

Vertical Profiles of Microbial Trace Element Utilization in Marine Environments

Yinzhen Xu

College of Life Sciences and Oceanography, Shenzhen University

Qiong Liu

College of Life Sciences and Oceanography, Shenzhen University

Yan Zhang (✉ zhangyan@szu.edu.cn)

College of Life Sciences and Oceanography, Shenzhen University

Research Article

Keywords: Trace element, Marine metagenome, Metalloprotein, Selenoprotein, Depth, Evolution

Posted Date: January 12th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-141603/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **Vertical profiles of microbial trace element utilization**
2 **in marine environments**

3 Yinzhen Xu^{1,2,3,4}, Qiong Liu^{1,2,4}, Yan Zhang^{1,2,4,*}

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

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

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

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

14
15 * Corresponding author: Yan Zhang, College of Life Sciences and Oceanography,
16 Shenzhen University, Shenzhen, 518055, Guangdong Province, P. R. China. Tel:
17 86-755-26922024; Fax: 86-755-86713951; E-mail: zhangyan@szu.edu.cn

18
19
20
21

22 **Abstract**

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

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

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

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

49

50 **Keywords**

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

52

53

54

55

56

57

58

59

60

61

62

63

64 **Background**

65 Biological trace elements are required by all organisms in very small quantities and
66 their excess or insufficiency may interfere with various metabolic and physiologic
67 processes of organisms [1-3]. These micronutrients (mostly metals) include iron (Fe),
68 zinc (Zn), copper (Cu), manganese (Mn), nickel (Ni), cobalt (Co), molybdenum (Mo),
69 tungsten (W), chromium (Cr), selenium (Se), and several other elements. Utilization
70 of trace elements is generally rather complex. Fe and Zn are essential to nearly all
71 organisms [4, 5]; however, the utilization of many other elements is quite diverse and
72 scattered throughout all domains of life [6, 7]. Metal ions are usually required as
73 cofactors for the assembly of metalloproteins. The majority of them are directly
74 incorporated into their cognate binding sites on proteins. A small number of metals
75 have to become part of prosthetic groups or cofactors prior to their insertion into
76 target proteins, such as Mo and Co that are mainly used in the forms of molybdopterin
77 (Mo cofactor) and vitamin B₁₂ (cobalamin), respectively [8, 9]. Se, one of the most
78 important metalloid elements, is involved in a variety of redox reactions in many
79 organisms [10]. It mainly exists as selenocysteine (Sec) which is a nonstandard amino
80 acid encoded by TGA (normally a stop codon) in all selenoproteins [11, 12].

81

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

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

91

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

103

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

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

122

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

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

133

134 **Results**

135 Based on the metagenomic data derived from 66 marine samples from different
136 locations around the world (details of the samples are shown in Additional file 1:
137 Table S1; geographic locations of these samples are shown in Additional file 2: Figure
138 S1), we generated the first map illustrating vertical patterns of trace element
139 utilization in marine microbial communities. These samples spanned a large depth
140 gradient of 10 to 3000 m, which belong to epipelagic (<200 m), mesopelagic (\geq 200 m
141 and \leq 1000 m), and bathypelagic (>1000 m and <4000 m) zones of the world oceans.

142 A total of 63,697 metalloprotein and selenoprotein genes were predicted, which is so
143 far the largest dataset of genes encoding trace element-dependent proteins (all
144 predicted metalloprotein and selenoprotein sequences are stored in Additional file 3).

145

146 **General analysis of metalloprotein genes and metalloproteomes**

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

153

154 *Copper*

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

166

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

181

182 *Molybdenum and tungsten*

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

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

193

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

203

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

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

212

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

222

223 *Nickel*

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

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

237

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

244

245 *Cobalt*

246 Co is an essential metal constituent of cobalamin (or called vitamin B₁₂) which is a
247 complex organometallic cofactor needed for a variety of B₁₂-dependent enzymes,
248 such as methylmalonyl-CoA mutase (MCM), B₁₂-dependent class II ribonucleotide
249 reductase (RNR II), methionine synthase (MetH), dehalogenase, and some other
250 important enzymes [36-39]. In this study, we identified 17,605 Co-dependent protein

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

262

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

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

273

274 **Analysis of selenoprotein genes and selenoproteomes**

275 Se is a necessary trace element for many vital physiological functions in a variety of
276 organisms, which exerts its biological role through selenoproteins [10-12]. It has been
277 suggested that the number of selenoprotein families detected in bacteria is much
278 larger than that in eukaryotes or archaea [17, 20, 22]. Although aquatic habitat has
279 been thought to promote the evolution of new selenoprotein genes in microorganisms
280 [22, 25], the relationship between depth and the distribution of selenoprotein genes in
281 marine microbial communities is not yet clear. Here, we analyzed the occurrence of
282 all selenoprotein families (including both known and previously predicted) in the
283 metagenomic dataset. Distributions of selenoproteomes in different locations (projects)
284 and samples are shown in Table 1 and Figure 1, respectively.

285

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

292 accounted for approximately 70% of all selenoprotein genes.

293

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

305

306 **Interactions among trace element utilization in marine microbes**

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

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

330

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

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

355

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

375

376 **Correlation analysis between water depth and metalloprotein/selenoprotein**
377 **families**

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

384

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

397 locations. Interestingly, depth-related protein families were only found in two areas:
398 RS_Stations and HOT (Table 2). The all-sample-based protein families identified
399 above were also significantly dependent on depth with the same patterns in one or
400 both locations. Moreover, additional depth-related metalloprotein and selenoprotein
401 families could be detected in individual areas, such as RNR II and B₁₂-dependent
402 methyltransferases in RS_Stations and Ni-Fe hydrogenase in HOT. These results
403 imply a complex and somewhat location-specific evolutionary history of trace
404 element-dependent proteins in terms of water depth.

405

406 **Discussion**

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

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

455

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

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

477

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

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

492

493 **Conclusions**

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

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

504

505 **Methods**

506 **Metagenomic data acquisition and assembly**

507 A total of 66 bacterial metagenomic datasets derived from 6 different locations around
508 the world ocean were downloaded from NCBI Sequence Read Archive with the
509 following accession numbers: PRJNA289734 (eight Red Sea water column stations,
510 45 samples), PRJNA193416 (Atlantis II Deep, Red Sea, 4 samples), PRJEB1798
511 (Landsort Deep, Baltic Sea, 3 samples), SRP001096 and SRP001048 (Sargasso Sea
512 BATS Station, 4 samples), PRJNA16339 (Hawaii open-ocean time-series Station
513 ALOHA, 6 samples), and PRJNA77801 (South China Sea SEATS Station, 4
514 samples).

515

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

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

524

525 **Identification of metalloprotein and selenoprotein genes**

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

531

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

540

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

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

554

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

563

564 **Correlation and clustering analyses**

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

572

573 **List of abbreviations**

574 Fe: iron

575 Zn: zinc

576 Cu: copper

577 Mn: manganese

578 Ni: nickel

579 Co: cobalt

580 Mo: molybdenum

581 W: tungsten

582 Cr: chromium

583 Se: selenium

584 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: B₁₂-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

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

630

631 **Authors' contributions**

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

636

637 **Acknowledgements**

638 Not applicable.

639

640 **References**

- 641 1. Goldhaber SB. Trace element risk assessment: essentiality vs. toxicity. Regul
642 Toxicol Pharmacol. 2003;38(2):232-42.
- 643 2. Nordberg M, Nordberg GF. Trace element research-historical and future aspects.
644 J Trace Elem Med Biol. 2016;38:46-52.
- 645 3. Mehri A. Trace Elements in Human Nutrition (II) - An Update. Int J Prev Med.
646 2020;11:2.
- 647 4. King JC. Zinc: an essential but elusive nutrient. Am J Clin Nutr.

-
- 648 2011;94(2):679S-84S.
- 649 5. Oliveira F, Rocha S, Fernandes R. Iron metabolism: from health to disease. *J Clin*
650 *Lab Anal.* 2014;28(3):210-8.
- 651 6. Yannone SM, Hartung S, Menon AL, Adams MWW, Tainer JA. Metals in biology:
652 defining metalloproteomes. *Curr Opin Biotech.* 2012;23(1):89-95.
- 653 7. Zhang Y, Ying H, Xu Y. Comparative genomics and metagenomics of the
654 metallomes. *Metallomics.* 2019;11(6):1026-43.
- 655 8. Magalon A, Mendel RR. Biosynthesis and Insertion of the Molybdenum Cofactor.
656 *EcoSal Plus.* 2015;6(2).
- 657 9. Banerjee R, Ragsdale SW. The many faces of vitamin B12: catalysis by
658 cobalamin-dependent enzymes. *Annu Rev Biochem.* 2003;72:209-47.
- 659 10. Wrobel JK, Power R, Toborek M. Biological activity of selenium: Revisited.
660 *IUBMB Life.* 2016;68(2):97-105.
- 661 11. Mangiapane E, Pessione A, Pessione E. Selenium and selenoproteins: an
662 overview on different biological systems. *Curr Protein Pept Sci.*
663 2014;15(6):598-607.
- 664 12. Steinbrenner H, Speckmann B, Klotz LO. Selenoproteins: Antioxidant
665 selenoenzymes and beyond. *Arch Biochem Biophys.* 2016;595:113-9.
- 666 13. Zhang Y, Zheng J. Bioinformatics of Metalloproteins and Metalloproteomes.
667 *Molecules.* 2020;25(15):3366.
- 668 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

-
- 690 reveals new candidate genes involved in selenium metabolism in prokaryotes.
691 Genome Biol Evol. 2015;7(3):664-76.
- 692 24. Miller WG, Yee E, Lopes BS, Chapman MH, Huynh S, Bono JL, et al.
693 Comparative Genomic Analysis Identifies a Campylobacter Clade Deficient in
694 Selenium Metabolism. Genome Biol Evol. 2017;9(7):1843-58.
- 695 25. Zhang Y, Gladyshev VN. Trends in selenium utilization in marine microbial
696 world revealed through the analysis of the global ocean sampling (GOS) project.
697 PLoS Genetics. 2008;4(6):e1000095.
- 698 26. Toulza E, Tagliabue A, Blain S, Piganeau G. Analysis of the global ocean
699 sampling (GOS) project for trends in iron uptake by surface ocean microbes.
700 PLoS One. 2012;7(2):e30931.
- 701 27. Desai DK, Desai FD, Laroche J. Factors influencing the diversity of iron uptake
702 systems in aquatic microorganisms. Front Microbiol. 2012;3:362.
- 703 28. Farukh M. Comparative genomic analysis of selenium utilization traits in
704 different marine environments. J Microbiol. 2020;58(2):113-22.
- 705 29. Doxey AC, Kurtz DA, Lynch MD, Sauder LA, Neufeld JD. Aquatic
706 metagenomes implicate Thaumarchaeota in global cobalamin production. ISME J.
707 2015;9(2):461-71.
- 708 30. Antoine R, Rivera-Millot A, Roy G, Jacob-Dubuisson F. Relationships Between
709 Copper-Related Proteomes and Lifestyles in β Proteobacteria. Front Microbiol.
710 2019;10:2217.

-
- 711 31. Hille R, Hall J, Basu P. Cr. The mononuclear molybdenum enzymes. *Chem Rev.*
712 2014;114(7):3963-4038.
- 713 32. Moura JJ, Bernhardt PV, Maia LB, Gonzalez PJ. Molybdenum and tungsten
714 enzymes: from biology to chemistry and back. *J Biol Inorg Chem.*
715 2015;20(2):181-2.
- 716 33. Miralles-Robledillo JM, Torregrosa-Crespo J, Martínez-Espinosa RM, Pire C.
717 DMSO Reductase Family: Phylogenetics and Applications of Extremophiles. *Int*
718 *J Mol Sci.* 2019;20(13):3349.
- 719 34. Alfano M, Cavazza C. Structure, function, and biosynthesis of nickel-dependent
720 enzymes. *Protein Sci.* 2020;29(5):1071-89.
- 721 35. Evans PN, Boyd JA, Leu AO, Woodcroft BJ, Parks DH, Hugenholtz P, et al. An
722 evolving view of methane metabolism in the Archaea. *Nat Rev Microbiol.*
723 2019;17(4):219-32.
- 724 36. Marsh EN. Coenzyme B12 (cobalamin)-dependent enzymes. *Essays Biochem.*
725 1999;34:139-54.
- 726 37. Giedyk M, Goliszevska K, Gryko D. Vitamin B12 catalysed reactions. *Chem Soc*
727 *Rev.* 2015;44(11):3391-404.
- 728 38. Bridwell-Rabb J, Drennan CL. Vitamin B(12) in the spotlight again. *Curr Opin*
729 *Chem Biol.* 2017;37:63-70.
- 730 39. Tahara K, Pan L, Ono T, Hisaeda Y. Bjooc. Learning from B12 enzymes:
731 biomimetic and bioinspired catalysts for eco-friendly organic synthesis. *Beilstein*

-
- 732 J Org Chem. 2018;14(1):2553-67.
- 733 40. Leão PN, Vasconcelos MT, Vasconcelos VM. Role of marine cyanobacteria in
734 trace metal bioavailability in seawater. *Microb Ecol.* 2007;53(1):104-9.
- 735 41. Morel FM. The co-evolution of phytoplankton and trace element cycles in the
736 oceans. *Geobiology.* 2008;6(3):318-24.
- 737 42. Anderson RF. GEOTRACES: Accelerating Research on the Marine
738 Biogeochemical Cycles of Trace Elements and Their Isotopes. *Ann Rev Mar Sci.*
739 2020;12:49-85.
- 740 43. Bruland KW, Lohan MC. Controls of trace metals in seawater. In: Holland HD,
741 Turekian KK, editors. *Treatise on Geochemistry.* Amsterdam: Elsevier B.V.; 2003,
742 p. 23–47.
- 743 44. Kennedy J, Flemer B, Jackson SA, Lejon DP, Morrissey JP, O’gara F, et al.
744 Marine metagenomics: new tools for the study and exploitation of marine
745 microbial metabolism. *Mar Drugs.* 2010;8(3):608-28.
- 746 45. Wang DZ, Xie ZX, Zhang SF. Marine metaproteomics: current status and future
747 directions. *J Proteomics.* 2014;97:27-35.
- 748 46. Alma'abadi AD, Gojobori T, Mineta K. Marine Metagenome as A Resource for
749 Novel Enzymes. *Genomics Proteomics Bioinformatics.* 2015;13(5):290-5.
- 750 47. Behzad H, Ibarra MA, Mineta K, Gojobori T. Metagenomic studies of the Red
751 Sea. *Gene.* 2016;576(2 Pt 1):717-23.
- 752 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.

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

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793

794

795 **Figure legends**

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

805

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

811

812 **Figure 3. Correlation analysis of trace element utilization.** (A) All samples. (B)
813 Samples in the epipelagic and mesopelagic zones. Only significant correlations ($p <$
814 0.05) are shown. Positive and negative correlations are represented in white and black,
815 respectively. The size of the circle is proportional to the SCC values.

816

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

822

823

824

825

826

827

828

829

830

831

832

833

834

835

836

837 **Tables**838 **Table 1. Distribution of metalloprotein and selenoprotein genes in different**
839 **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

840

841

842

843

844

845

846

847

848

849

850

851

852

853

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

859 **Additional files**

860 **Additional file 1: Supplementary_tables.xlsx**

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

871

872 **Additional file 2: Supplementary_figures.pdf**

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

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

883

884 **Additional file 3: sequence_dataset.txt**

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

Figures

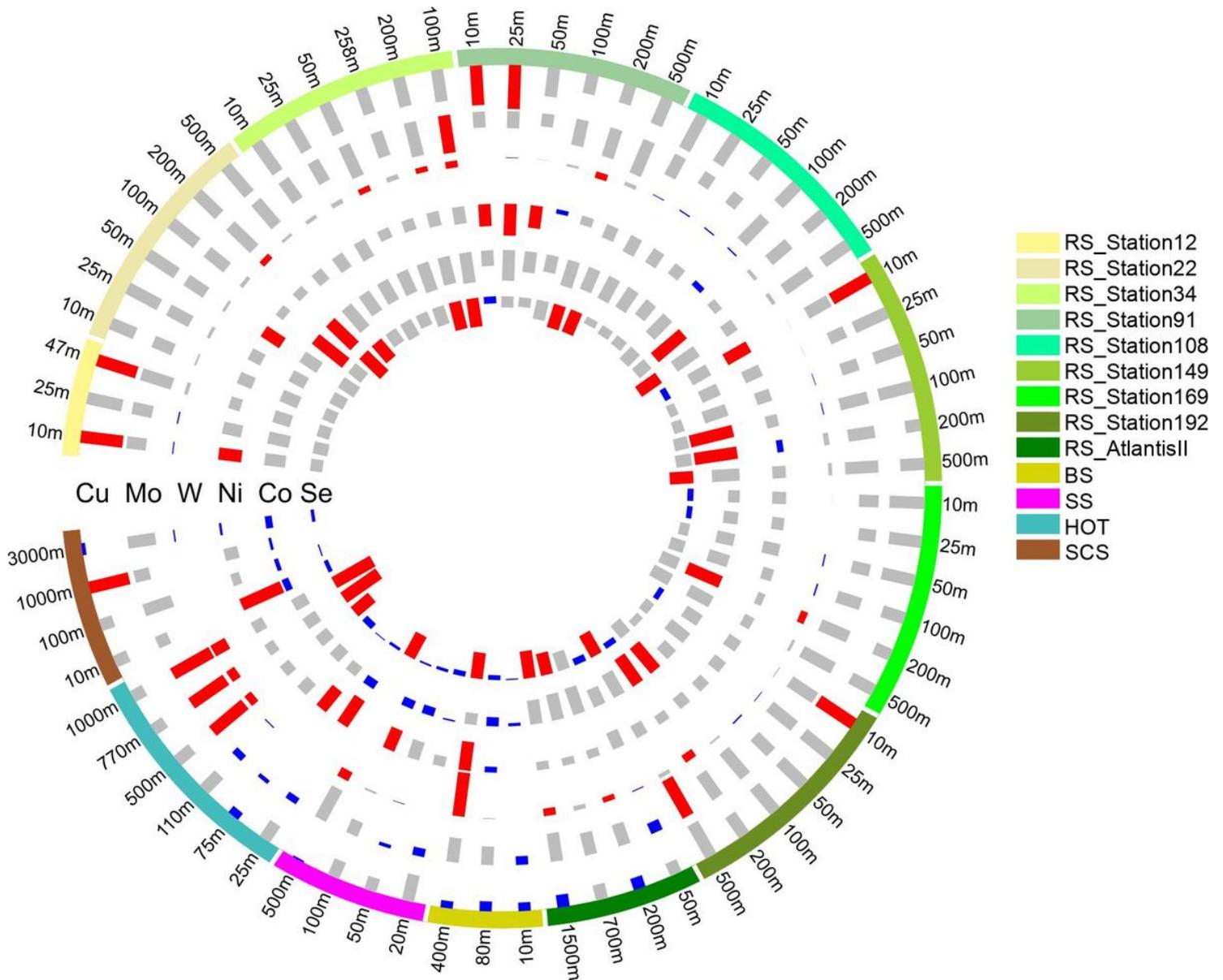


Figure 1

General distribution of metalloproteomes and selenoproteomes in marine metagenomic 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.

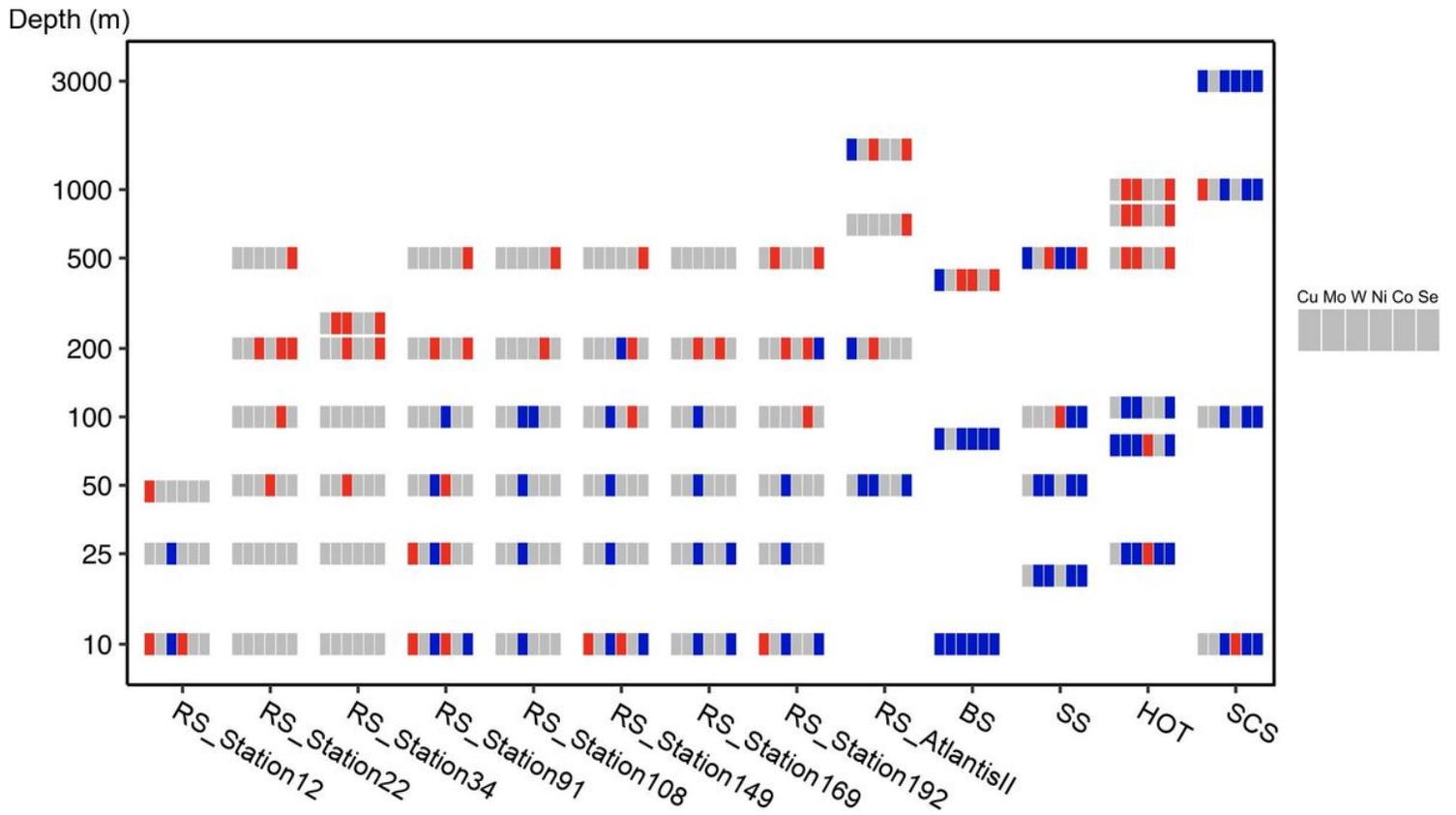
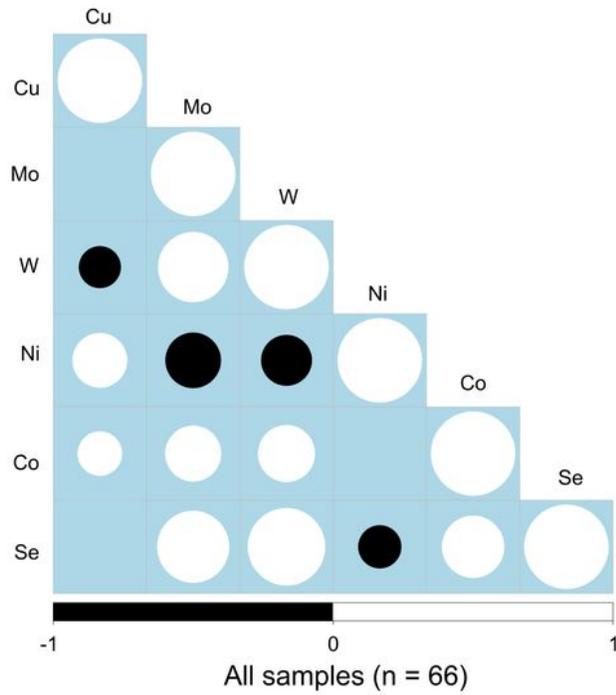
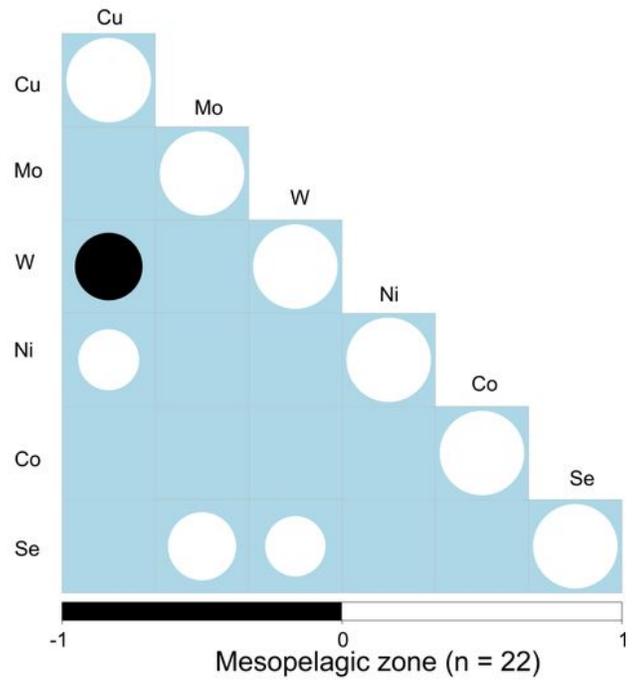
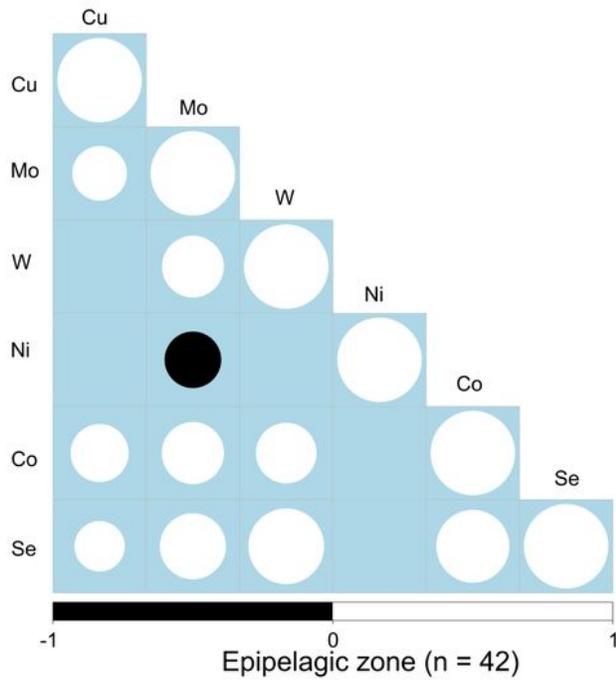


Figure 2

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.

A**B****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.

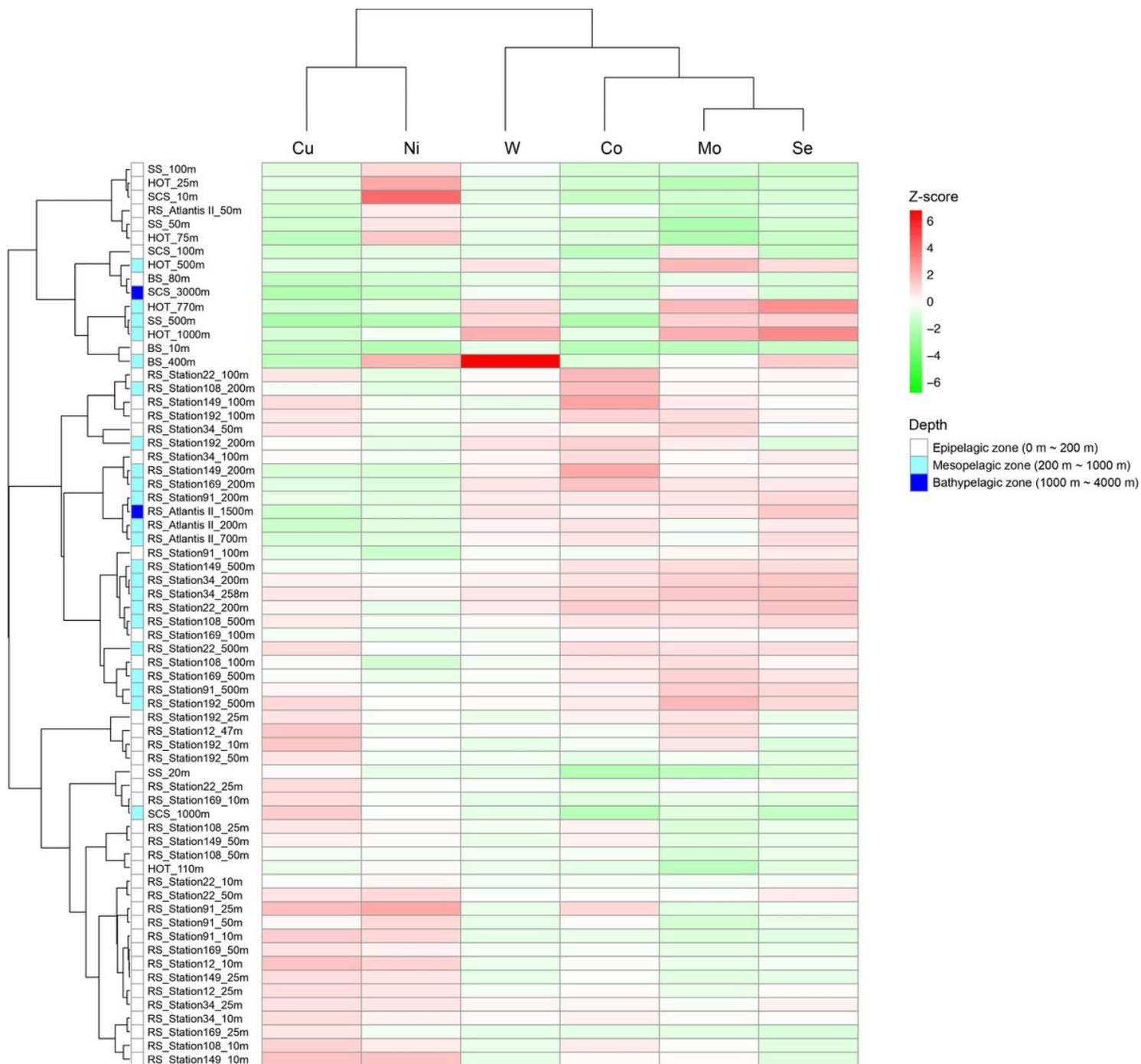


Figure 4

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.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementarytables.xlsx](#)
- [Supplementaryfigures.pdf](#)
- [sequencedataset.txt](#)