Electronic Supplementary Material (ESI) for Metallomics

Supplementary Information for

Histidine-rich proteins/motifs in prokaryotes: metal homeostasis and environmental habitat-related occurrence

Tianfan Cheng[1‡], Wei Xia[1‡], Panwen Wang[2], Feijuan Huang[1], Junwen Wang[2,3], Hongzhe Sun[1*]

[1]Department of Chemistry, The University of Hong Kong, Pokfulam Road., Hong Kong, P.R. China.

[2]Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd., Hong Kong, P.R. China.

[3]Centre for Genomic Sciences, LKS Faculty of Medicine, The University of Hong Kong, 21 Sassoon Rd., Hong Kong, P.R. China.

[*] To whom correspondence should be addressed: E-mail: hsun@hku.hk

[‡] These authors contribute equally to the paper.

Supplementary Figures:

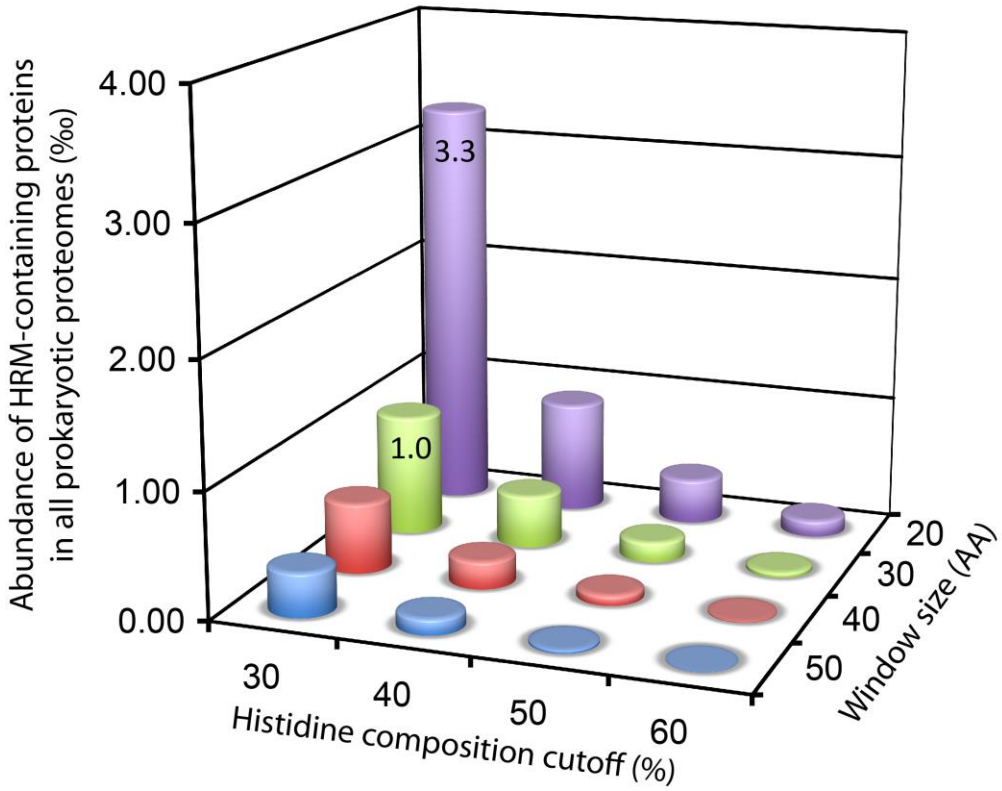


Fig. S1 Histidine rich protein (HRP) identification in prokaryotic proteomes by PSAT program. The window-frame size varied from 20 to 50 AA and the histidine content cutoff value ranged from 30% to 60%. The normalized abundances (‰) of HRPs in all available prokaryotic proteomes are shown in column chart. The highest abundance derived from least stringent criterion (20 AA/30%) and the abundance for 30 AA/30% setting are annotated accordingly.

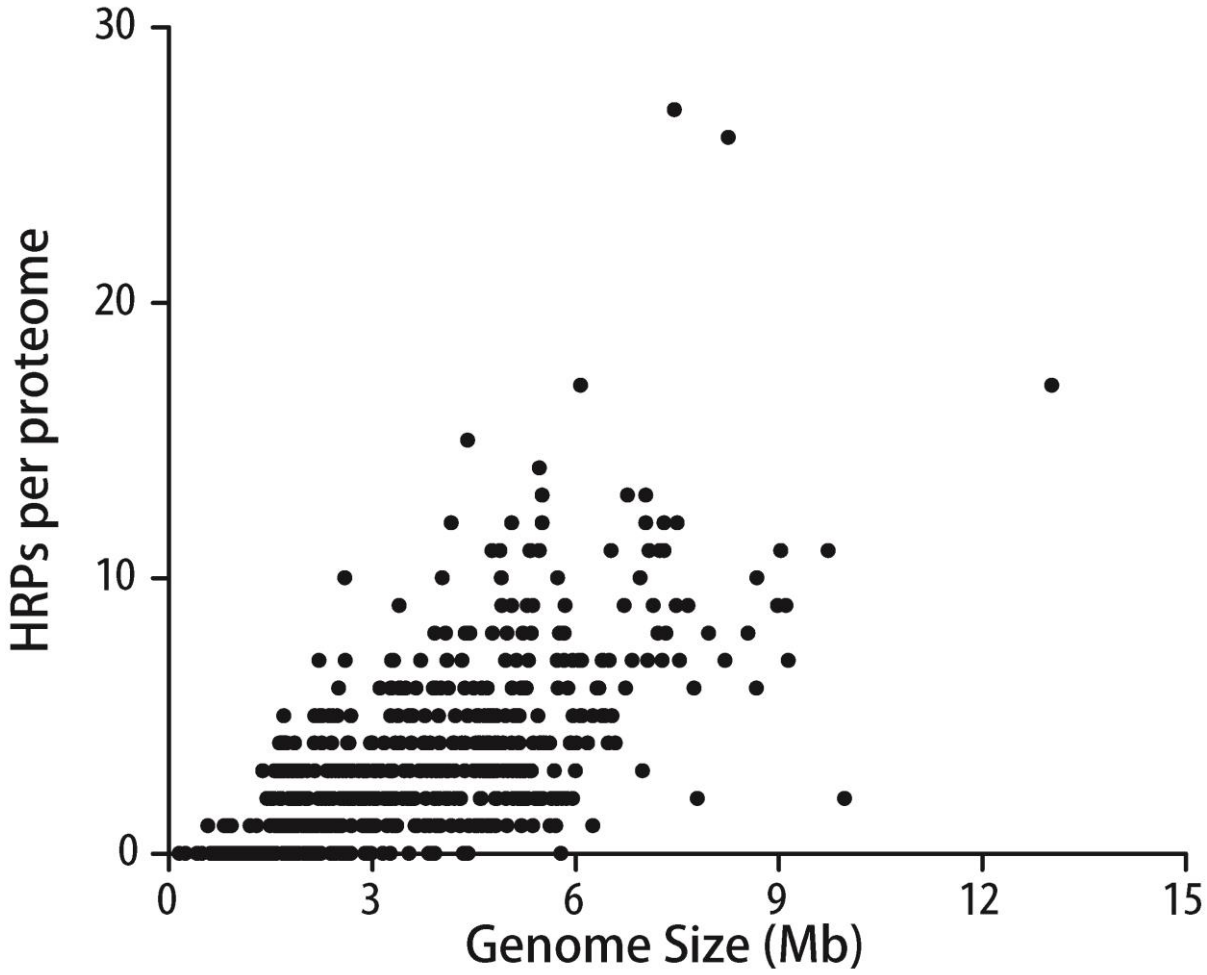
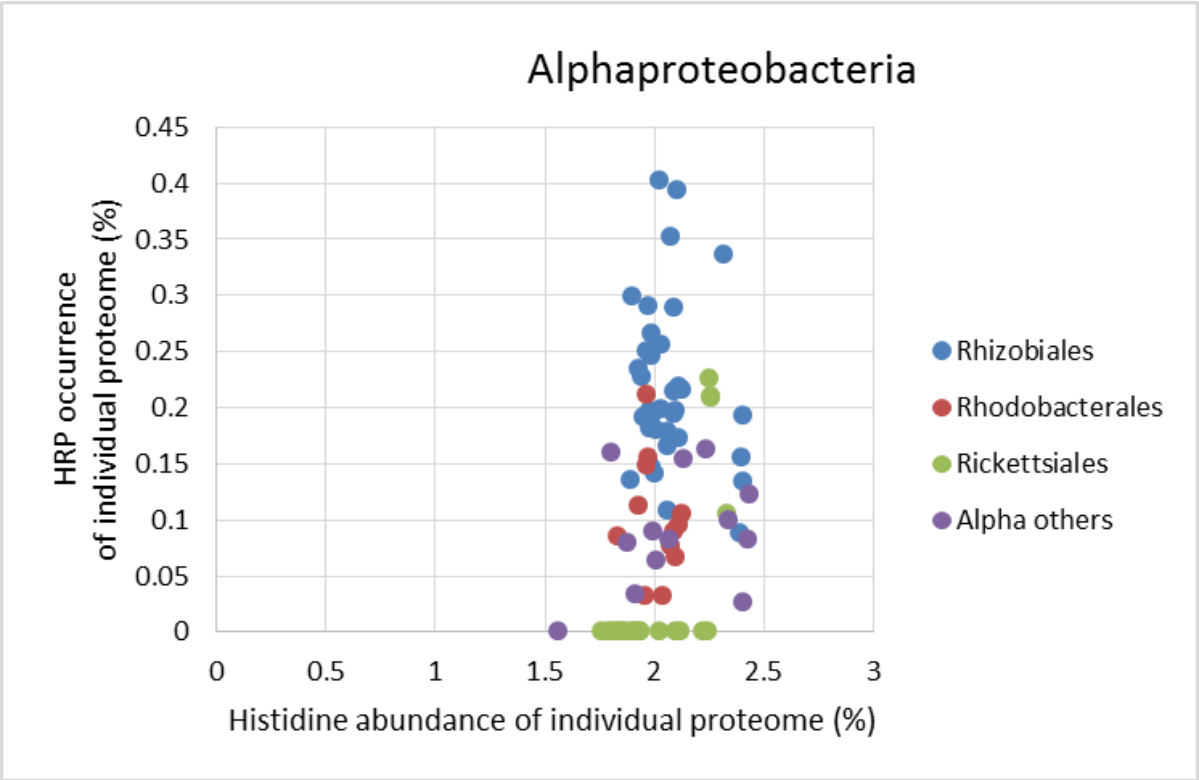
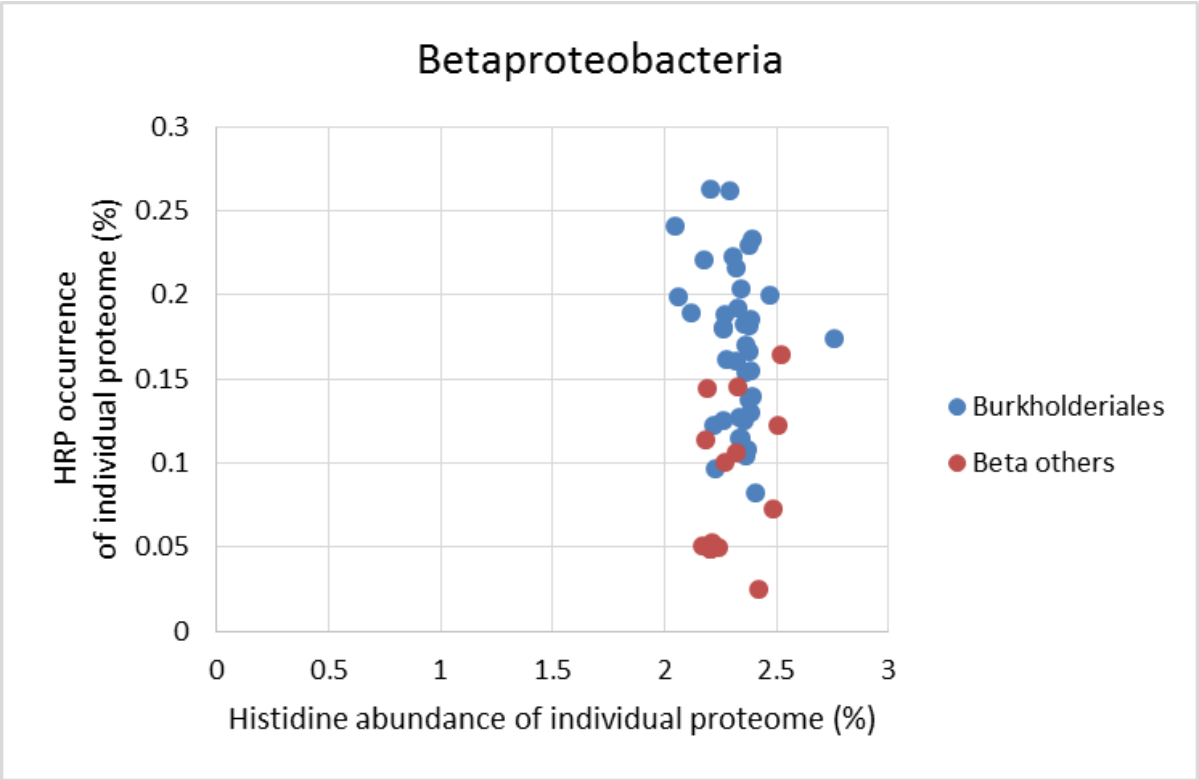
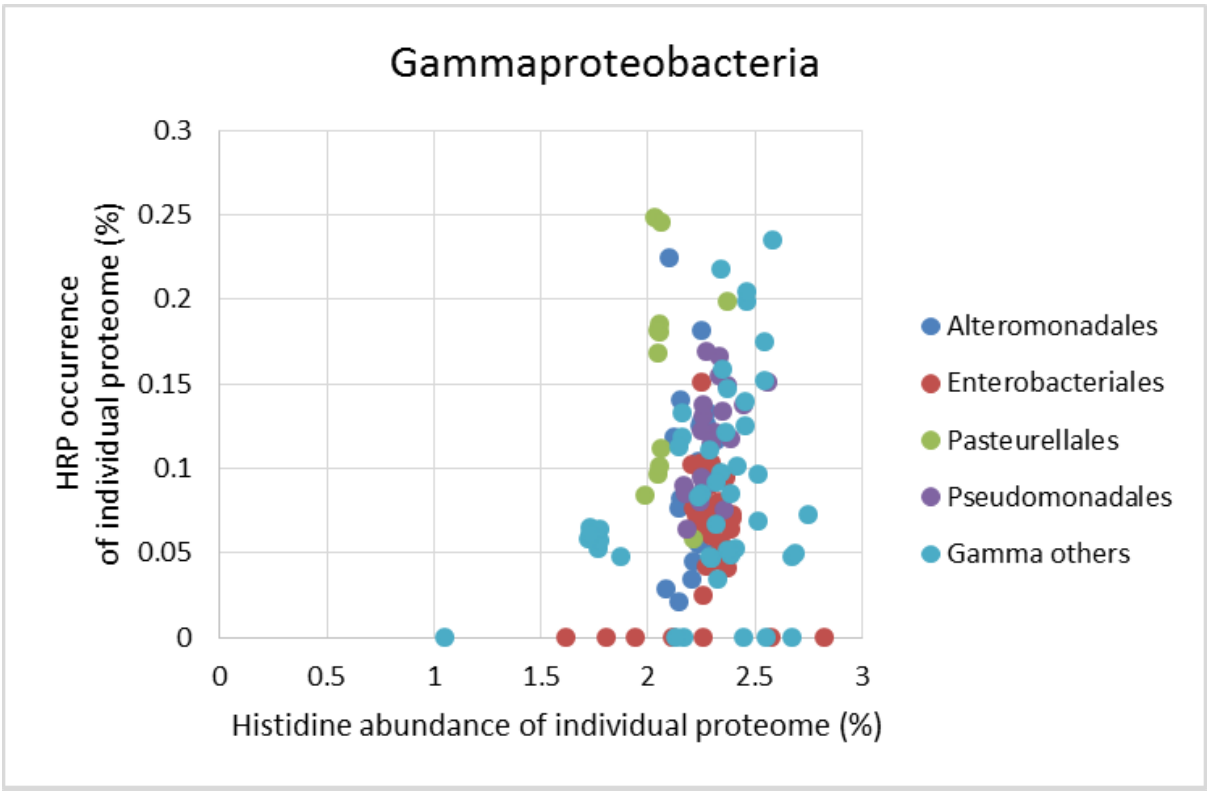
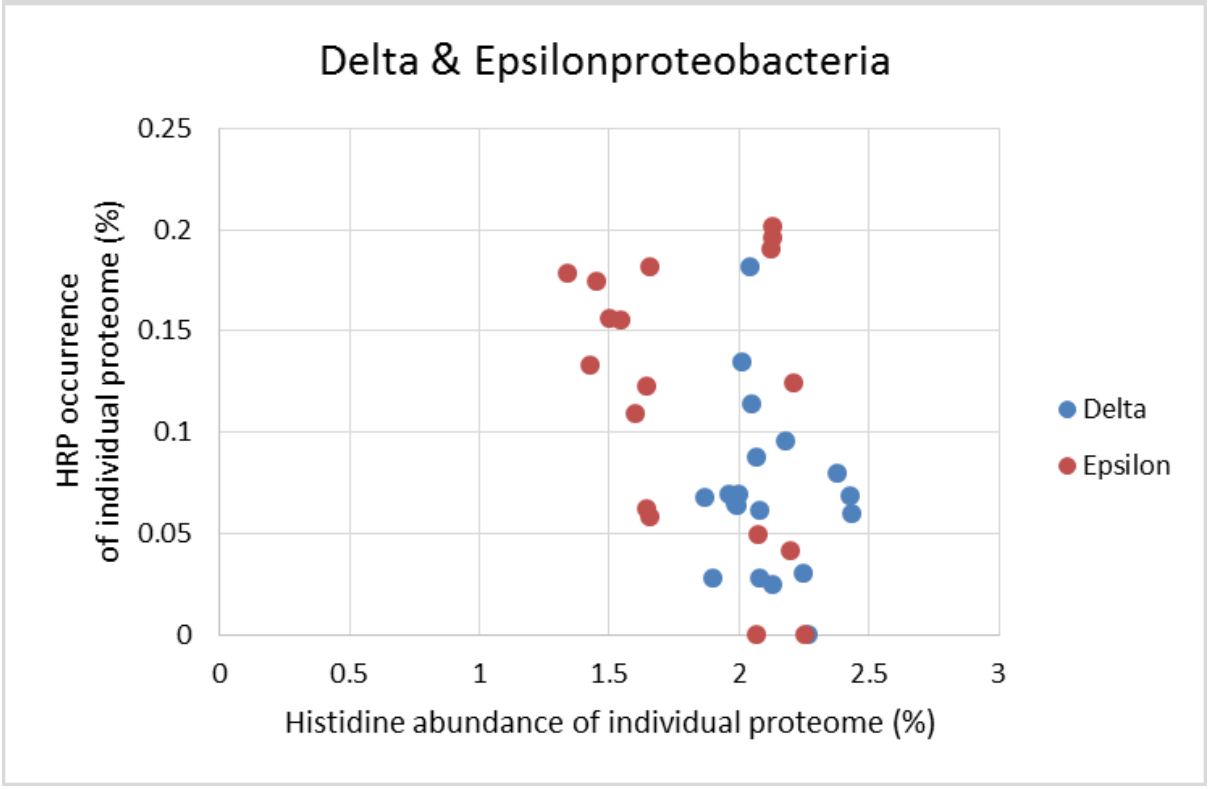


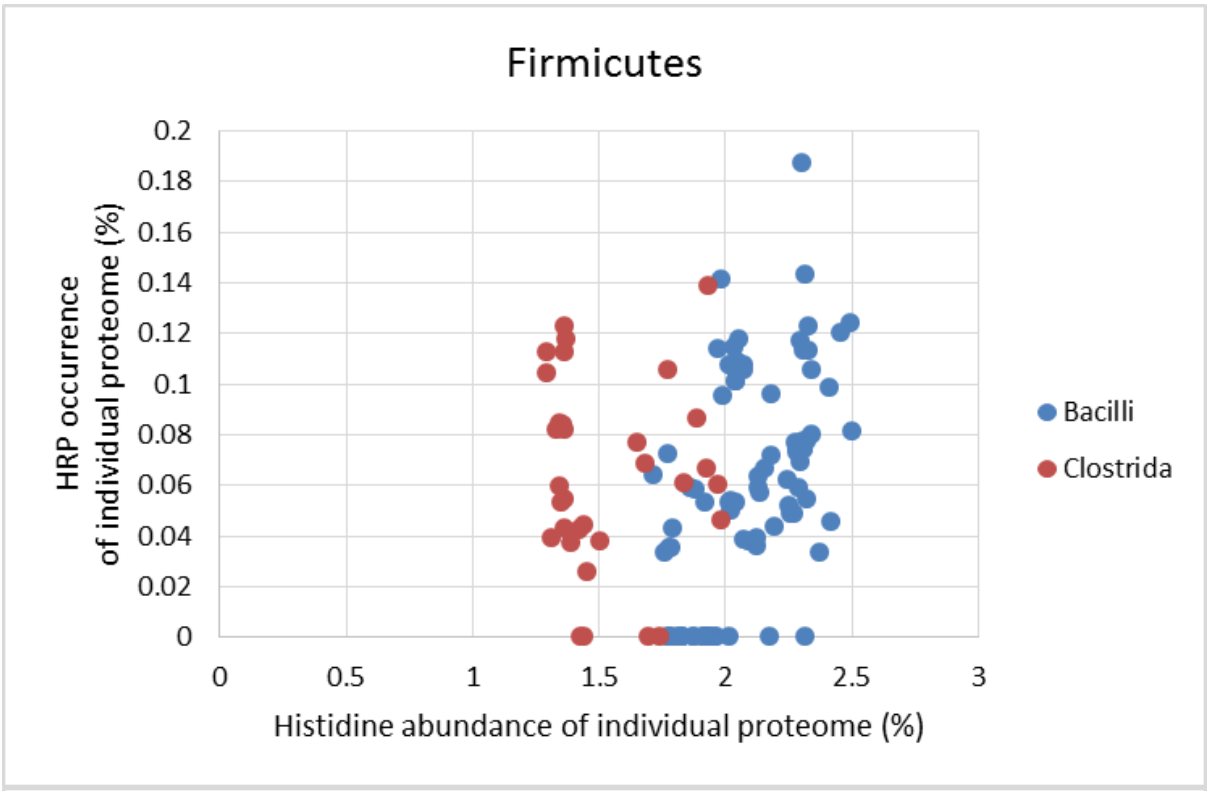
Fig. S2 The number of HRPs per prokaryotic proteome vs. chromosomal genome size. HRPs were searched by PSAT with the setting of 30 AA/30% (Table S1). Genome size was defined as the base pair numbers of chromosome(s) excluding plasmids and phages.

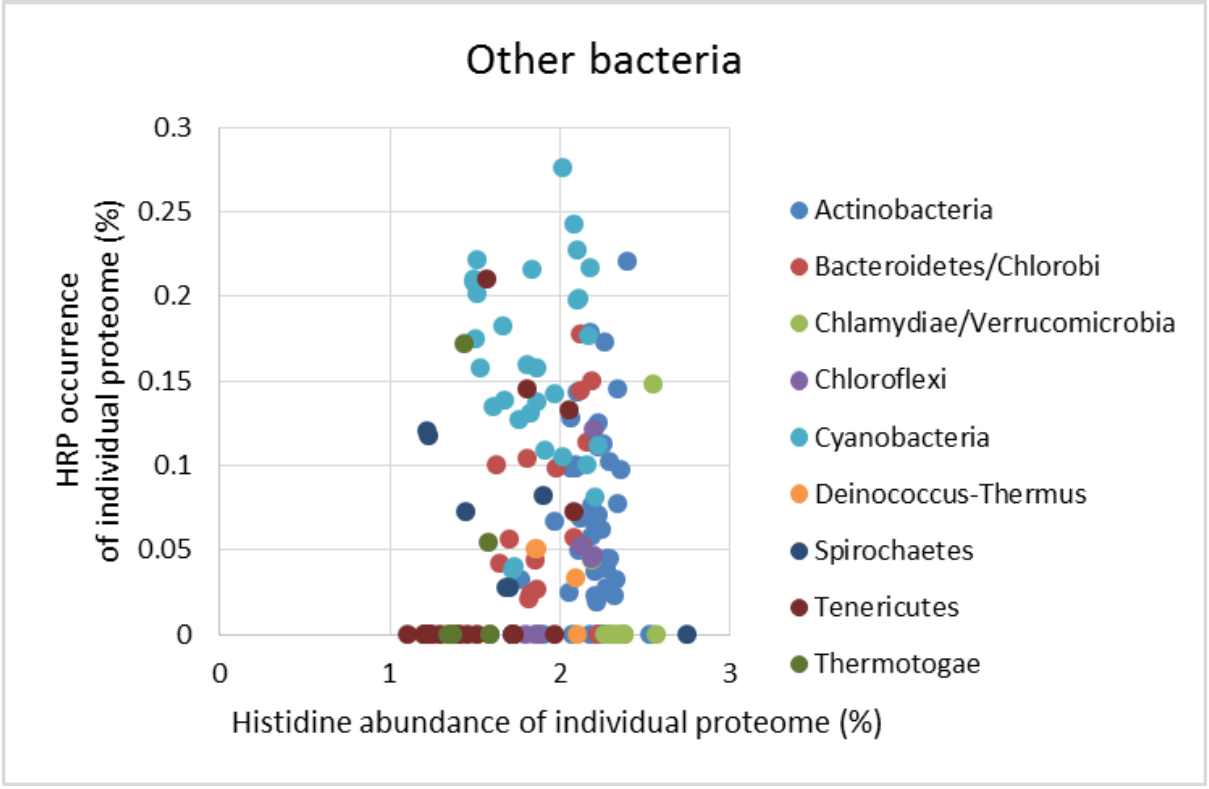












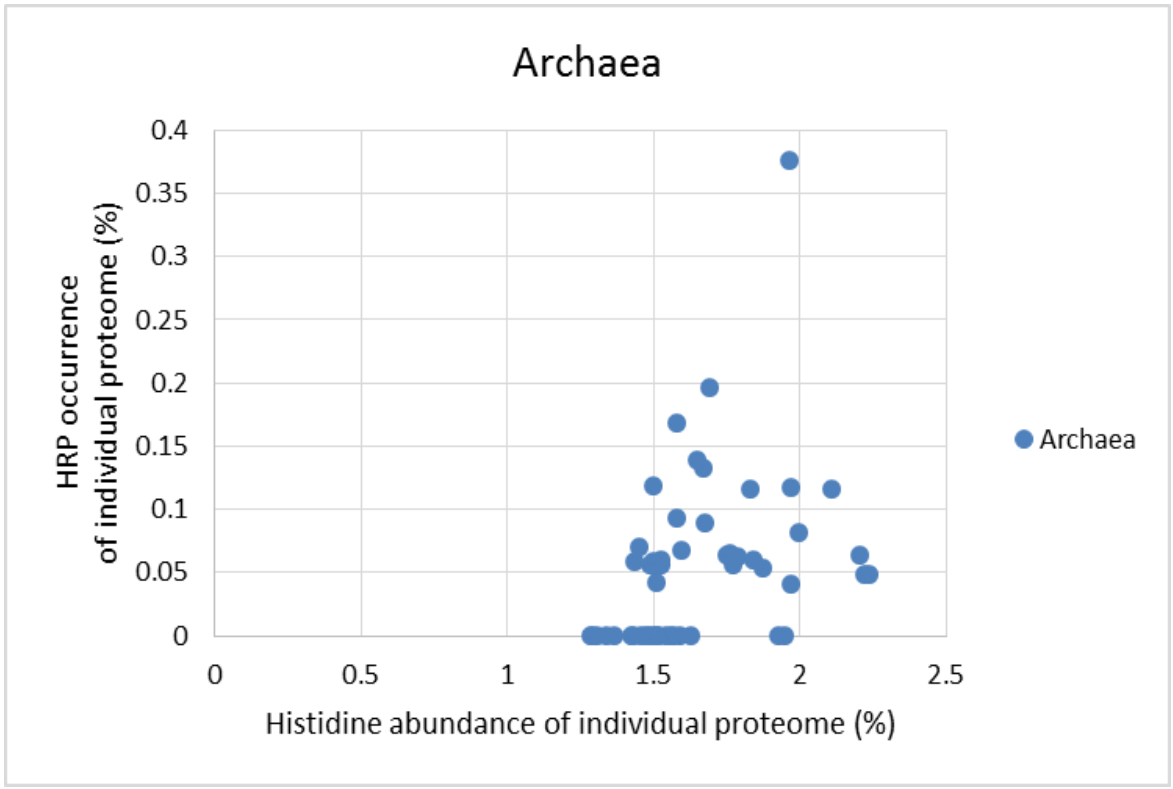


Fig. S3 The relationships between the occurrence percentage of HRPs and histidine abundance for individual proteomes of species from different taxonomical groups. HRPs were searched by PSAT with the setting of 30 AA/30% (Table S1). No significant correlations.

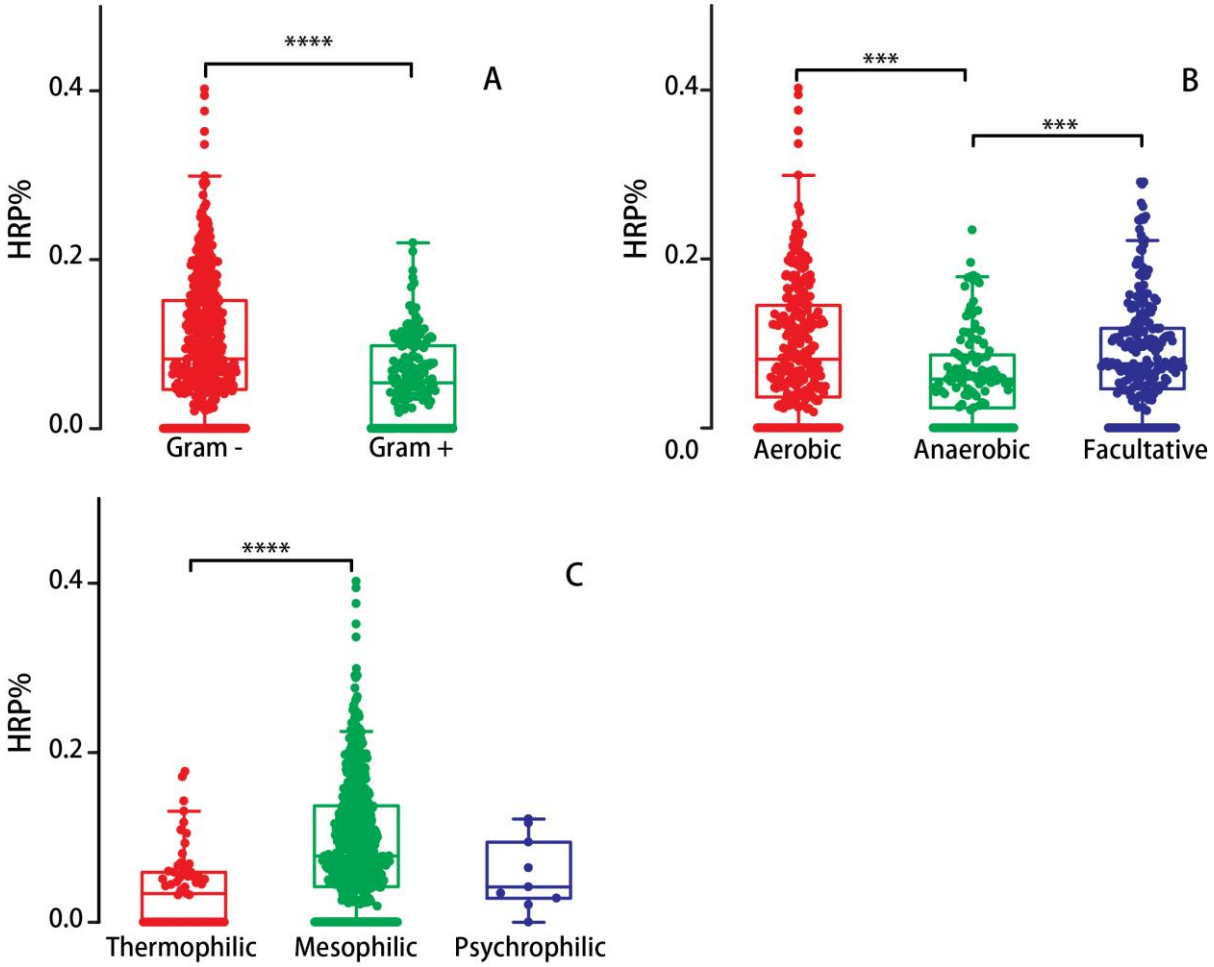


Fig. S4 The occurrence of HRPs in bacterial species is related other properties of species: (**A**) Bacteria types base on Gram staining, Gram negative (Gram-), Gram positive (Gram+); (**B**) Oxygen requirement; (**C**) Growth temperature range. The occurrence of HRPs per proteome is plotted as dots for individual habitat category and in box and whisker with the quantiles 5%, 25%, 50% (median), 75% and 95%. The annotations with "****" are for $P<0.0001$ and "***" for $P<0.0005$ (Mann-Whitney tests).

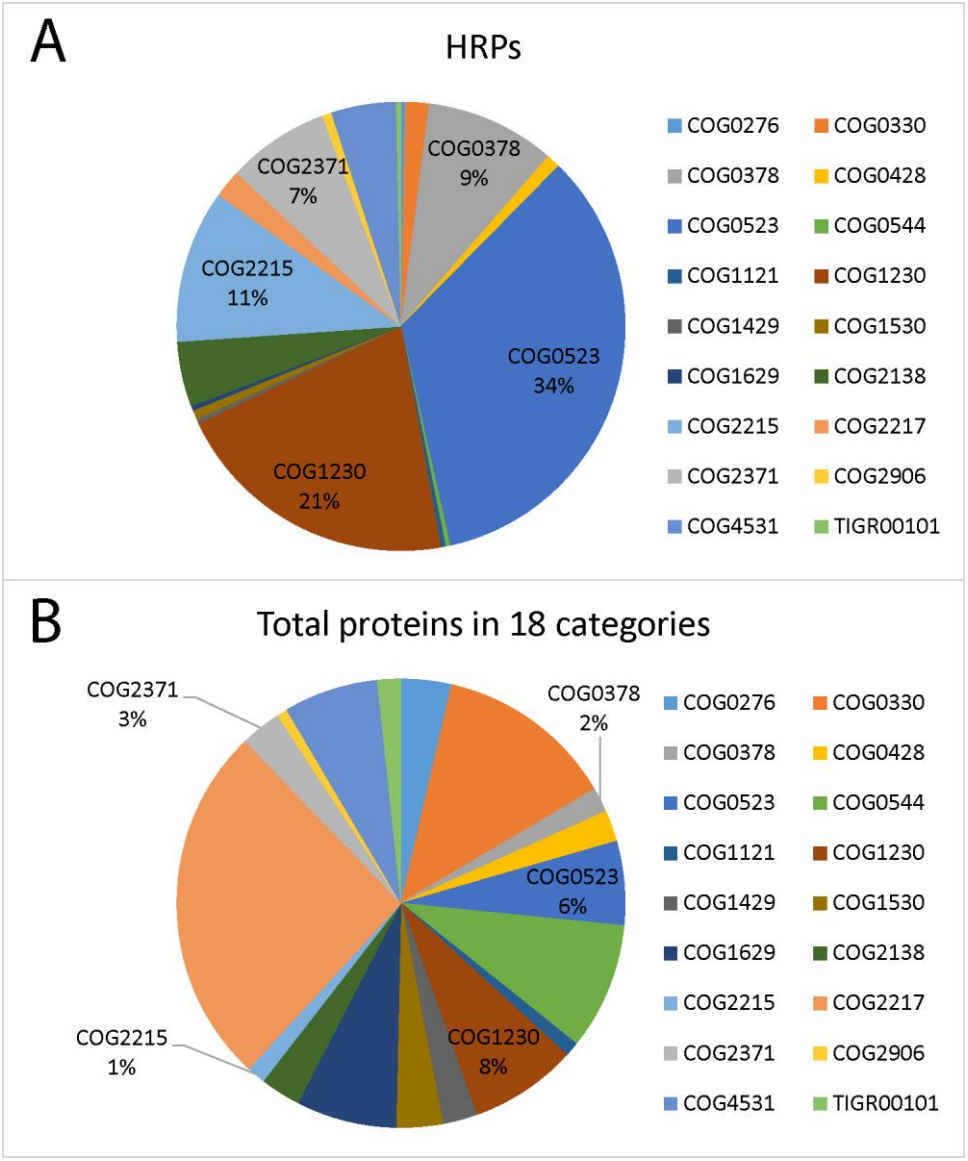


Fig. S5 Distribution of proteins in the 18 categories (as shown in the legends, table S6), excluding TIGR00900 (major facilitator superfamily, MFS). (**A**) Classification of identified HRPs (271 proteins) for CDD groups. (**B**) Classification of total proteins for each CDD group from a total of 675 proteomes. Five main protein categories in (a) are highlighted with COG numbers in both figures.

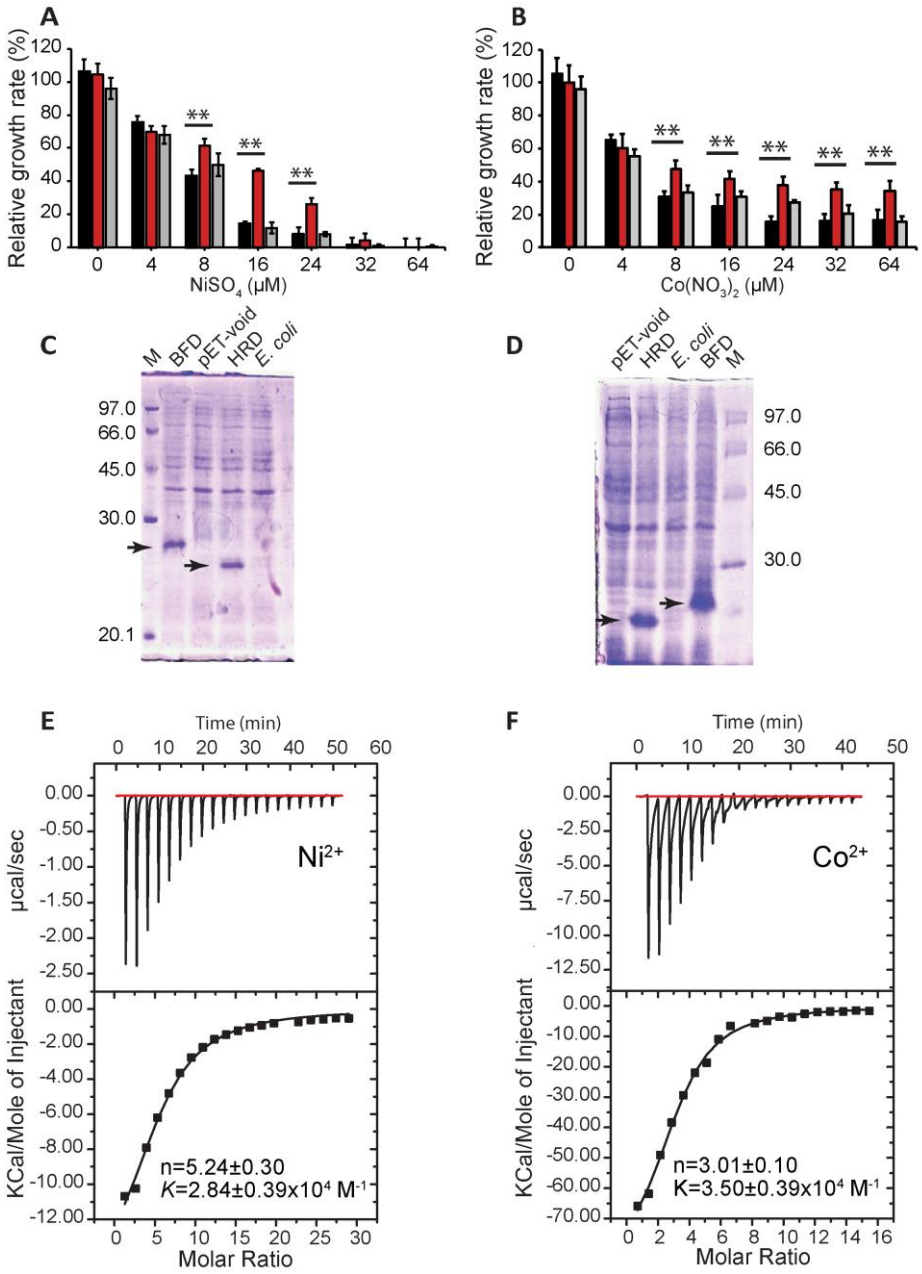


Figure S6 Metal resistance conferred by the overexpression of HRD of BFD and metal binding properties of BFD. (A) Ni^{2+} and (**B**) Co^{2+} resistance of *E. coli* cells with the expression of HRD (■) and FeSD (■) or void plasmid (■). The significance of difference is examined by Student's *t*-test ("**", $P<0.05$). SDS-PAGEs of *E. coli* cells cultured in the presence of 10 μM (**C**) Ni^{2+} and (**D**) Co^{2+}, without plasmid (*E. coli*), with tag-less pET-32a(+) plasmid (pET-void), with BFD-expressing plasmid (pET-BFD) and with HRD-expressing plasmid (pET-HRD). M: protein MW standard. Determination of binding constants of BFD toward Ni^{2+} (**E**) and Co^{2+} (**F**) by ITC in 25 mM Hepes, 125 mM NaCl with 0.5 mM TCEP at pH 7.2.

Supplementary Tables:

Table S1 Summary of HRM-containing proteins from a total of 675 bacterial and archaeal species. HRPs were searched by PSAT with the setting of 30 AA/30%. Taxonomical group, species name, total protein number, HRP number, HRP percentage, total amino acid number, total histidine number, histidine percentage, bacteria type (Gram staining), oxygen requirement and growth temperature range for each proteome are listed.

Additional file for Table S1.

Table S2 Protein sequences with the highest histidine contents motif in a 30 AA frame.

Species Name	*Acinetobacter baumannii* ATCC 17978
GenBank accession number	YP_001085982.1
Protein Sequence	MQDYSVSRHHQHQFDEGNPLAQKRILIATILTASMMVLEVFGGWF FNSMALLADGWHMSSHMLALGLAYFAYRAARHYSNDHRFSFGT WKIEILAGYSSAILLMVVAIFMAFQSVQRLFNPVEIFYNEAIPIAILG LVINLICAWLL**HDDGHHHHHHHHHHHHHHHHHHHHHEHGHHHH** DLNQKAAFLHVVADAVTSVFAIVALFAGKYFGWDFLDALLGILGA ILVAKWSFGLMKETGKTLLDAEMDHPVVDEIREVIAEFPKHVEITD IHVWKVAKGKFSCILALETDDISLNADQIRDALSIHDEIVHISVEINT LKPVYVPRETLA
Species Name	*Acinetobacter baumannii* AYE
GenBank accession number	YP_001715600.1
Protein Sequence	MTHLFSLVQFNILVGSILGLNFMQDYSVSRHHQHQFDEGNPLAQK RILIATILTASMMVLEVFGGWFFNSMALLADGWHMSSHMLALGLA YFAYRAARHYSNDHRFSFGTWKIEILAGYSSAILLMVVAIFMAFQSI QRLFNPVEIFYNEAIPIAILGLVINLICAWLLHDDG**HHHHHHHHHH HHHHHHHHHEHGHGHHHHDL**NQKAAFLHVVADAVTSVFAIVA LFAGKYFGWDFLDAILGILGAILVAKWSFGLMKETGKTLLDAEMD HPVVDEIREVIAEFPKHVEITDIHVWKVAKGKFSCILALETDDISLN ADQIRDALSIHDEIVHISVEINTLKPVYVPRET LA

Table S3 Summary of the niches of a total of 675 bacterial and archaeal species. Certain niche(s)/habitat(s) annotated in the Genomes OnLine Database (GOLD) (www.genomesonline.org) from the minimum information about a genome sequence (MIGS) specification. The "**1**" in a cell represents that one species belongs to the corresponding niche. The number of species for each niche was counted and is listed in the sub-table of "Habitat keywords". Each niche is attributed to one of six habitats.

Additional file for Table S3.

Table S4 Observed frequency of amino acids in histidine-rich-motifs (HRMs) sequences and all bacterial proteomes. The HRMs are identified by PSAT program with a window

size of 30 amino acids and cutoff of 50% (30 AA/50% details in Table S5). The observed average frequency of amino acids is retrieved from UniProt (http://www.uniprot.org).

	Observed frequency in HRMs (%)	**Average frequency (%)**
Histidine (H)	55.2	2.3
Aspartic acid (D)	14.5	5.5
Glycine (G)	9.8	7.1

Table S5 Summary of functions of a total of 373 HRPs. HRPs were search by PSAT with the setting of 30 AA/50%. The functional or putative annotation from NCBI Conserved Domain Database (CDD) was examined and further divided into several different subgroups. The Clusters of Orthologous Groups (COG) numbers are also noted when available. For COG0523's, subgroup and possible molecule related was also identified. The sequence of the identified histidine richest motif for the corresponding protein is also listed.

Additional file for Table S5.

Table S6 Summary of the occurrence of HRM-containing proteins for each CDD group. HRPs were searched by PSAT with the setting of 30 AA/50%. The number of HRPs for each CDD group in table S4 was counted and is listed. The total protein number for each CDD group was counted by keyword searching against the GenBank annotations for a total of 675 proteomes.

Additional file for Table S6.

Table S7 Summary of Pfam searching. All the protein sequences obtained from PSAT search with 30 AA/30% setting were searched in Pfam database.

Additional file for Table S7.