, R5
10	cZn1	F3, R7	F4, R2
11	cZn2	F3, R8	F4, R4
12	cZnN	F9, R11	F9, R8

Illumina MiSeq sequencing, sequence processing, and taxonomic analysis

Table 3 List of primers and their sequences used in this study

Forward Primer ID	Forward primer sequence	Reverse Primer ID	Reverse primer sequence
F1	NNNNacacacacCCAGCASCYGCGGTAATTCC	R1	NNNNgatcgcgaACTTTCGTTCTTGATYRA
F2	NNNNacagcacaCCAGCASCYGCGGTAATTCC	R2	NNNNcgctctcgACTTTCGTTCTTGATYRA
F3	NNNNgtgtacatCCAGCASCYGCGGTAATTCC	R3	NNNNgtcgtagaACTTTCGTTCTTGATYRA
F4	NNNNtatgtcagCCAGCASCYGCGGTAATTCC	R4	NNNNgtcacgtcACTTTCGTTCTTGATYRA
F5	NNNNtagtcgcaCCAGCASCYGCGGTAATTCC	R5	NNNNgactgatgACTTTCGTTCTTGATYRA
F6	NNNNtactatacCCAGCASCYGCGGTAATTCC	R6	NNNNagactatgACTTTCGTTCTTGATYRA
F7	NNNNactagatcCCAGCASCYGCGGTAATTCC	R7	NNNNgcgtcagcACTTTCGTTCTTGATYRA
F8	NNNNtgacatcaCCAGCASCYGCGGTAATTCC	R8	NNNNacgacgagACTTTCGTTCTTGATYRA
F9	NNNNacatgtgtCCAGCASCYGCGGTAATTCC	R9	NNNNcatcagtcACTTTCGTTCTTGATYRA
F10	NNNNgtacgactCCAGCASCYGCGGTAATTCC	R11	NNNNtctactgaACTTTCGTTCTTGATYRA
F11	NNNNatgatcgcCCAGCASCYGCGGTAATTCC		

Paired-end (PE) sequencing chemistry on the Illumina MiSeq platform was used, which produced raw reads of equal length from each end (2 . 250) of the amplicon. The NGS reads were assumed to be rDNA/rRNA derived as the amplification was executed using various sets of degenerate primers specific to the V4 region of 18S rDNA/rRNA. Both forward and reverse raw reads were first examined for their quality using FastQC (https:// www. bioin forma tics. babra ham. ac. uk) which includes analysis of GC bias, K-mer quality, and duplication levels. Afterward, datasets were processed using the Mothur software package (Schloss et al. 2009). The paired-end reads were merged to produce full-length amplicon sequences through the Mothur pipeline followed by sample-wise splitting and retrieval of sequences on basis of index sequence search. Further analyses of sequences were performed through the pipeline of the Parallel-Meta 3 suite (Jing et al. 2017). Parallel-Meta 3 processed and filtered sequences to produce the final dataset after ASV denoising to remove erroneous sequences generated during PCR and sequencing and chimera removal. A distance matrix based on the Meta-Storms scoring algorithm was generated, and sequences were clustered into operational taxonomic units (OTUs) at 3% dissimilarity cutoff or 97% similarity. Sequences were analyzed for phylotype identification and classified according to their taxonomic position. Most abundant unique sequences from each OTU were retrieved, and phylotype was deduced by aligning the reads to the SILVA 18S rRNA reference sequence database (https:// www. arb- silva. de/). The sequence similarity cutoff was set at 97% level, and the sequence alignment threshold was 0.99.

Statistical analysis of data

The taxa abundance table, community richness, and diversity indices were measured from OTU table by Parallel-Meta 3 suite through its built-in R Scripts packages. Sample-wise alpha diversity indices and statistical calculation of data produced by the Parallel-Meta 3 pipeline are used to generateplots by Origin Pro software. Multivariate statistical analysis for the study of β diversity was performed by calculating Meta-Storms Distances within and between the sample. Hierarchical clustering and heatmap analyses were based on pair-wised distance matrix scores as calculated by the Meta- Storms algorithm of all input samples and OTU profiles of multiple samples. PCA analysis and plot were generated in origin pro software on the basis of genus count data of all samples produced by Parallel-Meta 3. Marker analysis to find out key organisms in samples originating from PL and ZM locations was performed from the abundance table at the genus level.

Identification of metal-tolerant genes

Total RNA was isolated from 2 g soil, and cDNAs specific to eukaryotic mRNA were synthesized using the Mint-2 cDNA synthesis kit (Evrogen, Moscow, Russia) using polydT primer. The methods of RNA isolation, cDNA synthesis, size fractionation, and amplification of cDNAs were carried out as described in Yadav et al. (2014). cDNAs were ligated downstream of the *Saccharomyces cerevisiae* PGK1 promoter in a modified pFL61 yeast expression vector containing SfiIA and SfiIB sites (Minet et al. 1992). The cDNA libraries were transformed in Cd-sensitive *S. cerevisiae* mutant *ycf1*Δ by employing

the standard lithium acetate method (Gietz and Schiestl 2007). cDNA clones were screened for cadmium tolerance by functional complementation in *S. cerevisiae ycf1*Δ mutant (*Mat a; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0; YDR135::kanMX4*). Cadmium-tolerant cDNA clones were sequenced by the Sanger DNA sequencing method using vector-specific and internally designed primers. The sequences were compared for homologous sequences available in databases by using BLASTX. Few of the metal-tolerant cDNAs were assessed for their tolerance mechanisms by transforming them into metal-sensitive yeast mutants by functional complementation, and testing their abilities to tolerate high concentrations of metals such as Cd, Cu, Zn, and Co (Mukherjee et al. 2019a, b, 2021; Thakur et al. 2019, 2021).

Results and discussion

Diversity and composition of eukaryotic microbes in metal-contaminated soils

Change in soil physicochemical properties is linked to the change in the microbial diversity and composition. Similarly, the heavy metal pollution is also related to the change in the taxonomical structure of soil inhabiting microbes and its functional profile. These changes are also believed to be influenced by climatic conditions. Predicted heavymetal-contaminated soils from two geographically separated sites situated in two different climatic conditions were selected to study the variation in taxonomic diversity and composition of soil-inhabiting micro-eukaryotes. The physicochemical properties of the soil samples, including the level of different heavy metals, are presented in Table 1. The soil at PL was a sandy luvisol with average pH 7.06. The percentage organic carbon, total phosphorus, available phosphorus, and total nitrogen of this soil were 1.6%, 291 mg/kg, 14.2 mg/kg, and 0.12%, respectively. The soil at ZM was sandy ustochrepts with pH of 7.56, 2.27% organic carbon, 252.4 mg/kg total phosphorus, 1.92 mg/kg available phosphorus, and 0.18% total nitrogen. Both soils were found to be contaminated with high levels of heavy metals including toxic cadmium and lead. PL soil was also found contaminated with copper, while the ZM was found to be contaminated with lead.

To evaluate and quantify the eukaryotic microflora flourishing in these two soils, the V4 region of the 18S rRNA gene and its transcript was targeted in this study. Highquality total gDNA and RNA were extracted from all six soil samples. gDNA and cDNA synthesized from total soil RNA were subjected to targeted PCR to yield gDNA and cDNA-specific amplicons. Primers used for amplification were 5' linker sequence followed by an index sequence and an 18S V4 specific complementary sequence. The linker sequence increases the overall melting temperature, while the sample-specific index sequence was used for unveiling the sample-wise taxonomic diversity of soils. The reasons for selecting both rDNA and rRNA templates for targeted amplification were to analyze the specificity and compare the results obtained from these different types of nucleic acids. Twenty-four primer sets were used to amplify six genomic DNA and six cDNA templates in duplicates to constitute technical replicates (Tables 2 and 3). PCR with each primer set produced similar results as the amplified DNA products were ~ 380 bp in size. Earlier studies have exploited various regions of genes such as *rrn* operon (Guo et al. 2015), 18S and 23S regions, mitochondrial gene cytochrome c oxidase (Hebert et al. 2003), and rbcL gene from chloroplast to explore the microbial diversity studies. However, all these studies confined themselves to only a limited type of microbes in an environment (Vaulot et al. 2008; Eiler et al. 2013). It has been reported that the eukaryotic microbial community structures of complex environments like soil can be achieved by sequencing the V4 region of the 18S rDNA gene (Hugerth et al. 2014; Hadziavdic et al. 2014). Thus, the 18S rRNA gene is the most commonly preferred gene to study eukaryotic communities because it possesses numerous alternative hypervariable regions, i.e., V1-V9 and conserved regions. Numerous studies have explored different variable regions within the 18S rRNA gene and identified various primer combinations with the help of in silico sequence database coverage and taxonomic resolution. Also, these primers were validated for their feasibility with

environmental surveys and identified that V4 and V9 are the most efficient and frequently used regions (de Vargas et al. 2015; Stoeck et al. 2010; Lohan et al. 2016).

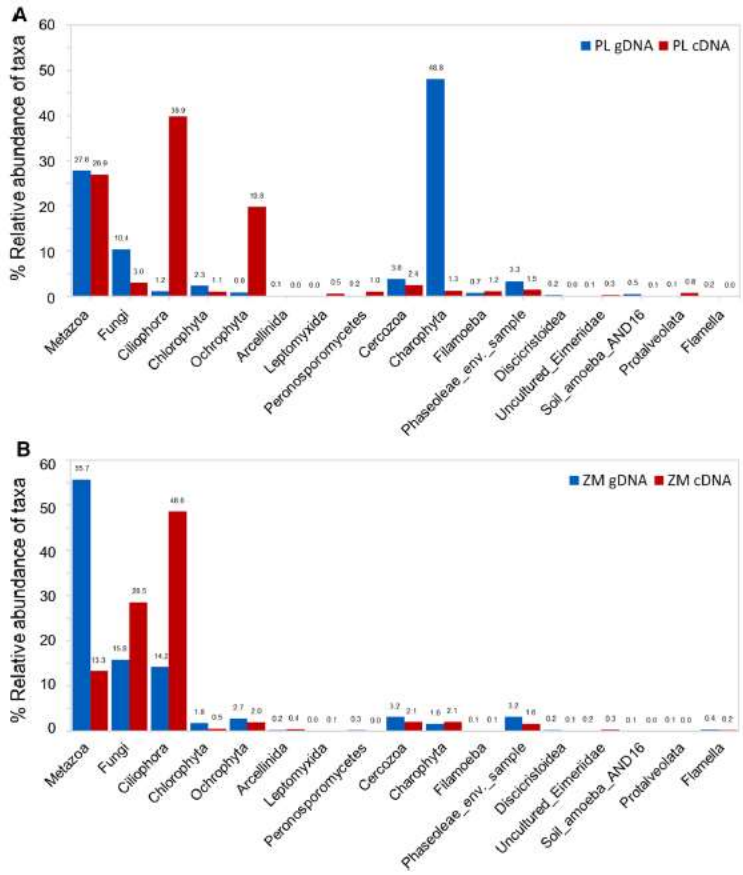
All the twenty-four PCR products were mixed in equimolar concentration and subjected to sequencing on the Illumina platform through paired-end sequencing chemistry. The degenerate primer sets used in this study were best suited to identify the eukaryotic taxa through 18S rDNA/ rRNA using read lengths of approximately 300 bp obtained by paired-end sequencing. Sequencing produced a vast amount of sequence data, which was further analyzed for the taxonomic origin of these sequences. A total of 10,28,789 sequences were generated after merging of paired-end reads from all the 24 amplicons. Out of these, 1,68,453 sequences were mapped to the 18S rRNA/rDNA reference dataset of SILVA after ASV filtration and chimera removal (Supp. Table S1). On average, 210 OTUs were generated per amplicon representing 12 cDNA and 12 gDNA samples. The number of average genera per sample was 140.

The predicted phylogenetic origin of sequence reads gave a wide range of biological insight into the diversity patterns of eukaryotes in the heavy metal polluted soil ecosystem. The taxa abundance at order level in both soil samples is illustrated in Fig. 1. It was observed that eukaryotic taxa actively maintaining homeostasis in metal-polluted soil of ZM and PL were dominated by metazoa and fungi. Fungi mainly constituted a substantial proportion of microflora in the soil, involving in the soil ecosystem's material cycle. Sequences originating from gDNA of PL soil show dominance of Charophyta. At genus level, 18S sequence amplified from gDNA sample from PL soil is dominated by a single genus *Populus*, a member of order Charophyta. The same genus is represented as a very minor taxon in cDNA amplicons of this soil. This may be due the reason that soil samples were collected from a *Populus* growing agroforestry land that contained functionally inactive microscopic components of the plant. These contributed comparatively a high proportion of specifically gDNA in the samples. In ZM samples, cDNA derived samples are showing similar taxa profile in all of the samples, while gDNA samples were showing inconsistency. These biases observed in samplesoriginating from the same location endorse the drawback of using gDNA in spite of RNA for targeted amplification of 18S sequences for study of the diversity of microbial eukaryotes in such cases. Data produced from the rRNA establish its efficacy and reliability in such type of studies.

Fungi mainly constituted a substantial proportion of microflora in the soil, involving in the soil ecosystem's material cycle. It also provides accurate information about actively proliferating taxa. The important phylum opisthokonta includes metazoa and fungi. Fungi mainly constituted a substantial proportion of microflora in the soil, involving in the soil ecosystem's material cycle. Along with these predominant phyla, both soils have also exhibited the presence of Amoebozoa (Fig. 1). Many studies have shown eukaryotic microbial communities from different environments. For instance, Tedersoo et al. (2014) reported the dominance of filamentous fungi and Amoebozoa in a soil environment compared to lakes, which are dominated by Ciliates and Chytrids. Studies exploring either 18S sequences amplified from soil DNA or soil cDNAs showed that soil hosts an incredible diversity of eukaryotes, all consistently showing the phyla metazoa and fungi (O'Brien et al. 2005; Lesaulnier et al. 2008; Urich et al. 2008). Damon et al. (2012) reported that the fungi contributed almost 40% of 18S rRNA soil sequences and 70% of the cDNA sequences. It has been reported that the eukaryotic microbial communities are affluent and active at the sampling sites based on the cDNA sample analysis. These results also suggest that various novel elements or species develop numerous mechanisms to combat metal toxicity in these sites. Different mechanisms generally contributed toward the microbial resilience to metals, such as Cd, Ni, Zn, Cu, Co, Pb, and As, involve the transfer of heavy metal tolerance genes by substitution across the metal sensitive strains and percentage of heavy metal bioavailability (Sobecky and Coombs 2009; Griffiths and Philippot, 2013).

Composition and diversity of the eukaryotic microbial community within the samples were analyzed. Alpha diversity of 18S rRNA/rDNA amplicon sets were calculated by Shannon, Simpson, and chao1 indices at the taxonomic genus level to examine the complexity of samples. Calculated indices were plotted against location and type of samples (Fig. 2A). The line within the box represents the median, while the whiskers represent the lowest and highest values within the interquartile range (IQR). Outliers are shown as empty dots. cDNA amplified communities feature, especially diverse taxonomical membership in both soil types, while that of gDNA revealed simple community structure. The microbial diversity was also compared across the different samples through beta-diversity analysis. Betadiversity estimates the similarity or distance between microbiome pairs and compares the overall taxonomic pattern of two different locations. Multivariate statistical analysis for the study of beta diversity was performed by calculating Meta-Storms Distances within and between the sample.

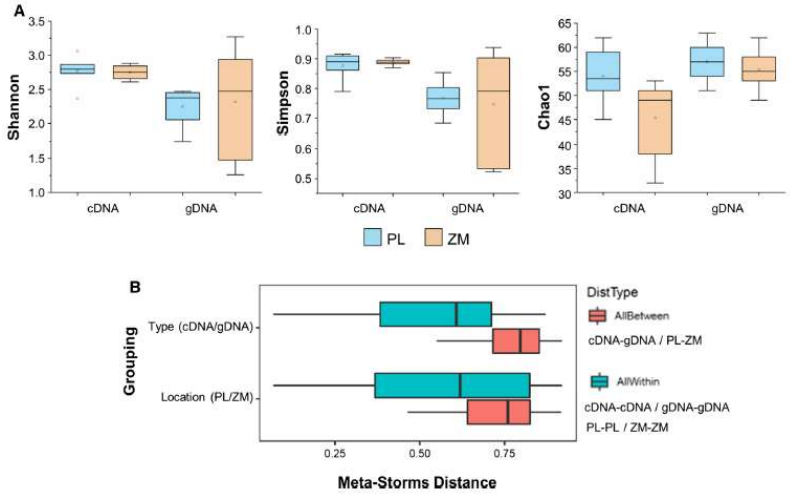
Fig. 1 Bar chart showing relative abundance of taxa at order level observed in PL **A** and ZM **B** soils. Taxa profiles corresponding to both gDNA and cDNA are shown



Pairwise analysis of distances between the datasets of the type of nucleic acid used for amplification and location of samples were performed. Dataset pairs of different type (cDNAgDNA) and location (PL-ZM) were found distant compared to dataset pairs of same type (gDNA-gDNA, cDNA-cDNA) and location (PL-PL, ZM-ZM) (Fig. 2B). Microbial communities observed between Indian and French soil are more diverse compared to samples coming from same source. Same is true for taxa observed from different nucleic acid. Location-wise difference in the taxonomic profiles indicates effect of climatic conditions, which not only affect the property of soil but also the composition of the taxa inhabiting it. Difference in the taxonomic profile originating from different type of nucleic acid is due to the presence of two different biotic components in the soil. Taxonomic profile that of gDNA is dependent on total including the inactive component.

Pair-wise distance matrix of all input samples and hierarchical clustering based on OTUs profiles of multiple samples were used to plot heatmap to compare the similarity of the sub-samples (Fig. 3B). The heat map result shows the highest degree of similarity among sub-samples of the same locations. Contrastively, the PL and ZM samples are highly dissimilar. This again proves that there is a substantial impact of the climate on the soil eukaryotic diversity and community structure. This also agrees with the PCA plot, wherein two components represent substantial differences among samples (Fig. 3A). PCA analysis was based on observed genus count in soil samples of two locations and type of template used for the study. Calculated principal components were plotted for different sample. It was observed that the datasets from two soils samples coming from different locations were clearly separated, while all the location-specific sub-samples were closely arranged. Abundant genera of each location were analyzed for biomarker identification. This analysis at the genus level suggested that each location is characterized and distinguished by the relative abundance of just a few taxa. For instance, taxa *Glomeromycetes, Leptopharynx, Glomerales, Bryometopus, Chlamydomyxa* in the total eukaryotic microbial community. The active marker genera of PL soil were *Balantidion, Polycostidae, Eimeriidae, Filamoeba, Epispathidium* and *Chrysophyte* (Supplementary Fig. S1). All these findings showed that various soil abiotic and biotic factors impacted the eukaryotic diversity flourishing in these geographically separated metal-polluted soils. Various global studies have confirmed that the majority of the eukaryotic microbes are not ubiquitous, and variations in the distribution of species illustrate strong bio-geographical outlines at large geographical scales (de Vargas et al. 2015; Talbot et al. 2014). For instance, saprophytic fungi present in soils of tropical biomes showed more species richness and a remarkable taxonomic difference from those present in soils of temperate biomes (Tedersoo et al. 2014). Bates et al. (2013) observed that environmental factors such as the availability of water, nutrients, and soils could also affect the taxonomic distribution among the eukaryotic microbial communities. Furthermore, a worldwide analysis of eukaryotic microbial communities by meta-barcoding also drew attention to a complete understanding of eukaryotic diversity and underpins the point that eukaryotic diversity is still not well described. For instance, a global survey of protist diversity in oceans observed ten times more OTUs than the number of already described marine planktonic species (de Vargas et al. 2015). All these data and factors help in understanding the microbial diversity in environments.

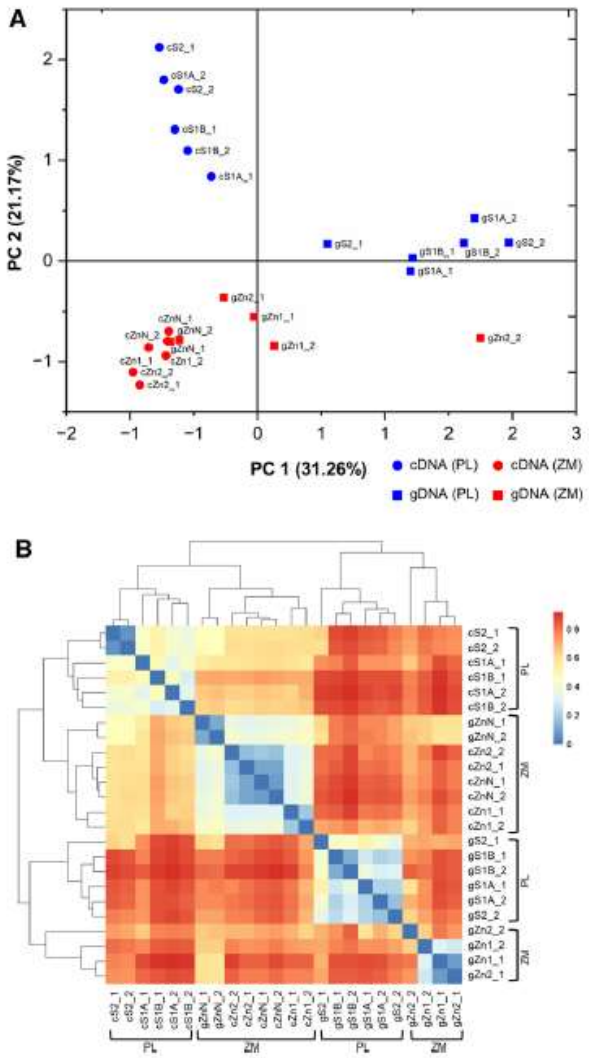
Fig. 2 Qualitative measure of alpha and beta diversity of eukaryotic microbial communities in soil. **A** : Shannon, Simpson and Chao1 index analysis for the measurement of alpha diversity within the sample. **B** : Beta diversity analysis of samples



Diversity of metal-tolerant eukaryotic genes expressed in contaminated soil The rRNA pool of a soil sample provides an indirect view of the functional status of the microbial communities flourishing in

the soil environment. Thus, to monitor numerous functional genes or active mechanisms, phenomena, or pathways among the eukaryotic soil community, exploring the total mRNA content of the soil sample was a necessary step. Thus, these soil samples were explored to mine the metal-tolerant eukaryotic genes by functional metatranscriptomic approach by constructing the size-fractionated cDNA libraries (Yadav et al, 2014). The sized eukaryotic cDNA libraries were screened for metal-tolerant genes by transforming the recombinant plasmid libraries into metal sensitive yeast mutants (Yadav et al. 2014). Out of 112 cDNA inserts, 38 cDNAs were identified as unique sequences, while others were either identified more than once or could not be read (Table 4). As shown in Table 4, these sequences reported their phylogenetic relationships among different eukaryotic lineages such as Protozoa, Ciliates, Fungi, Acanthamoeba, Choanoflagellate, Ciliophora, Nematode, Plantae, and Animalia. Some of these genes were selected randomly and characterized to provide metal tolerance. These genes include ubiquitin fusion protein, von Willebrand factor type D domain, heat shock protein type I like protein, serine protease inhibitor, aldehyde dehydrogenase, and DHHC palmitoyl transferase-like protein (Thakur et al. 2018, 2019, 2021; Mukherjee et al. 2019a, 2019b, 2020, 2021). These characterized cDNAs have shown high tolerance toward different heavy metals like Cd, Cu, Zn, and Co.

Fig. 3 PCA plot analysis based on observed genus count in samples of two locations and type of template used for the study of eukaryotic microbial diversity **A**. Hierarchical clustering and heatmap analysis ofdifferent samples **B**



Degradation of non-native proteins or proteolytic inactivation of regulatory proteins is a primary cellular response of living organisms in metal-contaminated environments. Various studies reported that Hsp40 in association with Hsp70 prevents aggregation of proteins and assists in refolding or degradation of non-native proteins under normal and stressed environments such as metal toxicity (Frydman 2001; Suzuki et al. 2001). In a similar study, Woo et al. (2009) observed elevated expression of ubiquitin encoding cDNA OjaUB to remove numerous aberrant proteins synthesized after exposure to different concentrations of Cd, Cu, Zn, Ni, Cr, and Ag. Similarly, cDNA PLCe10 encoding von Willebrand factor type D domain (VWD domain) like protein showed identity to VWD domain of vitellogenin, an egg yolk protein already been used as a biomarker for metal toxicity in an aqueous environment (Jubeaux et al. 2012). Furthermore, cDNA PLCg49 showing homology to serine protease inhibitors have conferred tolerance to all the hypersensitive yeast mutants. Protease inhibitors (PIs) have already been studied in plants with respect to various stress-inducible factors. They have emerged as an essential gene family reported to be highly expressed and positively responsible for modulating tolerance of plants against abiotic stresses (Shitan et al. 2007). The expression of PIs in response to abiotic stresses such as drought, salinity, heavy metals, UV radiation, and wounding suggests numerous interlinked metabolic processes that ultimately lead to their induction under adverse conditions. The DHHC palmitoyl transferase-like protein (cDNA PLCg39), the multi-metal tolerance phenotype conferred to the hypersensitive yeast mutants, was attributed to the presence of Zn fingers in the putative protein structure. Various responses toward metal stresses studied in zinc finger transcripts are positively correlated with an increase in the relative expression of Zn finger transcripts (Dai et al. 2007; Dixit and Dhankar 2011). These findings have concluded that both metagenomic analysis and functional metatranscriptomic analysis of soil samples support and validate each other's results and give a complete picture of taxonomic diversity and functional diversity with numerous novel gene sequences active at the time of sampling. This study also summarized the variation in soil eukaryotes' total and functional diversity at the time of sampling, which strongly highlights the presence of numerous unknown, undescribed phenomena and factors which are actively working toward rejuvenating such polluted soils.

Conclusion

As the metagenomic approach exposes both active and inactive affiliates of microbial communities present in an environment, the actual activity and metabolic role of individual members of a microbiota flourishing in the environment remains unknown. In this work, combined 18S rDNA/18S rRNA-based metagenomics was employed to explore active eukaryotic communities thriving in a metal-contaminated soil environment. This work showed the effects of potentially toxic metals on soil eukaryotic diversity by targeting V4 hypervariable regions of 18S rRNA/rDNA through amplicon sequencing. Amplicons sequence analysis of both the sites, i.e., French soil site PL and Indian soil site ZM, reported that the majority of the significant eukaryotic phylum contributed actively to maintain the homeostasis in metal-polluted soil sites but their genera diversity and composition is different. Further, the higher expression level of genes corresponding to eukaryotic microbial communities active at the time of sampling suggested that these genes play important roles in metal homeostasis. Furthermore, it was observed that such contaminants and climatic conditions in the soil affect the eukaryotic diversity. These findings showed that microbial diversity and community structure are significantly influenced due to location specific conditions, which significantly change the ecological environment of soil. Marker organisms and expressed eukaryotic genes identified in this study can be helpful to characterize similar conditions in the environment.

Table 4 Details of cadmium-tolerant cDNA clones identified through functional metatranscriptomic analysis of the soil

cDNA ID	Clone Name	Seq length	Annotation BLASTX/Swiss Prot	E value	Organism NR
PLCc37	*ycf1*ΔPLCc37	1342	No significant similarity found	2e-34	*Oxytricha trifallax*
2PLCc38	*ycf1*ΔPLCc38	1667	Aldehyde dehydrogenase (ALDH) family 7 member of	3e-151	*Pseudocohnilembus persalinus*
PLCc43	*ycf1*ΔPLCc43	1039	DNJA2_MOUSE RecName: Full = DnaJ homolog subfamily A member 2	2.00E-103	*Acanthamoeba castellanii*
PLCc52	*ycf1*ΔPLCc52	1513	Hypothetical protein	0.77 ND	not determined)
PLCc53	*ycf1*ΔPLCc53	928	Hypothetical protein	0.006	ND
PLCc62	*ycf1*ΔPLCc62	1041	MIOX4_ARATH RecName: Full = Inositol oxygenase 4	0	*Populus*
PLCd20	*ycf1*ΔPLCd20	1180	No significant similarity found	NA	NA
PLCd27	*ycf1*ΔPLCd27	907	Hypothetical protein	8.7	ND
PLCd31	*ycf1*ΔPLCd31	1312	No hit found	3.00E-64	NA
PLCd37	*ycf1*ΔPLCd37	1323	Hypothetical protein	3.00E-85	*Neofusicoccum parvum*
PLCd39	*ycf1*ΔPLCd39	1184	Hypothetical protein	ND	*Drosophila ananassae*
PLCd43	*ycf1*ΔPLCd43	1392	Gamma glutamylcysteinyl transferase [*Arabidopsis thaliana*]		*Arabidopsis thaliana*
PLCd49	*ycf1*ΔPLCd49	1737	LAMB1_HUMAN RecName: Full = Laminin subunit beta-1	1.00E-09	*Paramecium tetraurelia*
PLCd56	*ycf1*ΔPLCd56	1479	Hypothetical protein 4.4 ND		
PLCe10	*ycf1*ΔPLCe10	1234	VIT1_CAEEL RecName: Full = Vitellogenin-1	3.00E-35	*Ascaris suum*
PLCe11	*ycf1*ΔPLCe11	1157	Hypothetical protein	NA	NA
PLCe13	*ycf1*ΔPLCe13	1002	Hypothetical protein 4.1		ND
PLCe38	*ycf1*ΔPLCe38	1126	sp|Q6DE96.1|IWS1A_XENLA Rec-Name: Full = Protein IWS1 homolog A; AltName: Full = IWS1-like	4.00E-23	*Citrus sinensis*
PLCe42	*ycf1*ΔPLCe42	1345	EFTU_RECAM RecName: Full = Elongation factor Tu, mitochondrial	0	*Andalucia godoyi*
PLCf17	*ycf1*ΔPLCf17	1554	PP2C_LEICH RecName: Full = Protein phosphatase 2C	7.00E-21	*Monosiga brevicollis MX1*
PLCf19	*ycf1*ΔPLCf19	1203	No significant similarity found	NA	NA
PLCf26	*ycf1*ΔPLCf26	1442	No hit found	NA	NA
PLCg9	*ycf1*ΔPLCg9	1191	PSD3_SCHPO RecName: Full = Phosphatidylserine decarboxylase proenzyme 3	2.00E-159	*Reticulomyxa filosa*
PLCg39	*ycf1*ΔPLCg39	1355	DHHC zinc finger domain containing protein	1e-59	*Acanthamoeba castellanii*
PLCg49	*ycf1*ΔPLCg49	1409	Serine Protease Inhibitor (serpin) family	1e-67	*Hypsibius dujardini*
PLCg52	*ycf1*ΔPLCg52	1212	GL12_ARATH RecName: Full = Putative germin-like protein subfamily 1 member	8.00E-18	*Populus trichocarpa*
PLCg56	*ycf1*ΔPLCg56	1272	sp|Q941D6.1|HDA14_ARATH Rec-Name: Full = Histone deacetylase 14 [Arabidopsis thaliana] homologue a HDA1 de *S. cerevisiae* avec une e value de 1e-47	1e-114 (Min)	*Nannochloropsis gaditana*
PLCg62	*ycf1*ΔPLCg62	1134	No significant similarity found	NA	NA
PLCg71	*ycf1*ΔPLCg71	1669	Hypothetical protein	3e-13	*Trichoderma harzianum*
PLBa11	*ycf1*ΔPLBa11	512	No significant similarity found	NA	NA
PLBa15	*ycf1*ΔPLBa15	573	Transcription factor (Conus magus)	9.00E-29	*Conus magus*
PLBb10	*ycf1*ΔPLBb10	806	Senescence-associated protein [Gossypium australe]	8.00E-36	*Gossypium australe*
PLBb18	*ycf1*ΔPLBc18	807	No significant similarity found	NA	NA
PLBe1	*ycf1*ΔPLBe1	694	Ubiquitin domain containing protein [Acanthamoeba castellanii str. Neff]	1.00E-134	*Acanthamoeba castellanii*
PLBe6	*ycf1*ΔPLBe6	681	Reticulocyte binding protein homologue 1 (RH1), partial [*Plasmodium ovale curtisi*]	1.00E-51	*Plasmodium ovale curtisi*
PLBf11	*ycf1*ΔPLBf11	811	Putative reverse transcriptase [*Zingiber officinale*]	5.00E-20	*Zingiber officinale*
PLAe2	*ycf1*ΔPLAe2	324	2-amino-4-hydroxy-6-Hydroxymethyldihydropteridine diphosphokinase [*Algoriphagus* sp. F21]	7.6	*Algoriphagus* sp. F21
PLAg22	*ycf1*ΔPLAg22	489	Hypothetical protein BVRB_022340 [*Beta vulgaris* subsp. *vulgaris*]	9.00E-10	*Beta vulgaris* subsp. *vulgaris*

Supplementary Information The online version contains supplementary material available at https:// doi. org/ 10. 1007/ s13762- 022- 04635-5. Acknowledgements The authors are thankful to Indo-French Centre for the Promotion of Advanced Research, New Delhi, for financial support under project No. 4709-1.

Author contributions BT and RKY performed all the sequencing experiments. BT analyzed the NGS data, wrote the manuscript, and. PS and KMM helped in the acquisition of the data and contributed to the data analysis. MSR, RM, and LFT conceived the study and designed experiments and manuscript revision. MSR and LFT, the principal investigator for the grant and for constructing the experimental design, supervised the project and helped to write the manuscript. All authors contributed to the article and approved the submitted version. Funding This research received a grant from Indo-French Centre for the Promotion of Advanced Research, New Delhi, under project No. 4709–1.

Data availability The authors declare that all data supporting the findings of this study are available within this article and the sequence information in the NCBI database.

References

Arkadeep Mukherjee M, Reddy S (2020) Metatranscriptomics: an approach for retrieving novel eukaryotic genes from polluted and related environments. 3 Biotech. https:// doi. org/ 10. 1007/ s13205- 020- 2057-1

Bailly J, Fraissinet-Tachet L, Verner MC, Debaud JC, Lemaire M, Wesolowski-Louvel M et al (2007) Soil eukaryotic functional diversity, a metatranscriptomic approach. ISME J 1:632–642

Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R et al (2013) Global biogeography of highly diverse protistan communities in soil. ISME J 7:652–659

Dai X, Xu Y, Ma Q, Xu W, Wang T, Xue Y et al (2007) Overexpression of an R1R2R3 MYB gene, OsMYB3R-2, increases tolerance to freezing, drought, and salt stress in transgenic *Arabidopsis*. Plant Physiol 143:1739–1751

Damon C, Lehembre F, Oger-Desfeux C, Luis P, Ranger J, Fraissinet- Tachet L, Marmeisse R (2012) Metatranscriptomics reveals the diversity of genes expressed by eukaryotes in forest soils. PLoS ONE 7(1):e28967. https:// doi. org/ 10. 1371/ journ al. pone. 00289 67

De Vargas C, Audic S, Henry N, Decelle J, Mahe F, Lagos, (2015) Eukaryotic plankton diversity in the sunlight ocean. Science 348:1261605

Dixit AR, Dhankher OP (2011) A novel stress-associated protein 'AtSAP10'from Arabidopsis thaliana confers tolerance to nickel, manganese, zinc, and high temperature stress. PLoS ONE 6:e20921

Eilers KG, Lauber CL, Knight R, Fierer N (2010) Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil. Soil Biol Biochem 42:896–903

Eiler A, Drakare S, Bertilsson S, Pernthaler J, Peura S et al (2013) Unveiling distribution patterns of freshwater phytoplankton by a Next Generation Sequencing based approach. PLOS ONE 8:e53516

Fierer N, Bradford MA, Jackson RB (2007) Toward an ecological classification of soil bacteria. Ecology 88:1354–1364

Frydman J (2001) Folding of newly translated proteins in vivo: the role of molecular chaperones. Annu Rev Biochem 70:603–647

Gans J, Wolinsky M, Dunbar J (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. Science 309:1387–1390

Gietz RD, Schiestl RH (2007) High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc 2:31–34

Griffiths BS, Philippot L (2013) Insights into the resistance and resilience of the soil microbial community. FEMS Microbiol Rev 37:112–129

Guo L, Sui Z, Zhang S, Ren Y, Liu Y (2015) Comparison of potential diatom "barcode" genes (18S and ITS rDNA, COI, *rbcL*) and their effectiveness in discriminating and determining species taxonomy in Bacillariophyta. Int J Syst Evol Microbiol 65:1369–1380

Hadziavdic K, Lekang K, Lanzen A (2014) Characterization of the 18S rRNA gene for designing universal eukaryote specific primers. PLoS ONE 9:e87624

Hebert PD, Ratnasingham S, De Waard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc B 270:S96–S99

Hugerth LW, Muller EE, Hu YO, Lebrun LA, Roume H, Lundin D (2014) Systematic design of 18S rRNA gene primers for determining eukaryotic diversity in microbial consortia. PLoS ONE 9:e95567

Jing G, Sun Z, Wang H, Gong Y, Huang S, Ning K, Xu J (2017) Su X (2017) Parallel-META 3: Comprehensive taxonomical and functional analysis platform for efficient comparison of microbial communities. Sci Rep 12(7):40371

Jubeaux G, Audourd-Combe F, Simon R, Tutundijn R, Salvador A, Geffard O et al (2012) Vitellogenin- like proteins among invertebrate species diversity: potential of proteomic mass spectrometry for biomarker development. Environ Sci Technol 46:6315–6323

Kitson RE, Mellon MG (1954) Colorimetric determination of phosphorus as molybdivanadophosphoric acid. Ind Eng Chem 16:379–383

Lehembre F, Doillon D, David E, Perrotto S, Baude J, Foulon J et al (2013) Soil metatranscriptomics for mining eukaryotic heavy metal resistance genes. Environ Microbiol 15:2829–2840

Leimena MM, Ramiro-Garcia J, Davids M, van den Bogert B, Smidt H, Smid EJ et al (2013) A comprehensive metatranscriptome analysis pipeline and its validation using human small intestine microbiota datasets. BMC Genomics 14:530

Lesaulnier C, Papamichail D, McCorkle S, Ollivier B, Skiena S et al (2008) Elevated atmospheric CO2 affects soil microbial diversity associated with trembling aspen. Environ Microbiol 10:926–941

Lohan KP, Fleischer RC, Carney KJ, Holzer KK, Ruiz GM (2016) Amplicon-based pyrosequencing reveals high diversity of protistan parasites in ships' ballast water: implications for biogeography and infectious diseases. Microbial Ecol 71:530–542

Manoharan L, Kushwaha SK, Ahren D, Hedlund K (2017) Agricultural land use determines functional genetic diversity of soil microbial communities. Soil Biol Biochem 115:423–432

Marmeisse R, Kellner H, Fraissinet-Tachet L, Luis P (2017) Discovering protein-coding genes from the environment: Time for the eukaryotes? Trends Biotechnol 35:824–835

Mico C, Recatala L, Peris M, Sanchez J (2006) Assessing heavy metal sources in agricultural soils of a European Mediterranean area by multivariate analysis. Chemosphere 65:863–872

Minet M, Dufour ME, Lacroute F (1992) Complementation of *Saccharomyces cerevisiae* auxotrophic mutants by *Arabidopsis thaliana* cDNAs. Curr Genet 2:417–422

Mukherjee A, Yadav R, Marmeisse R, Fraissinet-Tachet L, Reddy MS (2019a) Heavy metal hypertolerant eukaryotic aldehyde dehydrogenase isolated from metal contaminated soil by metatranscriptomics approach. Biochimie 160:183–192

Mukherjee A, Yadav R, Marmeisse R, Fraissinet-Tachet L, Reddy MS (2019b) Detoxification of toxic heavy metals by serine protease inhibitor isolated from polluted soil. Int Biodeterior Biodegr 143:104718

Mukherjee A, Thakur B, Pandey AK, Marmeisse R, Fraissinet-Tachet L, Reddy MS (2021) Multi-metal tolerance of DHHC palmitoyl transferase-like protein isolated from metal contaminated soil. Ecotoxicology 30:67–79

O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R (2005) Fungal community analysis by large-scale sequencing of environmental samples. Appl Environ Microbiol 71:5544–5550

Olsen SR (1954) Estimation of available phosphorus in soils by extraction with sodium bicarbonate. United States Department of Agriculture, Washington

Piper CS (1966) Soil and plant analysis; A laboratory manual of methods for the examination of soils and the determination of the inorganic constitutents of plants. Hans Publishers; Bombay

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464:59

Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, Brietbart M et al (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4:739

Rondon MR, August PR, Bettermann AD, Brady SF, Grossman TH, Liles MR et al (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl Environ Microbiol 66:2541–2547

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB et al (2009) Introducing mothur: open-source, platformindependent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol 75:7537–7541

Shitan N, Horiuchi KI, Sato F, Yazaki K (2007) Bowman-Birk proteinase inhibitor confers heavy metal and multiple drug tolerance in yeast. Plant Cell Physiol 48:193–197

Sobecky PA, Coombs JM (2009) Horizontal gene transfer in metal and radionuclide contaminated soils. Methods Mol Biol 532:455–472

Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, Breiner HW et al (2010) Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19:S21–S31

Suzuki N, Koizumi N, Sano H (2001) Screening of cadmium responsive genes in *Arabidopsis thaliana*. Plant Cell Environ 24:1177–1188

Talbot JM, Bruns TD, Taylor JW, Smith DP, Branco S, Glassman SI et al (2014) Endemism and functional convergence across the North American soil mycobiome. Proc Natl Acad Sci USA 111:6341–6346

Tedersoo L, Bahram M, Polme S, Koljalg U, Yorou NS, Wijesundera R (2014) Global diversity and geography of soil fungi. Science 346:1256688

Thakur B, Yadav R, Fraissinet-Tachet L, Mermeisse R, Reddy MS (2018) Isolation of multi-metal tolerant ubiquitin fusion protein from metal polluted soil by metatranscriptomic approach. J Microbiol Methods 152:119–125

Thakur B, Yadav R, Vallon L, Marmeisse R, Fraissinet-Tachet L, Reddy MS (2019) Multi-metal tolerance of von Willebrand factor type D domain isolated from metal contaminated site by metatranscriptomics approach. Sci Total Environ 661:432–440