Vertical profiles of microbial trace element utilization in marine environments

Yinzhen Xu[1,2,3,4], Qiong Liu[1,2,4], Yan Zhang[1,2,4,*]

[1] Shenzhen Key Laboratory of Marine Bioresources and Ecology, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518055, Guangdong Province, P. R. China

[2] Key Laboratory of Optoelectronic Devices and Systems of Ministry of Education and Guangdong Province, College of Optoelectronic Engineering, Shenzhen University, Shenzhen, 518060, Guangdong Province, P. R. China

[3] College of Food Engineering and Biotechnology, Hanshan Normal University, Chaozhou, 521041, Guangdong Province, P. R. China

[4] Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions, Shenzhen, 518055, Guangdong Province, P. R. China

* Corresponding author: Yan Zhang**,** College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518055, Guangdong Province, P. R. China. Tel: 86-755-26922024; Fax: 86-755-86713951**;** E-mail: zhangyan@szu.edu.cn

Abstract

Background: Biological trace elements are used in small amounts but are required for all living organisms. They are key components of many proteins and enzymes involved in important biological processes. Many trace element-dependent proteins have been characterized in various microbes, but very little is known about their occurrence, functions, and interactions in microbial communities in marine environments, especially in depth-related marine ecosystems.

Results: In this study, by analyzing metagenomic data from different geographic locations and water depths in the oceans around the world, we identified the distribution of genes encoding trace element-dependent proteins (for copper, molybdenum, tungsten, cobalt, nickel, and selenium) in a variety of marine samples from the upper ocean to the deep sea, which demonstrates vertical patterns of trace element utilization in marine microbes. More than 63,000 metalloprotein and selenoprotein genes belonging to nearly 100 families were predicted, constituting the largest environmental metalloprotein and selenoprotein gene dataset reported so far. Further examination of the interactions among trace elements revealed significant correlations between some of them (especially molybdenum or tungsten and selenium) and more active elemental crosstalk in the epipelagic zone of the ocean. Comparison of the patterns of trace element utilization across samples suggested that additional unknown factors might play a more important role in shaping trace element utilization in marine microbes living in certain locations. Finally, analysis of the relationship

between water depth and metalloprotein/selenoprotein families revealed that the evolution of approximately half of the metalloprotein and selenoprotein families in marine microbial world could be influenced by ocean depth at either the global or the local level.

Conclusions: Our findings provide new insights into the utilization and functions of trace elements in marine microbes along a vertical gradient across the ocean.

Keywords

Trace element; Marine metagenome; Metalloprotein; Selenoprotein; Depth; Evolution

Background

Biological trace elements are required by all organisms in very small quantities and their excess or insufficiency may interfere with various metabolic and physiologic processes of organisms [1-3]. These micronutrients (mostly metals) include iron (Fe), zinc (Zn), copper (Cu), manganese (Mn), nickel (Ni), cobalt (Co), molybdenum (Mo), tungsten (W), chromium (Cr), selenium (Se), and several other elements. Utilization of trace elements is generally rather complex. Fe and Zn are essential to nearly all organisms [4, 5]; however, the utilization of many other elements is quite diverse and scattered throughout all domains of life [6, 7]. Metal ions are usually required as cofactors for the assembly of metalloproteins. The majority of them are directly incorporated into their cognate binding sites on proteins. A small number of metals have to become part of prosthetic groups or cofactors prior to their insertion into target proteins, such as Mo and Co that are mainly used in the forms of molybdopterin (Mo cofactor) and vitamin B_{12} (cobalamin), respectively [8, 9]. Se, one of the most important metalloid elements, is involved in a variety of redox reactions in many organisms [10]. It mainly exists as selenocysteine (Sec) which is a nonstandard amino acid encoded by TGA (normally a stop codon) in all selenoproteins [11, 12].

In order to understand the roles that trace elements play in biological systems, it is necessary to identify genes encoding trace element-dependent proteins and their functions. To date, a large number of metalloproteins have been characterized in

various organisms [13]. The amounts of metalloprotein families differ greatly depending on which metal is used. For example, Zn is estimated to be used by hundreds of protein families [14]; however, less than ten metalloprotein families are known to be dependent on Mo or Ni [15, 16]. In addition, more than 70 selenoprotein families have been reported in a wide range of prokaryotes and eukaryotes, many of which were discovered by reliable bioinformatics tools [17, 18].

The past several years have witnessed a dramatic increase in the number of genome sequencing projects, which provides an enormous amount of genomic data for studying the metabolism and functions of trace elements and their evolutionary trends. Previously, several computational and comparative genomic studies have been carried out to examine different trace element utilization traits in sequenced prokaryotes and eukaryotes, which showed that the utilization of most trace elements is much more active in bacteria than in eukaryotes and that large metalloproteomes (the complete set of metalloproteins) for several metals (such as Cu, Mo, and Co) and selenoproteomes (the set of all selenoproteins) may associate with aquatic lifestyle [16, 19-24]. However, it is not yet clear whether or how different types of aquatic environments have influenced the use of trace elements.

Marine microbes have evolved for millions of years and are considered to be the most diverse organisms in the ocean, most of which have not been well studied in the

laboratory. Nowadays, the rapidly emerging area of microbial metagenomics enables scientists to explore species composition and diversity in a host of challenging environments including marine ecosystems. Several metagenome-based studies have been conducted to investigate the utilization of certain trace elements in different oceanic environments [25-29]. For example, analyses of selenoprotein genes and Fe uptake genes in surface ocean samples have provided a first glance at the utilization and roles of these elements in marine microbial communities [25-27]. Very recently, we performed a comparative metagenomic analysis of multiple trace elements using the metagenomic data from the Global Ocean Sampling (GOS) expedition (the largest and geographically most comprehensive marine metagenomic dataset), and found that the evolution of genes encoding trace element-dependent proteins could be affected by several aquatic environmental factors such as temperature and sample depth (unpublished data). Considering that GOS metagenomic data were derived from surface water bacterioplankton communities (mostly <5 m), it is essential to investigate the vertical profiles of trace element utilization in marine microbes over a much wider range of depths.

In this project, we analyzed the utilization of Cu, Mo, W, Ni, Co, and Se in marine microbial ecosystems by using multiple metagenomic datasets from seawater samples from different geographic locations and water depths. All known metalloprotein and selenoprotein genes were identified in each sample, and

metalloprotein/selenoprotein-rich and -poor samples were also predicted. Moreover, we investigated the interactions among trace elements and the patterns of trace element utilization across all samples. Finally, the influence of depth on the evolution of metalloprotein and selenoprotein families was also assessed. These data offer new insights into the general trends of trace element utilization and evolution in marine microbial world.

Results

Based on the metagenomic data derived from 66 marine samples from different locations around the world (details of the samples are shown in Additional file 1: Table S1; geographic locations of these samples are shown in Additional file 2: Figure S1), we generated the first map illustrating vertical patterns of trace element utilization in marine microbial communities. These samples spanned a large depth gradient of 10 to 3000 m, which belong to epipelagic (<200 m), mesopelagic (≥200 m and ≤1000 m), and bathypelagic (>1000 m and <4000 m) zones of the world oceans. A total of 63,697 metalloprotein and selenoprotein genes were predicted, which is so far the largest dataset of genes encoding trace element-dependent proteins (all predicted metalloprotein and selenoprotein sequences are stored in Additional file 3).

General analysis of metalloprotein genes and metalloproteomes

We analyzed all known metalloproteins and identified a large number of metalloprotein genes for Cu, Mo, W, Ni, and Co in marine microbial communities from different locations (Table 1). An overall view of the distribution of metalloproteomes in each sample is shown in Figure 1. A more clear illustration of metalloprotein-rich and -poor samples from each project according to the depths of the ocean is shown in Figure 2.

Copper

Cu is a vital metal for several metabolic enzymes involved in many biological processes. In this study, we identified 6,336 genes belonging to twelve known prokaryotic cuproprotein families (details are shown in Additional file 1: Table S2; the fraction of cuproproteome in each sample is shown in Additional file 2: Figure S2). Cytochrome *c* oxidase subunit I (COX I) and COX II are the most abundant cuproprotein families, which were detected in all examined metagenomic samples. In addition, genes encoding plastocyanin, nitrite reductase, Cu-Zn superoxide dismutase, azurin, and NADH dehydrogenase 2 were also observed in the majority of these samples. In contrast, only few sequences were found for Cu amine oxidase, nitrous oxide reductase, and tyrosinase, suggesting that these cuproproteins may be rarely used by marine microbes in these samples.

We found 7 cuproprotein-rich and 8 cuproprotein-poor samples based on the identification of the cuproproteome of each sample (Figure 2). Except for the sample collected at a depth of 1000 m in the South China Sea (SCS_1000m), all cuproprotein-rich samples were located at depths of less than 50 m (mostly at 10 m) at several Red Sea water column stations (RS_Stations) where no cuproprotein-poor sample could be detected. In contrast, the cuproprotein-poor samples were basically distributed at depths of more than 50 m (mostly ≥200 m in the mesopelagic zone) in other geographically distant areas. Considering that the dissolved oxygen concentration decreases with increasing of the sea water depth, our results are consistent with previous findings showing that molecular oxygen might promote Cu utilization in various microorganisms [19, 30]. Interestingly, all samples obtained from different depths of the Baltic Sea (BS_10m, BS_80m, and BS_400m) were cuproprotein-poor samples, suggesting a generally restricted utilization of this transition metal in that region.

Molybdenum and tungsten

Mo is a critical component of the active site of a number of molybdoproteins that catalyze key reactions in the global carbon, sulfur, and nitrogen metabolism. The currently known molybdoprotein families include dimethylsulfoxide reductase (DMSOR), sulfite oxidase (SO, including recently characterized MOSC-containing protein), xanthine oxidase (XO), aldehyde:ferredoxin oxidoreductase (AOR), and

Fe–Mo-containing nitrogenase [8, 31]. In a few prokaryotes, Mo is replaced by W to form tungstoproteins, which include nearly all enzymes of the AOR family and certain members of the DMSOR family in anaerobic bacteria and methanogenic archaea [32, 33]. In this study, only sequences of the AOR family were considered as tungstoproteins.

A total of 28,610 molybdoprotein genes were identified (Additional file 1: Table S3; the fraction of molybdoproteome in each sample is shown in Additional file 2: Figure S3). XO, DMSOR, and SO families could be detected in all or almost all samples. XO and DMSOR were the most abundant molybdoprotein families (50.3% and 35.9% of all molybdoprotein sequences, respectively). This is consistent with previous observations that organisms possessing members of DMSOR and XO families favor aerobic and aquatic living conditions [16]. In contrast, only one nitrogenase gene sequence could be found in these samples, implying that this enzyme is not essential for almost all marine bacteria.

We identified 5 molybdoprotein-rich and 7 molybdoprotein-poor samples. All molybdoprotein-rich samples were located in the mesopelagic zone, including three samples from the Hawaii open ocean area (HOT_500m, HOT_770m, and HOT_1000m) and two samples from RS_Stations (RS_Station34_258m and RS_Station192_500m). On the other hand, all molybdoprotein-poor samples were

obtained between 10 m and 110 m in the epipelagic zone of different locations. These data indicate that microorganisms living in deeper layers of the ocean may have a more active utilization of Mo.

With regard to tungstoproteins, we identified 958 AOR sequences in 57 samples, whose distribution showed large variations across samples (Additional file 1: Table S4; the fraction of tungstoproteomes in each sample is shown in Additional file 2: Figure S4). Fourteen tungstoprotein-rich and 32 tungstoprotein-poor samples were predicted. Similar to the utilization of molybdoproteins, More than 90% tungstoprotein-rich samples were located in the mesopelagic zone while almost all tungstoprotein-poor samples were collected at depths of less than 100 m from different locations. Surprisingly, only a single AOR gene was detected in all SCS samples, implying that W is basically not required for the marine microbes living in this area.

Nickel

Ni acts as a cofactor for a small number of metalloenzymes catalyzing key reactions in energy and nitrogen metabolism and in detoxification processes [34]. Here, we detected 4,640 genes encoding Ni-dependent proteins (Additional file 1: Table S5; the distribution of Ni-dependent metalloproteomes is shown in Additional file 2: Figure S5). Urease was the most frequently used Ni protein, which was detected in almost all samples and accounted for 76.5% of all Ni-dependent protein sequences. In contrast,

only few genes responsible for carbon monoxide dehydrogenase (CODH) and bifunctional CODH/acetyl-coenzyme A synthase (CODH/ACS) could be detected in 1 or 2 samples, suggesting that they are not essential for the majority of marine microorganisms. No methyl-coenzyme M reductase could be found in the whole marine metagenomic dataset, which is consistent with previous observation that this enzyme catalyzing the methane-forming step of the methanogenesis pathway is a methanogen-specific Ni-dependent protein [35].

By identifying Ni-dependent metalloproteomes for different samples, 11 Ni-protein-rich and 7 Ni-protein-poor samples were found. Similar to the distribution of cuproproteomes, approximately three-fourths of Ni-protein-rich samples are located at depths of 10 to 50 m whereas 71.4% Ni-protein-poor samples were collected at much deeper depths (100~3000 m). Thus, it seems that increased depth may inhibit the evolution of Ni protein-coding genes.

Cobalt

Co is an essential metal constituent of cobalamin (or called vitamin B_{12}) which is a complex organometallic cofactor needed for a variety of B_{12}-dependent enzymes, such as methylmalonyl-CoA mutase (MCM), B_{12}-dependent class II ribonucleotide reductase (RNR II), methionine synthase (MetH), dehalogenase, and some other important enzymes [36-39]. In this study, we identified 17,605 Co-dependent protein

genes (Additional file 1: Table S6; the distribution of Co-dependent metalloproteomes is shown in Additional file 2: Figure S6). Surprisingly, the majority (84.9%) of these sequences belonged to RNR II, whose number was ten times as many as that of the second most common Co protein family (MCM, 8.4%). As our recent analysis of surface ocean microbes has revealed that almost half of RNR II gene sequences originated from viruses (unpublished data), we then examined possible taxonomic compositions of these RNR II sequences and found that 19.2% of them might have a viral origin (Additional file 2: Figure S7). Moreover, the majority of viral RNR II sequences were detected in the epipelagic zone (10~110 m) of different locations, suggesting that viral-mediated, Co-dependent nucleotide biosynthesis may play a more important role in generating microbial diversity in upper layers of the ocean.

We identified 8 Co-protein-rich and 11 Co-protein-poor samples in current metagenomic dataset. The Co-protein-rich samples were located in the epipelagic zone of several RS_Stations. In contrast, the Co-protein-poor samples were collected from other distant locations of the ocean (except Red Sea) with a very wide range of water depth (10~3000 m), including all samples from the Sargasso Sea (SS) and the South China Sea. Although a strong relationship between Co and Ni utilization traits in prokaryotes has been previously reported [20], both Co-protein-rich and Co-protein-poor samples were located at various depths of different oceanic areas,

implying a different effect of depth gradient on Co-dependent metalloproteomes in marine microbes when compared with Ni.

Analysis of selenoprotein genes and selenoproteomes

Se is a necessary trace element for many vital physiological functions in a variety of organisms, which exerts its biological role through selenoproteins [10-12]. It has been suggested that the number of selenoprotein families detected in bacteria is much larger than that in eukaryotes or archaea [17, 20, 22]. Although aquatic habitat has been thought to promote the evolution of new selenoprotein genes in microorganisms [22, 25], the relationship between depth and the distribution of selenoprotein genes in marine microbial communities is not yet clear. Here, we analyzed the occurrence of all selenoprotein families (including both known and previously predicted) in the metagenomic dataset. Distributions of selenoproteomes in different locations (projects) and samples are shown in Table 1 and Figure 1, respectively.

We identified 5,548 selenoprotein genes belonging to 60 selenoprotein families (Additional file 1: Table S7; the fraction of selenoproteomes in each sample is shown in Additional file 2: Figure S8). The prominent selenoprotein families include selenoprotein W-like (13.3%), AhpD-like (11.5%), peroxiredoxin (Prx, 9.8%), selenophosphate synthetase (8.2%), UGSC-containing (7.2%), and a variety of Prx- and thioredoxin(Trx)-like proteins. Genes encoding the top 10 selenoprotein families

accounted for approximately 70% of all selenoprotein genes.

A total of 16 selenoprotein-rich and 19 selenoprotein-poor marine samples were identified, which were distributed in different oceanic regions (Figure 2). All selenoprotein-rich samples were located in the mesopelagic zone (mostly ≥500 m). In contrast, except for two deep-sea samples from the South China Sea, SCS_1000m (no selenoprotein gene could be detected) and SCS_3000m, all selenoprotein-poor samples were found in the epipelagic zone (~90% ⩽100 m, more than 50% ⩽25 m) of different areas. This may suggest that, similar to the use of Mo, many microorganisms in deep-sea environments may have a stronger ability to use Se. Considering that the oxygen concentration is often lower in the deeper water, our observation is consistent with previous hypothesis that low oxygen level might contribute to the evolution of selenoprotein genes [22].

Interactions among trace element utilization in marine microbes

Integrated analysis of metalloprotein and selenoprotein genes across different samples may help to better understand the interactions among trace elements in marine microbial communities. In this study, we used Spearman correlation coefficient (SCC) to evaluate the relationship between trace elements based on the fractions of metalloprotein and selenoprotein genes in each sample. Except for Cu-Mo, Cu-Se, and Ni-Co, all element pairs were found to be significantly correlated when using all

samples ($p < 0.05$, Figure 3A). Among them, eight element pairs (Cu-Ni, Cu-Co, Mo-W, Mo-Co, Mo-Se, W-Co, W-Se, and Co-Se) were positively correlated whereas four pairs (Cu-W, Mo-Ni, W-Ni, and Ni-Se) were negatively correlated. To further explore element-element interactions in different zones of the ocean, we divided these samples into two categories: epipelagic and mesopelagic subgroups (Figure 3B, the bathypelagic subgroup was excluded as it only contained two samples). In the epipelagic zone (42 samples), a total of 10 element pairs were found to be significantly correlated, eight of which showed the same correlation trend as in the all-sample group in Figure 3A. Significantly positive correlations between Cu and Mo as well as Cu and Se were only observed in this zone, implying the presence of specific crosstalk between the utilization of these elements in marine bacteria living in the near-surface layer. In the mesopelagic zone (22 samples), only four element pairs (Cu-W, Cu-Ni, Mo-Se, and W-Se) were significantly correlated, suggesting a generally weakened interaction between trace elements in deeper layers of the ocean. It should be noted that positive correlations between Mo and Se as well as W and Se are conserved in both subgroups and the all-sample group, suggesting a strong and depth-independent relationship between these elements.

We further compared the patterns of trace element utilization across all marine samples to explore possible linkages between samples with different geographic locations and water depths. The fractions of metalloprotein and selenoprotein genes in

each sample were z-score transformed and then analyzed by hierarchical clustering approach. The correlation heatmap of all samples is shown in Figure 4. Considering that most of the samples were collected from different sites in the Red Sea, it is not surprising that trace element utilization in this area is more conserved than other sea areas. The Red Sea cluster could be further divided into epipelagic and mesopelagic subgroups, indicating similar patterns of trace element utilization in marine microbial communities in the same zones. However, several samples from the epipelagic zone of different RS_Stations (mostly 100 m) were found to be clustered with all RS_Station samples from the mesopelagic zone, suggesting that some other factors may have been important in shaping trace element utilization of extant marine microbes in the Red Sea. Most samples from other locations were also clustered according to different zones they belong to. Interestingly, a small number of samples showed quite different patterns when compared to other samples from the same area, such as RS_AtlantisII_50m (a sample collected at 50 m of Atlantis II Deep, Red Sea, which clustered with samples from the epipelagic zone of other distant areas), SS_20m, and HOT_110m (the latter two were clustered within the group of samples mostly from the epipelagic zone of the Red Sea). This may suggest that geographic location is not so important for maintaining trace element utilization in microbes at these places. The only two bathypelagic samples were found to be grouped in different clusters of the mesopelagic zone, implying a relatively similar pattern of trace element utilization between mesopelagic and bathypelagic zones.

We also analyzed those “-rich” and “-poor” samples which represent highly active or restricted utilization of corresponding elements (60 samples in total). Among them, 38 samples were considered as trace element-dependent protein-rich/-poor samples involving multiple elements, most of which showed consistent trends for the utilization of these elements (*i.e.*, 9 and 11 samples for metalloprotein/selenoprotein-rich and -poor samples, respectively) (Figure 2). All of these metalloprotein/selenoprotein-rich samples were derived from the mesopelagic zone of the Red Sea and Hawaii open-ocean areas, while more than 90% (10 out of 11) metalloprotein/selenoprotein-poor samples were collected from the epipelagic zone of different locations. Several samples from the mesopelagic zone of RS_Station22 and RS_Station34 and of HOT showed the most active trace element utilization among all examined samples, especially for Mo, W, and Se. In contrast, several samples from the epipelagic zone of different locations (excluding the Red Sea) appeared to have a quite limited utilization of the majority of these elements, including BS_10m and BS_80m which had a restricted utilization of all or almost all examined trace elements (Figure 2). Surprisingly, the bathypelagic-zone sample SCS_3000m was also observed to have a poor utilization of Cu, W, Ni, Co, and Se, implying that trace element utilization is generally limited in this place probably due to the decreased availability of most trace elements at such a deep depth.

Correlation analysis between water depth and metallopeotein/selenoprotein families

Although the distributions of metalloproteomes or selenoproteomes are related to water depth, it is unclear which proteins have been influenced. To investigate the vertical characteristics for the evolution of metalloprotein/selenoprotein families in marine microbial world, SCC was then used to identify significant correlations between different metalloprotein/selenoprotein families and depth based on all samples and samples from different locations.

A total of 26 protein families (including 10 metalloprotein and 16 selenoprotein families) appeared to be significantly correlated with depth when using all samples (Table 2). The majority of them showed significant positive correlation, such as XO (the most abundant molybdoprotein), AOR (the only tungstoprotein), and 15 out of the 16 depth-related selenoprotein families, which may partially explain the preference of Mo, W, and Se utilization observed in deeper ocean environments (such as the mesopelagic zone). In contrast, the most abundant cuproproteins (COX I, COX II, and plastocyanin) and Ni-dependent metalloproteins (urease and superoxide dismutase SodN) appeared to be negatively correlated with depth, which is consistent with the above observation that the utilization of the two metals is quite active in the epipelagic zone. We further examined the relationship between metalloprotein/selenoprotein families and water depth using samples from different

locations. Interestingly, depth-related protein families were only found in two areas: RS_Stations and HOT (Table 2). The all-sample-based protein families identified above were also significantly dependent on depth with the same patterns in one or both locations. Moreover, additional depth-related metalloprotein and selenoprotein families could be detected in individual areas, such as RNR II and B_{12}-dependent methyltransferases in RS_Stations and Ni-Fe hydrogenase in HOT. These results imply a complex and somewhat location-specific evolutionary history of trace element-dependent proteins in terms of water depth.

Discussion

The ocean's microbiome plays an important role in the biological and geochemical cycling of trace elements and nutrients in marine systems [40-42]. It has been known that trace element abundances in modern oceans vary both laterally and vertically [43]. Therefore, the bioavailability of trace elements and their utilization may greatly influence marine microbial composition and function. In recent years, a growing number of local and global-scale metagenomic sequencing projects have been performed for exploring marine microbial communities in diverse marine environments and locations, which may provide new insights into the diversity and metabolic activities of microorganisms in the ocean [44-47]. Characterization of genes encoding trace element-dependent proteins in marine metagenomic dataset may help to understand the utilization and functions of these micronutrients in microbial

418 populations and their potential linkages with marine ecological systems. Although
419 previous studies have examined the distribution of certain metalloprotein or
420 selenoprotein genes in surface ocean microbes [25-29], the vertical profiles of trace
421 element utilization in marine microbial communities along depth gradients have not
422 been analyzed yet.
423
424 To investigate the vertical distribution of metalloproteins and selenoproteins in marine
425 microbial world, in this study, we collected all previously published metagenomic
426 datasets containing samples from diverse locations with different depths. Raw
427 shortgun sequencing reads were used to identify metalloprotein (for Cu, Mo, W, Ni,
428 and Co) and selenoprotein genes. For the first time, our results provide a
429 comprehensive analysis of metalloproteomes and selenoproteomes which are used by
430 marine microorganisms at different water depths.
431
432 We built the first map for depicting the vertical profiles of trace element utilization in
433 marine microbes across a large depth gradient in the global ocean. More than 63,000
434 metalloprotein and selenoprotein genes were identified, which is the largest
435 environmental metalloproteome and selenoproteome dataset reported to date. In
436 general, the utilization of Cu, Ni, and Co was highly active in the epipelagic zone
437 (especially <50 m) while Mo, W, and Se were more frequently used in deeper layers
438 of the ocean such as the mesopelagic zone (especially ≥500 m). Exceptions could also

be observed for certain elements, such as highly active Cu utilization in SCS_1000m, highly active Ni utilization in BS_400m and restricted Se utilization in SCS_1000m and SCS_3000m. Analysis of the geographical distribution of metalloprotein and selenoprotein families revealed a quite restricted utilization of all or almost all of examined trace elements in the epipelagic zone of the Baltic Sea as well as the South China Sea (regardless of depth), suggesting that additional factors may inhibit the utilization of trace elements in microbes living in these areas. Consistent with our recent observations, a significant number of RNR II sequences originated from viral genomes (mostly bacteriophages), which may serve as auxiliary genes in the biological and ecological processes of host bacteria to influence the microbial diversity and biogeochemical cycling in the ocean [48, 49]. We also checked other metalloprotein and selenoprotein families and found that the cuproprotein plastocyanin also had a high portion of viral sequences (37.5%, almost all of which were detected at the depths of 10~110 m), which is also consistent with the previous observation [50]. Other protein families examined in this study either lacked or only had very few viral sequences in the metagenomic dataset.

An additional interesting finding of our study is the interactions between different trace elements. Strong positive correlations between Mo and Se as well as W and Se are the most prominent examples which were quite stable across all samples and within different zones of samples. A correlation between Mo and Se utilization has

been previously reported in sequenced prokaryotes most of which were non-marine organisms [51]. Our data suggest that such a relationship may also be present in marine microbial populations. The number of significantly correlated element pairs in the epipelagic zone was larger than that in the mesopelagic zone, implying a more active crosstalk between the utilization of trace elements in marine microbes found in the near-surface layer of the ocean. Clustering analysis of metalloproteomes and selenoproteomes across all samples revealed that several samples obtained from different zones and distant locations have similar patterns of trace element utilization (such as RS_Station169_10m and SCS_1000m) while several other samples obtained from same locations and zones have quite distinct patterns (such as HOT_110m when compared to all other samples from the same zone and location). This suggests that additional unknown factors may play a more important role in determining the evolution of trace element utilization trait than geographic location and depth. Moreover, the findings that samples only possessing enriched metalloprotein/selenoprotein genes were obtained from the mesopelagic zone of Red Sea and Hawaii open ocean imply a highly active utilization of trace elements in these areas.

Analysis of the relationship between depth and metalloprpotein/selenoprotein families could help to identify genes whose evolution is influenced by ocean depth. The occurrence of most of these depth-related protein families (especially several

widespread metalloproteins for Mo and W and selenoproteins) appeared to be positively correlated with depth; however, the occurrence of several abundant Cu- and Ni-dependent metalloproteins showed negative correlation with depth. Further examination of such relationship in different locations revealed that additional metalloprotein and selenoprotein families were significantly correlated with depth in specific areas. Thus, our study not only provides new clues to the evolutionary trends of metalloprotein and selenoprotein genes in marine microbes along a vertical gradient in the ocean, but also helps understand how these microorganisms have adapted to their local environments from the viewpoint of trace element utilization. Future efforts are needed to identify depth-independent environmental factors that could affect the evolution of metalloprotein and/or selenoprotein genes.

Conclusions

In this study, we used metagenomic data derived from different locations around the world ocean to identify the metalloproteomes (Cu, Mo, W, Ni, and Co) and selenoproteomes in a large number of marine microbial samples. We generated the largest environmental metalloprotein and selenoprotein gene dataset which contains more than 63,000 gene sequences, and demonstrated vertical profiles of trace element utilization in marine microbes. In addition, interactions among trace element utilization and the relationship between water depth and metalloprpotein/selenoprotein families were also analyzed at a much wider range of

scales, which provide valuable insights towards our understanding of the complex and dynamic evolution of trace element utilization in marine microbial communities.

Methods

Metagenomic data acquisition and assembly

A total of 66 bacterial metagenomic datasets derived from 6 different locations around the world ocean were downloaded from NCBI Sequence Read Archive with the following accession numbers: PRJNA289734 (eight Red Sea water column stations, 45 samples), PRJNA193416 (Atlantis II Deep, Red Sea, 4 samples), PRJEB1798 (Landsort Deep, Baltic Sea, 3 samples), SRP001096 and SRP001048 (Sargasso Sea BATS Station, 4 samples), PRJNA16339 (Hawaii open-ocean time-series Station ALOHA, 6 samples), and PRJNA77801 (South China Sea SEATS Station, 4 samples).

Quality control (such as trimming adapters and removing low-quality reads and possible contaminations) on raw metagenomic sequencing data was performed by using KneadData (version 0.7.3) with default parameters [52]. After that, the MEGAHIT tool (version 1.2.9) was used for *de novo* assembly of metagenomes [53]. Contigs and singletons that are longer than 150 bp were remained. We then used the BWA-MEM algorithm [54] to map all reads back to the assembly and used samtools

(version 1.10) [55] to extract the mapping information for each read. The assembly and mapping statistics are briefly shown in Additional file 1: Table S8.

Identification of metalloprotein and selenoprotein genes

We collected all known metalloproteins (for Cu, Mo, W, Ni, and Co) and selenoproteins from published resources and databases [16-22, 25]. In addition, we reviewed recent literatures for newly reported metalloproteins and selenoproteins. A list of known metalloprotein and selenoprotein families is shown in Additional file 1: Table S9.

Representative sequences of each metalloprotein family were used as seeds to search against each metagenomic dataset for homologs via TBLASTN with default parameters. Additional homologs were identified using repetitive TBLASTN searches. A possible open reading frame (ORF) was predicted for each nucleotide sequence identified above. All metalloprotein sequences were further verified by examining the presence of conserved domains (such as COG, Pfam, and TIGR) of corresponding metalloprotein families. Known metal-binding residues or motifs (if available) were also examined to help identify metalloproteins.

To predict selenoprotein genes in metagenomic datasets, we adopted the same strategy that was previously used for the identification of selenoprotein genes in the

GOS dataset [25]. In brief, representative sequences of bacterial selenoprotein families were initially used to search for selenoprotein homologs via TBLASTN with default parameters. The Sec/TGA pairs were then selected and the minimum ORF constraint was examined for each TGA-containing nucleotide sequence (TGA was translated to Sec). The presence of a possible Sec insertion sequence (SECIS) element immediately downstream of the Sec-encoding TGA codon was examined for questionable sequences using bSECISearch [56]. Considering that homologs in which the Sec residue is replaced by cysteine (Cys) can be found for almost all selenoprotein families [22], all remaining selenoprotein sequences were further confirmed by searching against the NCBI non-redundant protein database for the presence of Cys-containing homologs via BLASTP.

The fraction of genes encoding metalloproteins or selenoproteins in each sample was normalized using the number of reads in the corresponding ORF divided by the total number of cleaned and mapped reads obtained for the sample. Metalloprotein/selenoprotein-rich (containing at least 1.5 times the average level of metalloprotein or selenoprotein genes across all samples) and metalloprotein/selenoprotein-poor samples (containing no more than half of the average level of metalloprotein or selenoprotein genes) were designated by using the same criteria as previously described [25].

Correlation and clustering analyses

SCC was used to evaluate the relationship between trace elements based on the fraction of metalloprotein and selenoprotein genes in different samples. Significantly correlated element pairs ($p < 0.05$) were presented using the R package corrplot (version 0.84). In addition, SCC was also used to assess the association between metalloprotein/selenoprotein families and water depth with a predefined threshold of significance ($p < 0.05$, SCC > 0.5). Hierarchical clustering and heatmap analysis were performed via the R packages ggdendro, ggplot2, and pheatmap.

List of abbreviations

Fe: iron

Zn: zinc

Cu: copper

Mn: manganese

Ni: nickel

Co: cobalt

Mo: molybdenum

W: tungsten

Cr: chromium

Se: selenium

Sec: selenocysteine

585 GOS: Global Ocean Sampling

586 COX I: Cytochrome *c* oxidase subunit I

587 COX II: Cytochrome *c* oxidase subunit II

588 SCS: South China Sea

589 RS_Station: Red Sea water column station

590 BS: Baltic Sea

591 DMSOR: dimethylsulfoxide reductase

592 SO: sulfite oxidase

593 XO: xanthine oxidase

594 AOR: aldehyde:ferredoxin oxidoreductase

595 HOT: Hawaii open ocean area time-series

596 CODH: carbon monoxide dehydrogenase

597 CODH/ACS: bifunctional CODH/acetyl-coenzyme A synthase

598 MCM: methylmalonyl-CoA mutase

599 RNR II: B_{12}-dependent class II ribonucleotide reductase

600 MetH: methionine synthase

601 SS: Sargasso Sea

602 Prx: peroxiredoxin

603 Trx: thioredoxin

604 SCC: Spearman correlation coefficient

605 ORF: open reading frame (ORF)

606 SECIS: Sec insertion sequence

607 Cys: cysteine

608

609 # Declarations

610 **Ethics approval and consent to participate**

611 Not applicable.

612

613 **Consent for publication**

614 Not applicable.

615

616 **Availability of data and materials**

617 Sequences of metalloproteins and selenoproteins analyzed during this study are

618 available as supplementary data.

619

620 **Competing interests**

621 The authors declare that they have no competing interests.

622

623 **Funding**

624 This work was supported by National Natural Science Foundation of China (grant

625 number 31771407), Guangdong Basic and Applied Basic Research Foundation (grant

626 number 2019A1515011938), Science and Technology Innovation Committee of

Shenzhen Municipality (grant number JCYJ20180305124023495), and Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (grant number 2019SHIBS0003).

Authors' contributions

YX and YZ designed the study. YX performed all computational analyses (such as database search, identification of metalloprotein/selenoprotein genes and correlation analysis) and wrote the manuscript. QL and YZ edited the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Goldhaber SB. Trace element risk assessment: essentiality vs. toxicity. Regul Toxicol Pharmacol. 2003;38(2):232-42.
2. Nordberg M, Nordberg GF. Trace element research-historical and future aspects. J Trace Elem Med Biol. 2016;38:46-52.
3. Mehri A. Trace Elements in Human Nutrition (II) - An Update. Int J Prev Med. 2020;11:2.
4. King JC. Zinc: an essential but elusive nutrient. Am J Clin Nutr.

2011;94(2):679S-84S.

5. Oliveira F, Rocha S, Fernandes R. Iron metabolism: from health to disease. J Clin Lab Anal. 2014;28(3):210-8.

6. Yannone SM, Hartung S, Menon AL, Adams MWW, Tainer JA. Metals in biology: defining metalloproteomes. Curr Opin Biotech. 2012;23(1):89-95.

7. Zhang Y, Ying H, Xu Y. Comparative genomics and metagenomics of the metallomes. Metallomics. 2019;11(6):1026-43.

8. Magalon A, Mendel RR. Biosynthesis and Insertion of the Molybdenum Cofactor. EcoSal Plus. 2015;6(2).

9. Banerjee R, Ragsdale SW. The many faces of vitamin B12: catalysis by cobalamin-dependent enzymes. Annu Rev Biochem. 2003;72:209-47.

10. Wrobel JK, Power R, Toborek M. Biological activity of selenium: Revisited. IUBMB Life. 2016;68(2):97-105.

11. Mangiapane E, Pessione A, Pessione E. Selenium and selenoproteins: an overview on different biological systems. Curr Protein Pept Sci. 2014;15(6):598-607.

12. Steinbrenner H, Speckmann B, Klotz LO. Selenoproteins: Antioxidant selenoenzymes and beyond. Arch Biochem Biophys. 2016;595:113-9.

13. Zhang Y, Zheng J. Bioinformatics of Metalloproteins and Metalloproteomes. Molecules. 2020;25(15):3366.

14. Maret W. Zinc and the zinc proteome. Met Ions Life Sci. 2013;12:479-501.

669 15. Sydor AM, Zamble DB. Nickel metallomics: general themes guiding nickel
670 homeostasis. Met Ions Life Sci. 2013;12:375-416.
671 16. Peng T, Xu Y, Zhang Y. Comparative genomics of molybdenum utilization in
672 prokaryotes and eukaryotes. BMC Genomics. 2018;19(1):691.
673 17. Santesmasses D, Mariotti M, Gladyshev VN. Bioinformatics of Selenoproteins.
674 Antioxid Redox Signal. 2020;33(7):525-36.
675 18. Santesmasses D, Mariotti M, Guigó R. Selenoprofiles: A Computational Pipeline
676 for Annotation of Selenoproteins. Methods Mol Biol. 2018;1661:17-28.
677 19. Ridge PG, Zhang Y, Gladyshev VN. Comparative genomic analyses of copper
678 transporters and cuproproteomes reveal evolutionary dynamics of copper
679 utilization and its link to oxygen. PLoS One. 2008;3(1):e1378.
680 20. Zhang Y, Gladyshev VN. General trends in trace element utilization revealed by
681 comparative genomic analyses of Co, Cu, Mo, Ni, and Se. J Biol Chem.
682 2010;285(5):3393-405.
683 21. Zhang Y, Rodionov DA, Gelfand MS, Gladyshev VN. Comparative genomic
684 analyses of nickel, cobalt and vitamin B12 utilization. BMC Genomics.
685 2009;10:78.
686 22. Peng T, Lin J, Xu YZ, Zhang Y. Comparative genomics reveals new evolutionary
687 and ecological patterns of selenium utilization in bacteria. ISME J.
688 2016;10(8):2048-59.
689 23. Lin J, Peng T, Jiang L, Ni JZ, Liu Q, Chen L, et al. Comparative genomics

reveals new candidate genes involved in selenium metabolism in prokaryotes. Genome Biol Evol. 2015;7(3):664-76.

24. Miller WG, Yee E, Lopes BS, Chapman MH, Huynh S, Bono JL, et al. Comparative Genomic Analysis Identifies a Campylobacter Clade Deficient in Selenium Metabolism. Genome Biol Evol. 2017;9(7):1843-58.

25. Zhang Y, Gladyshev VN. Trends in selenium utilization in marine microbial world revealed through the analysis of the global ocean sampling (GOS) project. PLoS Genetics. 2008;4(6):e1000095.

26. Toulza E, Tagliabue A, Blain S, Piganeau G. Analysis of the global ocean sampling (GOS) project for trends in iron uptake by surface ocean microbes. PLoS One. 2012;7(2):e30931.

27. Desai DK, Desai FD, Laroche J. Factors influencing the diversity of iron uptake systems in aquatic microorganisms. Front Microbiol. 2012;3:362.

28. Farukh M. Comparative genomic analysis of selenium utilization traits in different marine environments. J Microbiol. 2020;58(2):113-22.

29. Doxey AC, Kurtz DA, Lynch MD, Sauder LA, Neufeld JD. Aquatic metagenomes implicate Thaumarchaeota in global cobalamin production. ISME J. 2015;9(2):461-71.

30. Antoine R, Rivera-Millot A, Roy G, Jacob-Dubuisson F. Relationships Between Copper-Related Proteomes and Lifestyles in β Proteobacteria. Front Microbiol. 2019;10:2217.

31. Hille R, Hall J, Basu PJCr. The mononuclear molybdenum enzymes. Chem Rev. 2014;114(7):3963-4038.
32. Moura JJ, Bernhardt PV, Maia LB, Gonzalez PJ. Molybdenum and tungsten enzymes: from biology to chemistry and back. J Biol Inorg Chem. 2015;20(2):181-2.
33. Miralles-Robledillo JM, Torregrosa-Crespo J, Martínez-Espinosa RM, Pire C. DMSO Reductase Family: Phylogenetics and Applications of Extremophiles. Int J Mol Sci. 2019;20(13):3349.
34. Alfano M, Cavazza C. Structure, function, and biosynthesis of nickel-dependent enzymes. Protein Sci. 2020;29(5):1071-89.
35. Evans PN, Boyd JA, Leu AO, Woodcroft BJ, Parks DH, Hugenholtz P, et al. An evolving view of methane metabolism in the Archaea. Nat Rev Microbiol. 2019;17(4):219-32.
36. Marsh EN. Coenzyme B12 (cobalamin)-dependent enzymes. Essays Biochem. 1999;34:139-54.
37. Giedyk M, Goliszewska K, Gryko D. Vitamin B12 catalysed reactions. Chem Soc Rev. 2015;44(11):3391-404.
38. Bridwell-Rabb J, Drennan CL. Vitamin B(12) in the spotlight again. Curr Opin Chem Biol. 2017;37:63-70.
39. Tahara K, Pan L, Ono T, Hisaeda YJBjooc. Learning from B12 enzymes: biomimetic and bioinspired catalysts for eco-friendly organic synthesis. Beilstein

J Org Chem. 2018;14(1):2553-67.

40. Leão PN, Vasconcelos MT, Vasconcelos VM. Role of marine cyanobacteria in trace metal bioavailability in seawater. Microb Ecol. 2007;53(1):104-9.

41. Morel FM. The co-evolution of phytoplankton and trace element cycles in the oceans. Geobiology. 2008;6(3):318-24.

42. Anderson RF. GEOTRACES: Accelerating Research on the Marine Biogeochemical Cycles of Trace Elements and Their Isotopes. Ann Rev Mar Sci. 2020;12:49-85.

43. Bruland KW, Lohan MC. Controls of trace metals in seawater. In: Holland HD, Turekian KK, editors. Treatise on Geochemistry. Amsterdam: Elsevier B.V.; 2003, p. 23–47.

44. Kennedy J, Flemer B, Jackson SA, Lejon DP, Morrissey JP, O'gara F, et al. Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. Mar Drugs. 2010;8(3):608-28.

45. Wang DZ, Xie ZX, Zhang SF. Marine metaproteomics: current status and future directions. J Proteomics. 2014;97:27-35.

46. Alma'abadi AD, Gojobori T, Mineta K. Marine Metagenome as A Resource for Novel Enzymes. Genomics Proteomics Bioinformatics. 2015;13(5):290-5.

47. Behzad H, Ibarra MA, Mineta K, Gojobori T. Metagenomic studies of the Red Sea. Gene. 2016;576(2 Pt 1):717-23.

48. Enav H, Mandel-Gutfreund Y, Béjà O. Comparative metagenomic analyses reveal

753 viral-induced shifts of host metabolism towards nucleotide biosynthesis.
754 Microbiome. 2014;2(1):9.
755 49. Dwivedi B, Xue B, Lundin D, Edwards RA, Breitbart M. A bioinformatic
756 analysis of ribonucleotide reductase genes in phage genomes and metagenomes.
757 BMC Evol Biol. 2013;13:33.
758 50. Puxty RJ, Millard AD, Evans DJ, Scanlan DJ. Shedding new light on viral
759 photosynthesis. Photosynth Res. 2015;126(1):71-97.
760 51. Zhang Y, Gladyshev VN. Comparative genomics of trace elements: emerging
761 dynamic view of trace element utilization and function. Chem Rev.
762 2009;109(10):4828-61.
763 52. The KneadData tool. http://huttenhower.sph.harvard.edu/kneaddata. Accessed 20
764 July 2020.
765 53. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast
766 single-node solution for large and complex metagenomics assembly via succinct
767 de Bruijn graph. Bioinformatics. 2015;31(10):1674-6.
768 54. Li H. Toward better understanding of artifacts in variant calling from
769 high-coverage samples. Bioinformatics. 2014;30(20):2843-51.
770 55. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler
771 transform. Bioinformatics. 2010;26(5):589-95.
772 56. Zhang Y, Gladyshev VN. An algorithm for identification of bacterial
773 selenocysteine insertion sequence elements and selenoprotein genes.

Bioinformatics. 2005;21(11):2580-9.

Figure legends

Figure 1. General distribution of metalloproteomes and selenoproteomes in marine metegenomic samples. Sequential color schemes represent different geographic locations and stations. The six tracks (circles) within the sample name circle (from outside to inside) represent the normalized occurrence of metalloproteomes (Cu, Mo, W, Ni, and Co) and selenoproteomes, respectively. The length of each column represents the normalized ratio of the fraction of metalloproteome/selenoproteome in each sample to the average of corresponding proteomes. Metalloprotein/selenoprotein-rich and -poor samples are highlighted in red and blue, respectively.

Figure 2. Vertical distribution of metalloprotein/selenoprotein-rich and -poor samples in different locations. Each sample is represented by a rectangle which is divided into six parts indicating different trace elements. Metalloprotein/selenoprotein-rich and -poor samples are highlighted in red and blue, respectively. Other samples are shown in grey.

Figure 3. Correlation analysis of trace element utilization. (A) All samples. (B) Samples in the epipelagic and mesopelagic zones. Only significant correlations ($p < 0.05$) are shown. Positive and negative correlations are represented in white and black, respectively. The size of the circle is proportional to the SCC values.

Figure 4. The correlation heatmap of marine samples. Red and green colors represent increased and decreased levels of metalloproteomes/selenoproteomes (transformed into z-scores), respectively, when compared to the average level for each element. Columns and lines represent trace elements and samples, respectively. Different zones that samples belong to are also indicated.

Tables

Table 1. Distribution of metalloprotein and selenoprotein genes in different marine metagenomic datasets

Accession number (location)	Num. of samples	Num. of metalloprotein genes					Num. of selenoprotein genes
		Cu	Mo	W	Ni	Co	
PRJNA289734 (Red Sea water column stations)	45	4,859	19,977	508	3265	12,513	3,774
PRJNA193416 (Atlantis II Deep, Red Sea)	4	718	3,420	112	700	3,376	705
PRJEB1798 (Landsort Deep, Baltic Sea)	3	57	385	73	89	163	56
SRP001096 & SRP001048 (Sargasso Sea)	4	91	223	7	116	95	21
PRJNA16339 (Hawaii open ocean)	6	575	4,375	257	436	1,415	987
PRJNA77801 (South China Sea)	4	36	230	1	34	43	5
Total	66	6,336	28,610	958	4,640	17,605	5,548

854 **Table 2. Metalloprotein and selenoprotein families associated with water depth**

Trace element	Protein family	Spearman correlation coefficient		
		All samples (n = 66)	Red Sea water column Stations (n = 45)	Hawaii open ocean (n = 6)
Cu	Particulate methane monooxygenase	0.66	0.85	-
	Nitrite reductase	0.64	0.82	0.94
	Cytochrome c oxidase subunit I	-0.6	-0.71	-
	Plastocyanin	-0.6	-0.61	-
	Cytochrome c oxidase subunit II	-0.66	-0.87	-
Mo	Xanthine oxidase	0.69	0.79	1.00
W	Aldehyde:ferredoxin oxidoreductase	0.64	0.66	0.99
Ni	Urease	-0.54	-0.56	-
	Superoxide dismutase SodN	-0.62	-0.79	-0.89
	Lactate racemase	-	0.63	0.94
	Ni-Fe hydrogenase	-	-	0.85
Co	PpaA	-0.63	-0.75	-
	B_{12}-dependent methyltransferases	-	0.56	-
	B_{12}-dependent class II ribonucleotide reductase	-	0.53	-
	Methionine synthase	-	-	0.84
Se	Peroxiredoxin (Prx)	0.70	0.83	0.94
	Formate dehydrogenase alpha subunit	0.69	0.86	0.88
	Prx-like, UGC-containing	0.65	0.83	-
	COG0737 UshA	0.63	0.79	0.94
	Thiol:disulfide isomerase-like protein	0.61	0.81	-
	Hypothetical protein 1	0.59	0.8	0.85
	Glutathione S-transferase	0.57	0.82	-
	Prx-like thiol:disulfide oxidoreductase	0.56	0.71	0.88
	Arsenate reductase 1 UxxS-containing	0.56	0.79	0.94
	Rhodanase related sulfurtransferase	0.56	0.5	0.88
	OS_HP1	0.56	0.66	0.85
	Hypothetical protein GOS_C	0.53	0.65	0.94
	ULPU-containing	0.53	0.62	0.85
	Proline reductase	0.52	0.8	0.99
	Deiodinase-like	0.51	0.57	0.85
	Thioredoxin(Trx)-like	-0.64	-0.72	-
	UGSC-containing	-	0.72	0.88
	Arsenate reductase 2 UxxC-containing	-	0.7	0.88
	Molybdopterin biosynthesis protein MoeB	-	0.67	-
	Distant homolog of new Trx-like protein AACY01531301	-	0.61	-
	DsbA-like	-	0.6	0.94
	Hypothetical protein GOS_B	-	0.55	0.94
	Glycine reductase selenoprotein B	-	0.52	-
	Thioredoxin	-	-0.59	-
	Thiol:disulfide interchange protein	-	-	0.94
	OsmC-like protein	-	-	0.85
	Prx-like	-	-	0.85
	Unknown protein (UAAU)	-	-	0.85
	Glutathione peroxidase	-	-	0.85

855

856

857

858

Additional files

Additional file 1: Supplementary_tables.xlsx

This file contains nine supplementary tables: Table S1 shows all marine metagenomic samples and related information. Table S2 shows the distribution of cuproprotein genes in metagenomic samples. Table S3 shows the distribution of molybdoprotein genes in metagenomic samples. Table S4 shows the distribution of tungstoprotein genes in metagenomic samples. Table S5 shows the distribution of Ni-dependent protein genes in metagenomic samples. Table S6 shows the distribution of Co-dependent protein genes in metagenomic samples. Table S7 shows the distribution of selenoprotein genes in metagenomic samples. Table S8 shows the assembly statistics for metagenomic samples. Table S9 shows the list of known metalloproteins (Cu, Mo, W, Ni, and Co) and selenoproteins in prokaryotes.

Additional file 2: **Supplementary_figures.pdf**

This file contains eight supplementary figures: Figure S1 shows geographic locations of marine metagenomic samples. Figure S2 shows the distribution of cuproproteomes in marine metagenomic samples. Figure S3 shows the distribution of molybdoproteomes in marine metagenomic samples. Figure S4 shows the distribution of tungstoproteomes in marine metagenomic samples. Figure S5 shows the distribution of Ni-dependent proteomes in marine metagenomic samples. Figure S6 shows the distribution of Co-dependent proteomes in marine metagenomic samples.

Figure S7 shows putative taxonomic affiliation of RNR II sequences detected in the metagenomic dataset. Figure S8 shows the distribution of selenoproteomes in marine metagenomic samples.

Additional file 3: sequence_dataset.txt

This file contains all metalloprotein and selenoprotein sequences analyzed in this study.