scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Mining of novel secondary metabolite biosynthetic gene clusters from acid mine drainage

Ling Wang[1,2,6], Wan Liu[2,6], Jieliang Liang[3,6], Linna Zhao[2], Qiang Li[2,4], Chenfen Zhou[2], Hui Cen[2], Qingbei Weng[1,5,7] ✉ & Guoqing Zhang[1,2,7] ✉

Acid mine drainage (AMD) is usually acidic (pH < 4) and contains high concentrations of dissolved metals and metalloids, making AMD a typical representative of extreme environments. Recent studies have shown that microbes play a key role in AMD bioremediation, and secondary metabolite biosynthetic gene clusters (smBGCs) from AMD microbes are important resources for the synthesis of antibacterial and anticancer drugs. Here, 179 samples from 13 mineral types were used to analyze the putative novel microorganisms and secondary metabolites in AMD environments. Among 7,007 qualified metagenome-assembled genomes (MAGs) mined from these datasets, 6,340 MAGs could not be assigned to any GTDB species representative. Overall, 11,856 smBGCs in eight categories were obtained from 7,007 qualified MAGs, and 10,899 smBGCs were identified as putative novel smBGCs. We anticipate that these datasets will accelerate research in the field of AMD bioremediation, aid in the discovery of novel secondary metabolites, and facilitate investigation into gene functions, metabolic pathways, and CNPS cycles in AMD.

Background & Summary

Acid mine drainage (AMD) is a type of acidic (pH < 4) and metal-enriched water that results from the accelerated oxidative dissolution of exposed minerals, principally sulfides, and is associated with mining[1,2]. The strong acidity and heavy metal toxicity of AMD has caused severe pollution to surrounding water systems and soils[2–4], making AMD one of the most serious environmental problems arising during the mining of mineral resources[5,6]. Metabolically-active acidophilic microorganisms have been observed in AMD[7,8], including microbes primarily from the Bacteria (such as Proteobacteria, Nitrospirae, Actinobacteria, Firmicutes, and Acidobacteria) and Archaea domains[9].

Microbes in AMD play a key role in the bioremediation of AMD environments[10,11]. For example, *Acidithiobacillus*[12], one of the most common genera in AMD, includes microbes with chemolithotrophic metabolisms that are able to oxidize Fe^{2+} and sulfur compounds (such as *Acidithiobacillus ferrooxidans*, *Acidithiobacillus ferridurans*, and *Acidithiobacillus ferrivorans*)[9,13,14], or oxidize sulfur compounds alone (such as *Acidithiobacillus caldus*, *Acidithiobacillus thiooxidans*, and *Acidithiobacillus albertensis*)[15–17]. Sulfate-reducing bacteria (SRB), a group of diverse anaerobic microorganisms that are ubiquitous in natural habitats, have been utilized in AMD remediation[11].

Secondary metabolite biosynthetic gene clusters (smBGCs) found in AMD microbes are important resources for the synthesis of antibacterial and anticancer drugs[18,19]. A previous study reported that microbes including *Penicillium* sp., *Penicillium rubrum*, *Penicillium solitum*, *Penicillium clavigerum*, *Chaetomium funicola*, and *Pithomyces* sp. were separated and cultivated from water and sediment samples in a pit lake formed by the former Berkeley copper mine, among which worthwhile secondary metabolites were found[20]. For example, berkelic

[1]School of Life Sciences, Guizhou Normal University, Guiyang, Guizhou, 550025, China. [2]National Genomics Data Center, CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai, 200031, China. [3]School of Life Sciences, South China Normal University, Guangzhou, Guangdong, 510631, China. [4]Suzhou BiomeMatch Therapeutics Co., Ltd. No.351, Guoshoujing Road, Shanghai, 201203, China. [5]Qiannan Normal University for Nationalities, Duyun, Guizhou, 558000, China. [6]These authors contributed equally: Ling Wang, Wan Liu, Jieliang Liang. [7]These authors jointly supervised this work: Qingbei Weng, Guoqing Zhang. ✉e-mail: wengqb@126.com; gqzhang@picb.ac.cn

Mineral type	Sample number	Base number (Gb)	Base number per sample (Gb)	Country	Data source
Antimony	8	500.66	62.58	China	NODE: OEP001841
Arsenic	3	30.60	10.20	China	NODE: OEP001841
Coal	13	78.31	6.02	China and USA	SRA: SRP218093, SRP226684
Copper	39	1,660.60	42.58	Brazil, China, Germany, United Kingdom, and USA	NODE: OEP001841; SRA: SRP093762, SRP149873, SRP201756, SRP288126
Iron	17	62.81	3.69	USA	SRA: SRP009106
Lead-Zinc	24	1,638.82	68.28	China	NODE: OEP001841; SRA: ERP002170
Lignite	1	5.00	5.00	Germany	SRA: SRP093591
Magnetite	5	598.44	119.69	China	NODE: OEP001841
Nickel-Copper	15	49.15	3.28	Canada	SRA: SRP102076
Polymetallic	33	2,495.94	75.62	China and Sweden	NODE: OEP001841; SRA: SRP132763
Pyrite	13	477.59	36.74	China and Germany	NODE: OEP001841; SRA: SRP096619
Pyrite-Copper	6	454.49	75.75	China	NODE: OEP001841
Tin-Zinc	2	163.63	81.81	China	NODE: OEP001841

Table 1. Data information for each mineral type. Giga base is a unit of length for DNA molecules, consisting of one billion nucleotides; abbreviated Gb, or Gbp for giga base pair (http://en.wikipedia.org/wiki/Base_pair).

acid, a secondary metabolite of *Penicillium sp*., had anti-OVCAR-3 activity in *NCI-DTP60*; berkeleydione, the terpenoid secondary metabolite of *Penicillium rubrum*, showed selective activity against non-small cell lung cancer NCI-H460 in *NCI-DTP 60*; and $CHCl_3$ extracted from *Penicillium solitum* strongly inhibited MMP-3 and caspase-1. In addition, cyclodipeptide synthases (CDPSs) that were capable of synthesizing cyclodipeptide, a precursor of 2,5-diketopiperazines, were found to be produced by 23 metagenome-assembled genomes (MAGs) (LMSG_G000006317.1–LMSG_G000006339.1) in *Diplorickettsiaceae* in this study[21–23]. Therefore, mining smBGCs from AMD may reveal valuable secondary metabolites[18].

In this study, data were collected and the GTDB species representative assignment for the binned MAGs and putative novel smBGCs of 111 samples from nine mineral types were analyzed. The same method was used to reanalyze public metagenomic datasets consisting of 68 samples of eight mineral types from seven countries. In total, this study obtained the analysis results of metagenomic datasets covering 179 samples of 13 projects across 13 mineral types from seven countries (Table 1, Supplementary Table 1, Figs. 1a,2). A total of 7,007 MAGs mined from the datasets exceeded the medium-quality level of the MIMAG standard[24], including 981 MAGs determined to be high quality (Table 2, Supplementary Table 2). Further taxonomic analysis by GTDB-Tk showed that 1,394 MAGs were classified into 150 existed genera, while 5,613 MAGs were not assigned to existed genera; total of 667 MAGs could be assigned to 154 GTDB species representatives, while 6,340 MAGs were not assigned (Fig. 3, Supplementary Table 2). Overall, 11,856 smBGCs in eight categories were obtained from 7,007 MAGs (Table 3, Supplementary Table 3), and 10,899 smBGCs were identified as putative novel smBGCs for discovering novel secondary metabolites by querying each smBGC sequence against the NCBI nucleotide sequence collection (Supplementary Table 3). The analysis of the number of smBGCs in all mineral types showed that the greatest number of smBGCs was found in polymetallic mines, and the second largest number was found in copper mines. The descending order of smBGC abundance in the remaining mineral types was as follows: lead-zinc mines, antimony mines, pyrite-copper mines, pyrite mines, coal mines, nickel-copper mines, magnetite mines, tin-zinc mines, iron mines, arsenic mines, and lignite mines (Fig. 4a).

Methods

The workflow of data processing is depicted in Supplementary Fig. 1.

Date source. AMD metagenomic datasets of 179 samples from 13 mineral types obtained from seven countries were used to analyze GTDB species representative assignment for the binned MAGs and putative novel smBGCs (Table 1, Supplementary Table 1, Figs. 1a,2), including 68 public and 111 private samples. The datasets of 68 publicly available samples were downloaded from the SRA database (up to November 17, 2020) using the following search strategies: (((((Mine AMD) OR acid mine drainage) OR mine tailings) OR acidic stream) AND WGS [Strategy]) AND METAGENOMIC [Source] and (mine drainage metagenome [Organism]) AND WGS [Strategy] AND METAGENOMIC [Source], and the Illumina sequence data were kept. A total of 111 private samples across nine mineral types were collected and sequenced in this study. Among them, 87 samples across four mineral types newly collected in this study came from the same mineral types as the datasets downloaded from the SRA database, and 24 samples were obtained from five new mineral types. A total of 122 samples from 10 mineral types constituted the AMD metagenomic datasets for China (Table 1, Fig. 1b).

Quality control of raw data and metagenomic assembly. Trimmomatic is a flexible and efficient preprocessing tool used for reads processing of Illumina next-generation sequencing data, primarily for the filtering of adapter and low-quality sequences[25]. Quality control of the raw data for 179 samples in this study was performed using Trimmomatic (version 0.39) with Phred quality score cutoff of 20 and a minimum read

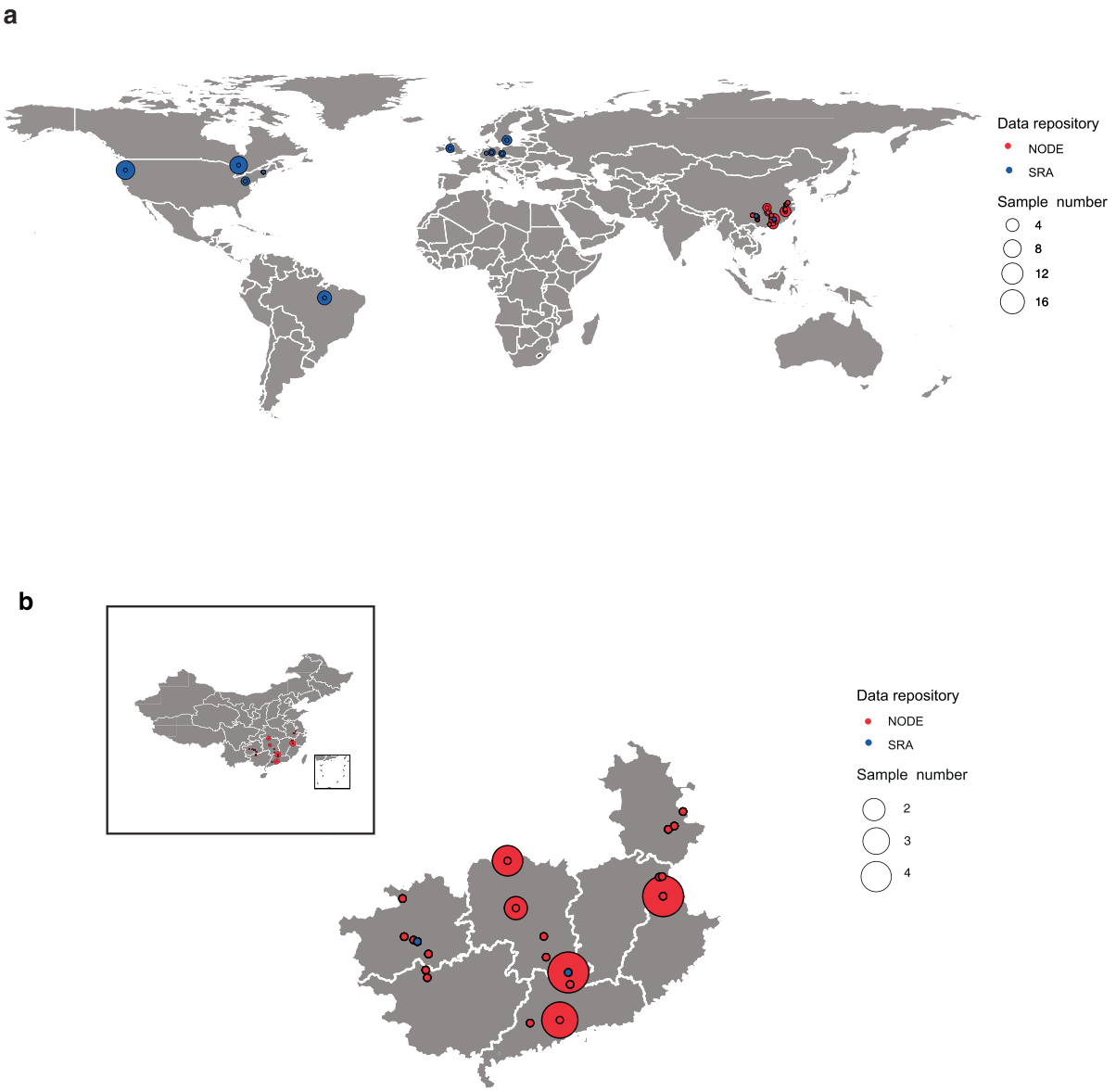

Fig. 1 Geographic distribution of sampling sites in this study. (**a**) Geographic distribution of sampling sites for all samples (the latitude and longitude of SRS1810936 was retrieved according to the geographic location of this sample). (**b**) Geographic distribution of sampling sites for the acid mine drainage (AMD) metagenomic datasets for China.

length of 50 to remove the low-quality sequences. MetaSPAdes performs better in assembly compared to the other assembly tools, but it is time-consuming and requires very high memory[26,27]. MEGAHIT and metaSPAdes are both widely used tools for metagenome assembly[28–30]. Although metaSPAdes can provide high-quality assemblies across diverse data sets, MEGAHIT can provide acceptable assemblies with low memory usage and computational time[31]. Therefore, by a comprehensive consideration of the large volume of AMD samples analyzed and the affordable computational resources, we chose MEGAHIT[28,29] as the software for metagenome assembly. The analysis of metagenome assembly was performed by MEGAHIT (version 1.2.9) in meta-sensitive mode to generate assembled contigs.

Metagenomic binning. Compared to original binning software, automated methods with multiple binning methods, such as MAGO, MetaWRAP or DAS Tool, combine the strengths of a flexible set of established binning algorithms to generate more or better bins[32–34]. MetaWRAP is a widely used tool for the metagenome binning of both environmental[35–41] and host-associated[42–44] samples, and it can obtain the largest number of high-quality draft genomes in tested datasets with relatively less computational requirements[33,45]. Additionally, MAGO used DAS Tool for bin refinement, and MetaWRAP outperformed DAS Tool for datasets of varied complexity[33]. Therefore, we selected MetaWRAP for metagenomic binning in this study. For each assembly, contigs were binned using the binning module (parameter: –maxbin2 –concoct –metabat2), consolidated into a de-replicated

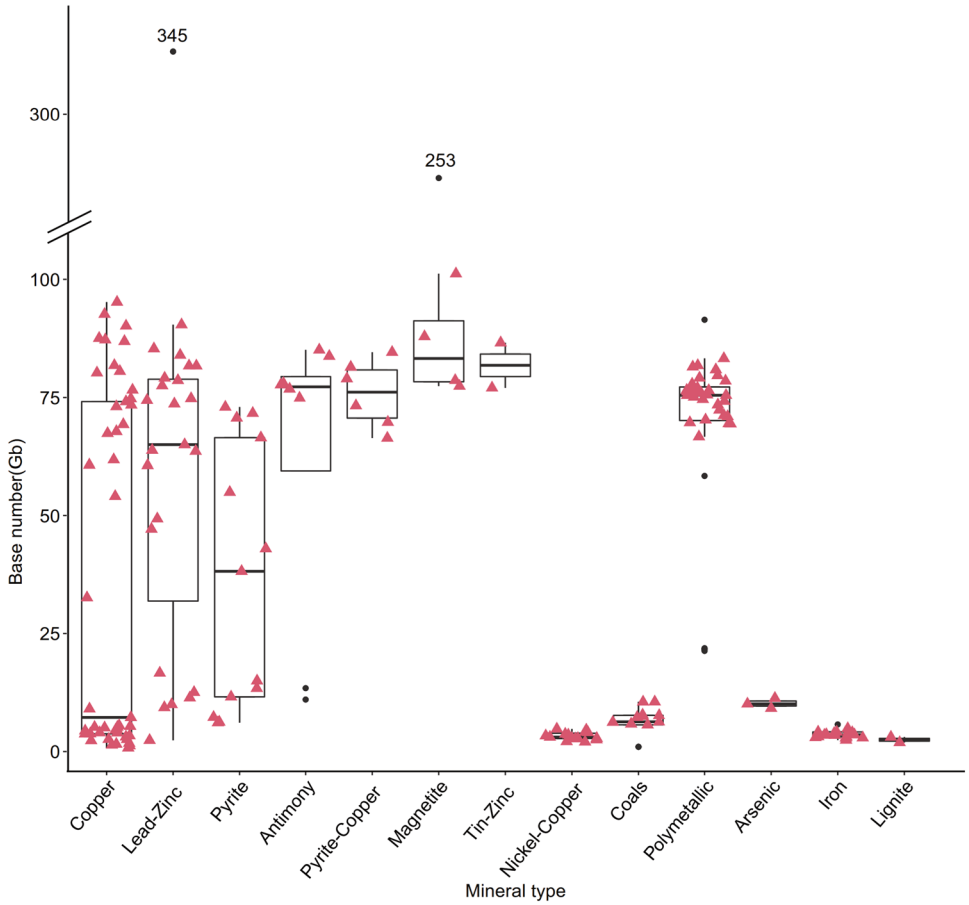

Fig. 2 Base number distributions of samples from 13 types of minerals. The median base number of samples was similar among lead-zinc mines, antimony mines, pyrite-copper mines, magnetite mines, tin-zinc mines, and polymetallic mines. The upper and lower whiskers extend from the hinge within 1.5 x the inter-quantile range to the highest and lowest values, respectively. The outlier points (black) are the ones outside that range.

Quality level	Completeness	Contamination	Quality score	Number
High quality	≥90%	≤5%	>50	981
Medium quality	50%~90%	≤5%	>50	6,026

Table 2. Quality control standards and metagenome-assembled genome (MAG) numbers in each quality level. High-quality MAG requires the presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.

bin set using the bin_refinement module (parameter: -c 50 -x 5), and the quality of bins was further improved by using the reassemble_bins module within MetaWRAP (version 1.3.2). A total of 8,035 binned MAGs were obtained from 179 samples by MetaWRAP taking 1224 hours of wall time using an HPC with multiple 2.10 GHz Intel Xeon E7-4380 CPUs and 2 TB of RAM.

The completeness and contamination of all MAGs were estimated using CheckM (version 1.1.2) with a lineage-specific workflow[46,47]. Based on these results, we selected 7,007 MAGs that were estimated to be at least 50% complete, with less than 5% contamination and that had a quality score of >50[36]. As additional indicators of completeness, we identified tRNA genes using tRNAscan-SE (version 2.0.9)[48] and rRNA genes using Infernal (version 1.1.2)[49] with models from the Rfam database[50]. Based on these results, we found that 981 of the 7,007 MAGs were classified as high quality based on the MIMAG standard (≥90% completeness, ≤5% contamination, ≥18/20 tRNA genes and the presence of 5S, 16S and 23S rRNA genes), with the remaining classified as medium quality (Table 2, Supplementary Table 2).

Taxonomic assignment for bacterial and archaeal genomes. GTDB-Tk is a computationally efficient tool providing objective taxonomic assignment for bacterial and archaeal genomes based on the Genome Taxonomy Database (GTDB, http://gtdb.ecogenomic.org), and it is widely used for the classification of draft genomes directly from environmental- and human-associated samples[51]. Taxonomic analysis of each MAG was initially assigned using GTDB-Tk (version 1.4.0) based on the GTDB taxonomy R05-RS95[52], and forty-eight

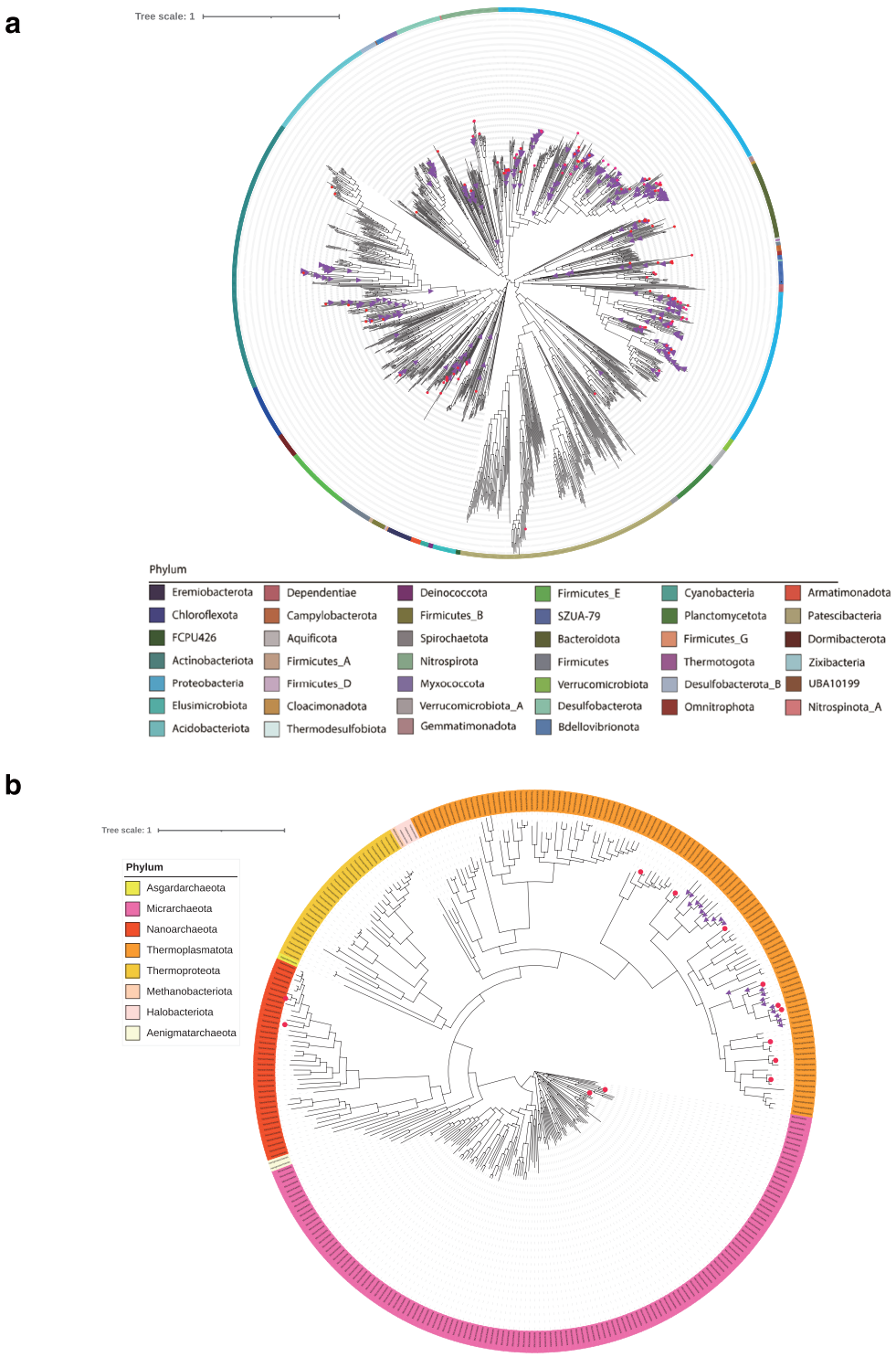

Fig. 3 Maximum-likelihood phylogenetic trees of bacterial and archaeal MAGs at the phylum level. Major lineages are assigned arbitrary colours and named. Lineages with GTDB representative species assignment are highlighted with red dots, while lineages with existed genera assignment (genus with NCBI taxonomy ID) are marked with purple triangles. (**a**) Maximum-likelihood phylogenetic trees of bacterial MAGs were inferred from a concatenated alignment of 120 bacterial single-copy marker genes. The tree includes 8 named archaeal phyla. (**b**) Maximum-likelihood phylogenetic trees of archaeal MAGs inferred from a concatenated alignment of 122 archaeal single-copy marker genes. The tree includes 40 named bacterial phyla.

phyla (eight archaeal phyla and 40 bacterial phyla) were obtained. GTDB-Tk analysis of 7,007 MAGs required 23 hours of wall time using an HPC with multiple 2.10 GHz Intel Xeon E7-4380 CPUs and 2 TB of RAM.

Type of smBGCs	Number of smBGCs	Percentage of smBGCs
Terpene	3,751	31.64%
RiPPs	1,864	15.72%
NRPS	1,738	14.66%
PKSother	936	7.89%
PKS I	250	2.11%
PKS-NRP_Hybrids	181	1.53%
Saccharides	1	0.01%
Others	3,135	26.44%

Table 3. Numbers and percentages of smBGCs in eight categories classified by BIG-SCAPE.

Based on the results of the GTDB-Tk analysis, a total of 1,707 MAGs were assigned to archaeal phyla, while 5,300 MAGs were assigned to bacterial phyla; 6,026 medium-quality MAGs were assigned to seven archaeal phyla and 38 bacterial phyla, while 981 high-quality MAGs were classified to four archaeal phyla and 31 bacterial phyla (Supplementary Table 2). In the genus level analysis, a total of 1,394 MAGs were classified into 150 extant genera, while 5,613 MAGs were not assigned. A total of 667 MAGs were assigned to GTDB representative genomes of 154 species, while 6,340 MAGs were not assigned to any GTDB species representative, data that would provide a large number of microbial resources for further research in the field of AMD bioremediation. *A. ferrooxidans*, *A. ferrivorans*, and *A. thiooxidan* have been demonstrated to be functional in AMD recovery[9,14,16]. In this study, *A. ferrooxidans* was found in copper mines, and *A. ferrivorans* and *A. thiooxidan* were found in polymetallic mines (Supplementary Table 2).

Constructing a phylogeny of nonredundant MAGs. dRep can reduce the computational time for pairwise genome comparisons by sequentially applying a fast, inaccurate estimation of genome distance and a slow, accurate measure of average nucleotide identity, thereby achieving a 28 fold increase in speed with perfect recall and precision compared to previously developed algorithms[53]. All of the produced 7,007 qualified bin sets were aggregated and de-replicated at 95% average nucleotide identity (ANI) using dRep (version 3.2.0, parameters: -comp 50 -con 5 -sa 0.95 –pa 0.9), resulting in a total of 1,992 species-level qualified MAGs[54]. These 1,992 de-replicated MAGs were further refined using a maximum-likelihood phylogeny inferred from a concatenation of 120 bacterial or 122 archaeal marker genes produced by GTDB-Tk[51]. Bacterial and archaeal approximate maximum likelihood trees were built using FastTree (version 2.1.10) with WAG + GAMMA models[47,55–57], and visualized by iTOL[58].

A striking feature of these trees is the large number of major lineages without assignment of a GTDB species representative (Fig. 3)[51]. There were 24 phyla in Bacteria without assignment of a GTDB species representative, and very limited MAGs were assigned to GTDB species representatives of Bacteria in the 16 phyla of Proteobacteria, Actinobacteriota, Nitrospirota, Firmicutes_E, Firmicutes, SZUA-79, Bacteroidota, Campylobacterota, Desulfobacterota, Spirochaetota, Firmicutes_B, Patescibacteria, Acidobacteriota, Aquificota, Bdellovibrionota, and Deinococcota (Fig. 3a). No MAGs were assigned to GTDB species representatives of Archaea in the phyla of Halobacteriota, Methanobacteriota, Thermoproteota, Asgardarchaeota, and Aenigmatarchaeota, and very limited MAGs were assigned to GTDB species representatives of Archaea in the phyla of Nanoarchaeota, Micrarchaeota, and Thermoplasmatota (Fig. 3b).

Mining of secondary metabolite biosynthetic gene clusters. Antibiotics & Secondary Metabolite Analysis Shell (antiSMASH, https://antismash.secondarymetabolites.org) is a tool that enables rapid identification, annotation, and analysis of smBGCs in genomes[59]. Since its first release in 2011, it has been the most widely used bioinformatics software for predicting smBGCs and the standard tool for smBGCs mining[60]. A total of 11,856 putative smBGCs were mined from 7,007 qualified MAGs across 13 mineral types using antiSMASH (version 5.1.2) called as follows: –cf-create-clusters –cb-general –cb-knownclusters –cb-subclusters –asf –pfam2go –smcog-trees –genefinding-tool prodigal, and in addition ignoring contigs with lengths shorter than 5 kb. antiSMASH analysis of 7,007 MAGs required 24 hours wall time using an HPC with multiple 2.60 GHz Intel (R) Xeon (R) Gold 6126 CPUs and 196 GB of RAM.

Using a threshold of 75% identity over 80% of the query length, 10,899 (91.93%) of 11,856 putative smBGCs were identified as putative novel smBGCs querying against the NCBI nucleotide sequence collection (downloaded 27 Jan 2021) by the command 'blastn' within the NCBI BLAST+ package (version 2.11)[61] with an E-value cutoff of 1×10^{-1}. Although many modular clusters were fragmented, we identified over 154 smBGC regions $>$50 kb in length and more than 1,834 $>$ 30 kb. These smBGCs were further classified into eight categories using BIG-SCAPE with default parameters[62]. Among these eight smBGC categories, terpene had the largest number and made up the highest percentage of smBGCs at 3,751 smBGCs and 31.64%, respectively (Table 3, Supplementary Table 3).

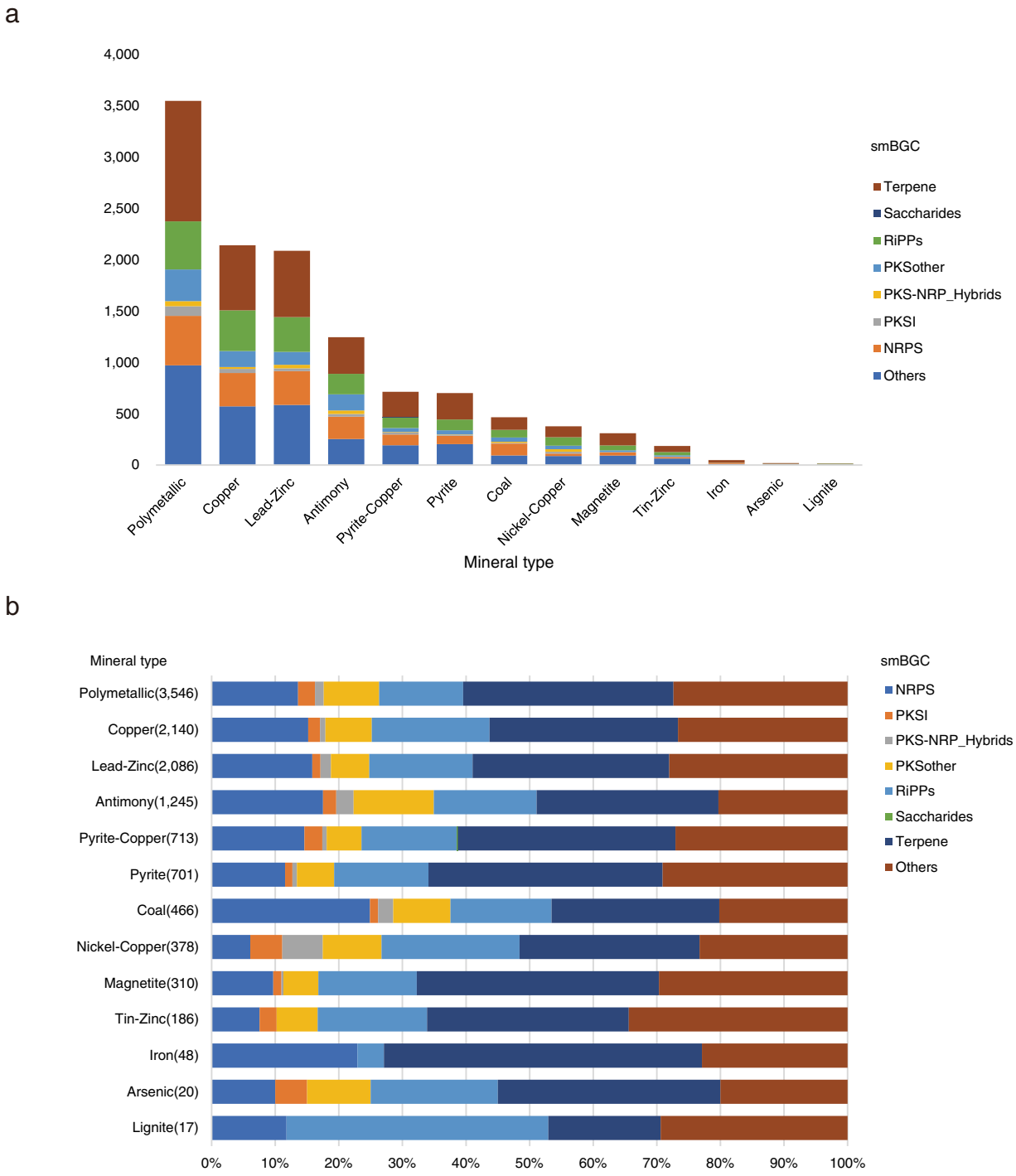

Fig. 4 Secondary metabolite biosynthetic gene cluster (smBGC) distributions in 13 types of minerals. (**a**) The number of smBGCs in different types of minerals. (**b**) Relative frequency of smBGC types across 13 types of minerals.

Data Records

The rawdata from the 111 private samples was deposited in NODE (https://www.biosino.org/node/project/detail/OEP001841)[63], GSA (CRA006735)[64], and NCBI SRA (PRJNA666025)[65]. A total of 7,007 MAGs with completeness ≥50%, contamination ≤5%, and had a quality score of >50 (the medium-quality level of the MIMAG standard) were obtained from 13 mineral types by metagenomic assembly and binning[47]. A total of 981 (14.00%) MAGs were assigned as high quality according to the MIMAG standard[24]. All 7,007 MAGs from the current study have been deposited in eLMSG (an eLibrary of Microbial Systematics and Genomics, https://www.biosino.org/elmsg/index) under accession numbers LMSG_G000004334.1–LMSG_G000011340.1[66], NODE (https://www.biosino.org/node/analysis/detail/OEZ008530)[67], and GenBank (PRJNA834572)[68].

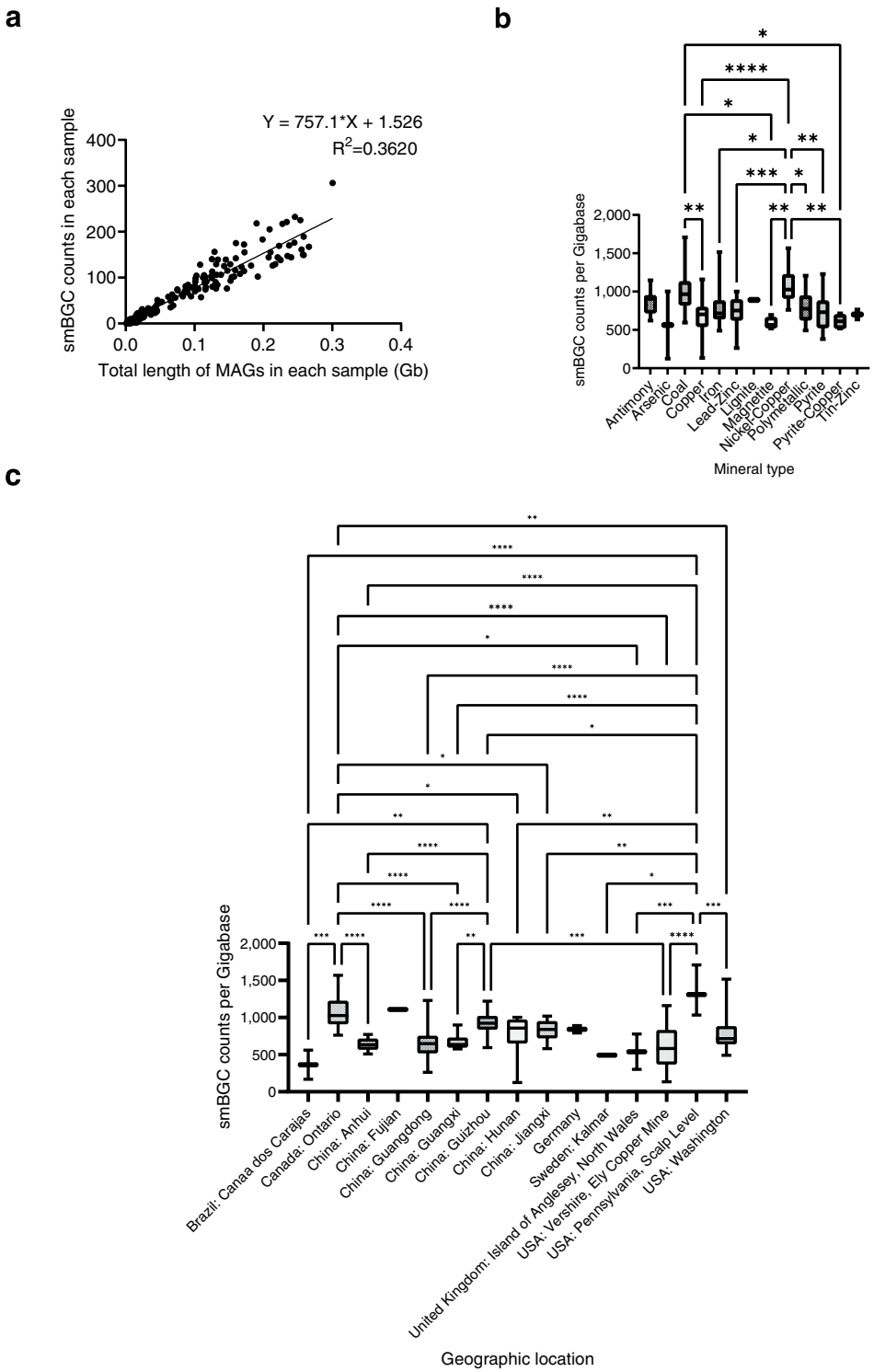

Fig. 5 The diversity of secondary metabolite biosynthetic gene clusters (smBGCs) in different mineral types and geographic locations. (**a**) Correlation between the total number of smBGCs in each sample and the total length of quality MAGs in each sample. (**b**) smBGC counts per Gigabase (the total number of smBGCs in each sample divided by the total length of quality MAGs in each sample) plotted according to mineral type. (**c**) smBGC counts per Gigabase (the total number of smBGCs in each sample divided by the total length of quality MAGs in each sample) plotted according to geographic location. Data were analyzed using one-way ANOVA followed by Turkey's test ($^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, and $^{****}P < 0.0001$).

All 11,856 putative smBGCs from 7,007 MAGs of 13 mineral types were deposited in NODE (https://www.biosino.org/node/analysis/detail/OEZ008529)[69] and GenBank (KFVK00000000)[70]. The classes of secondary metabolites synthesized by each smBGC across 13 mineral types were assigned (Fig. 4b). Non-ribosomal peptide synthetase (NRPS), post-translationally modified peptides (RiPPs), and terpene were found in all mineral

types. The 13 mineral types in this study had relatively low numbers of smBGCs in the remaining smBGC categories, including type I polyketide synthesase (PKS I), PKSother, and PKS-NRP_hybrids. Saccharides are only found in pyrite-copper mines.

Technical Validation

In order to ensure that the datasets from the SRA database only contained AMD metagenomic data, the metadata of these datasets from the SRA database and the scientific literature were manually curated. To select metagenomic datasets, only datasets for which the library strategy was WGS and the library source was METAGENOMIC were chosen. Because the pH values of AMD were usually 2–4[1], datasets such as SRS1650501-SRS1650503, SRS872561, SRS962537, SRS963313, SRS963552, SRS963574, SRS963594, SRS963611, and SRS963627, whose pH values were greater than 4, were removed to further filter the AMD metagenomic datasets. For datasets that did not provide pH values, metadata in the SRA database and in the scientific literature were reviewed to preserve only AMD metagenomic datasets[71–75].

The latitude and longitude of SRS1810936 was retrieved according to the geographic location of this sample. The mineral types of SRS5255199, SRS5255198, SRS5255197, and SRS2947527 were obtained through manual review of the metadata in the SRA database and scientific literature[76].

The smBGCs number and type varied even within the same dRep cluster (Supplementary Table 4). Therefore, we used the 7,007 MAGs before de-replication for the smBGCs prediction. A total of 6,026 from 7,007 MAG belonged to medium quality according to the MIMAG standard[24]. Using the draft genome for the smBGCs mining by using antiSMASH would cause the number of detected gene clusters to be artificially high, and some contigs with gene cluster fragments might be left undetected[77]. In order to obtain better smBGCs, we ignored contigs with lengths shorter than 5 kb to increase the chance of the smBGCs we mined to have roles in secondary metabolite synthesis[78]. Although many modular clusters were fragmented, we identified over 154 BGC regions >50 kb in length and more than 1,834 >30 kb.

We used linear regression to examine the sample size associated with the diversity of secondary metabolite biosynthetic gene clusters by GraphPad Prism (version 9.3.1). The total number of smBGCs in each sample showed a moderate positive correlation ($R^2 = 0.3620$) with the total length of quality MAGs in each sample (Fig. 5a), demonstrating that the number of smBGCs may also be affected by other factors.

The box plots of smBGC counts per Gigabase in different geographic locations or mineral types were generated using GraphPad Prism (version 9.3.1). One-way ANOVA followed by Turkey's test was used to analyze the differences among groups ($P < 0.05$) by GraphPad Prism (version 9.3.1). Notably, the smBGCs were most abundant in Canada: Ontario and USA: Pennsylvania, Scalp Level by the analysis of geographic location, while Coal mine and Nickel-Copper mine had relatively greater abundances of smBGCs according to the analysis of mineral type (Fig. 5b,c).

Usage Notes

The datasets analyzed in this study were the largest AMD metagenomic datasets considered to date. Among the 68 samples from the SRA database, only 11 (16%) of the samples were from AMD metagenomic datasets from China. Through the collection and sequencing of 111 AMD samples in this study, the metagenomic data of AMD in southeastern China were obtained. This complemented the publicly available datasets in order to provide a better overview of the putative novel microorganisms and secondary metabolite resources in the AMD environment. These datasets can be further employed in research on AMD bioremediation, the mining of novel secondary metabolites for drug synthesis, and for the analysis of gene functions, metabolic pathways, and CNPS cycles in AMD.

Code availability

The version and parameters of all of the bioinformatics tools used in this work are described in the Methods section.

Received: 19 May 2021; Accepted: 23 November 2022;
Published online: 09 December 2022

References

1. Nancucheo, I. *et al*. Recent Developments for Remediating Acidic Mine Waters Using Sulfidogenic Bacteria. *Biomed Res. Int.* **2017**, 7256582 (2017).
2. Grimalt, J. O., Ferrer, M. & Macpherson, E. The mine tailing accident in Aznalcollar. *Sci. Total Environ.* **242**, 3–11 (1999).
3. Glukhova, L. B. *et al*. Isolation, Characterization, and Metal Response of Novel, Acid-Tolerant Penicillium spp. from Extremely Metal-Rich Waters at a Mining Site in Transbaikal (Siberia, Russia). *Microb. Ecol.* **76**, 911–924 (2018).
4. Schmidt, U. Enhancing phytoextraction: the effect of chemical soil manipulation on mobility, plant accumulation, and leaching of heavy metals. *J. Environ. Qual.* **32**, 1939–1954 (2003).
5. Johnson, D. B. & Hallberg, K. B. The microbiology of acidic mine waters. *Res. Microbiol.* **154**, 466–473 (2003).
6. Denef, V. J., Mueller, R. S. & Banfield, J. F. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J.* **4**, 599–610 (2010).
7. Kuang, J. L. *et al*. Contemporary environmental variation determines microbial diversity patterns in acid mine drainage. *ISME J.* **7**, 1038–1050 (2013).
8. Fahy, A. *et al*. 16S rRNA and As-Related Functional Diversity: Contrasting Fingerprints in Arsenic-Rich Sediments from an Acid Mine Drainage. *Microb. Ecol.* **70**, 154–167 (2015).
9. Mendez-Garcia, C. *et al*. Microbial diversity and metabolic networks in acid mine drainage habitats. *Front. Microbiol.* **6**, 475 (2015).
10. Abinandan, S., Subashchandrabose, S. R., Venkateswarlu, K. & Megharaj, M. Microalgae-bacteria biofilms: a sustainable synergistic approach in remediation of acid mine drainage. *Appl. Microbiol. Biotechnol.* **102**, 1131–1144 (2018).

11. Qian, Z., Tianwei, H., Mackey, H. R., van Loosdrecht, M. C. M. & Guanghao, C. Recent advances in dissimilatory sulfate reduction: From metabolic study to application. *Water Res.* **150**, 162–181 (2019).
12. Williams, K. P. & Kelly, D. P. Proposal for a new class within the phylum Proteobacteria, Acidithiobacillia classis nov., with the type order Acidithiobacillales, and emended description of the class Gammaproteobacteria. *Int. J. Syst. Evol. Microbiol.* **63**, 2901–2906 (2013).
13. Hedrich, S. & Johnson, D. B. Acidithiobacillus ferridurans sp. nov., an acidophilic iron-, sulfur- and hydrogen-metabolizing chemolithotrophic gammaproteobacterium. *Int. J. Syst. Evol. Microbiol.* **63**, 4018–4025 (2013).
14. Hallberg, K. B., Gonzalez-Toril, E. & Johnson, D. B. Acidithiobacillus ferrivorans, sp. nov.; facultatively anaerobic, psychrotolerant iron-, and sulfur-oxidizing acidophiles isolated from metal mine-impacted environments. *Extremophiles* **14**, 9–19 (2010).
15. Chen, L. *et al*. Acidithiobacillus caldus sulfur oxidation model based on transcriptome analysis between the wild type and sulfur oxygenase reductase defective mutant. *PLoS One* **7**, e39470 (2012).
16. Gupta, A., Saha, A. & Sar, P. Thermoplasmata and Nitrososphaeria as dominant archaeal members in acid mine drainage sediment of Malanjkhand Copper Project, India. *Arch. Microbiol.* **203**, 1833–1841 (2021).
17. Yang, L. *et al*. Acidithiobacillus thiooxidans and its potential application. *Appl. Microbiol. Biotechnol.* **103**, 7819–7833 (2019).
18. Stierle, A. A. & Stierle, D. B. Bioactive secondary metabolites from acid mine waste extremophiles. *Nat. Prod. Commun.* **9**, 1037–1044 (2014).
19. Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nat. Rev. Microbiol.* **17**, 167–180 (2019).
20. Stierle, D. B., Stierle, A. A., Hobbs, J. D., Stokken, J. & Clardy, J. Berkeleydione and berkeleytrione, new bioactive metabolites from an acid mine organism. *Org. Lett.* **6**, 1049–1052 (2004).
21. Moutiez, M., Belin, P. & Gondry, M. Aminoacyl-tRNA-Utilizing Enzymes in Natural Product Biosynthesis. *Chem. Rev.* **117**, 5578–5618 (2017).
22. Gondry, M. *et al*. A Comprehensive Overview of the Cyclodipeptide Synthase Family Enriched with the Characterization of 32 New Enzymes. *Front. Microbiol.* **9**, 46 (2018).
23. Borthwick, A. D. 2,5-Diketopiperazines: synthesis, reactions, medicinal chemistry, and bioactive natural products. *Chem. Rev.* **112**, 3641–3716 (2012).
24. Bowers, R. M. *et al*. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat. Biotechnol.* **35**, 725–731 (2017).
25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
26. Forouzan, E., Shariati, P., Mousavi Maleki, M. S., Karkhane, A. A. & Yakhchali, B. Practical evaluation of 11 de novo assemblers in metagenome assembly. *J. Microbiol. Methods* **151**, 99–105 (2018).
27. Pasolli, E. *et al*. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell* **176**, 649–662 e620 (2019).
28. Li, D. *et al*. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**, 3–11 (2016).
29. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
30. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. metaSPAdes: a new versatile metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
31. Fritz, A. *et al*. CAMISIM: simulating metagenomes and microbial communities. *Microbiome* **7**, 17 (2019).
32. Sieber, C. M. K. *et al*. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat. Microbiol.* **3**, 836–843 (2018).
33. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP-a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).
34. Murovec, B., Deutsch, L. & Stres, B. Computational Framework for High-Quality Production and Large-Scale Evolutionary Analysis of Metagenome Assembled Genomes. *Mol. Biol. Evol.* **37**, 593–598 (2020).
35. Dong, X. *et al*. Thermogenic hydrocarbon biodegradation by diverse depth-stratified microbial populations at a Scotian Basin cold seep. *Nat. Commun.* **11**, 5825 (2020).
36. Xu, B. *et al*. A holistic genome dataset of bacteria, archaea and viruses of the Pearl River estuary. *Sci. Data* **9**, 49 (2022).
37. Zhou, L., Huang, S., Gong, J., Xu, P. & Huang, X. 500 metagenome-assembled microbial genomes from 30 subtropical estuaries in South China. *Sci. Data* **9**, 310 (2022).
38. Zhang, H. *et al*. Metagenome sequencing and 768 microbial genomes from cold seep in South China Sea. *Sci. Data* **9**, 480 (2022).
39. Lee, S. *et al*. Methane-derived carbon flows into host-virus networks at different trophic levels in soil. *Proc. Natl. Acad. Sci. U S A* **118**, e2105124118 (2021).
40. Bay, S. K. *et al*. Trace gas oxidizers are widespread and active members of soil microbial communities. *Nat. Microbiol.* **6**, 246–256 (2021).
41. Li, J. *et al*. Intracellular silicification by early-branching magnetotactic bacteria. *Sci. Adv.* **8**, eabn6045 (2022).
42. Yang, H. *et al*. ABO genotype alters the gut microbiota by regulating GalNAc levels in pigs. *Nature* **606**, 358–367 (2022).
43. von Schwartzenberg, R. J. *et al*. Caloric restriction disrupts the microbiota and colonization resistance. *Nature* **595**, 272–277 (2021).
44. Saheb Kashaf, S., Almeida, A., Segre, J. A. & Finn, R. D. Recovering prokaryotic genomes from host-associated, short-read shotgun metagenomic sequencing data. *Nat. Protoc.* **16**, 2520–2541 (2021).
45. Yang, C. *et al*. A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data. *Comput. Struct. Biotechnol. J.* **19**, 6301–6314 (2021).
46. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–1055 (2015).
47. Nayfach, S. *et al*. Publisher Correction: A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 520 (2021).
48. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res.* **49**, 9077–9096 (2021).
49. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
50. Kalvari, I. *et al*. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* **46**, D335–D342 (2018).
51. Chaumeil, P. A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2019).
52. Parks, D. H. *et al*. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
53. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
54. Nayfach, S. *et al*. Author Correction: A genomic catalog of Earth's microbiomes. *Nat. Biotechnol.* **39**, 521 (2021).
55. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
56. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).
57. Liu, K., Linder, C. R. & Warnow, T. RAxML and FastTree: comparing two methods for large-scale maximum likelihood phylogeny estimation. *PLoS One* **6**, e27731 (2011).

58. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**, W293–W296 (2021).
59. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
60. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
61. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
62. Navarro-Munoz, J. C. *et al.* A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **16**, 60–68 (2020).
63. data*NODE The National Omics Data Encyclopedia* https://www.biosino.org/node/project/detail/OEP001841 (2021).
64. data*GSA Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA006735 (2022).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP121625 (2022).
66. data*eLMSG an eLibrary of Microbial Systematics and Genomics* https://www.biosino.org/elmsg/amdDetail (2022).
67. data*NODE The National Omics Data Encyclopedia* https://www.biosino.org/node/analysis/detail/OEZ008530 (2022).
68. dataZhang, G. Q., Wang, L. & Liu, W. Mining of novel secondary metabolite biosynthetic gene clusters from acid mine drainage. *GenBank* https://identifiers.org/ncbi/insdc:KFVK01000000 (2022).
69. data*NODE The National Omics Data Encyclopedia* https://www.biosino.org/node/analysis/detail/OEZ008529 (2022).
70. dataZhang, G. Q., Wang, L. & Liu, W. Mining of novel secondary metabolite biosynthetic gene clusters from acid mine drainage. *GenBank* https://identifiers.org/ncbi/insdc:KFVK00000000 (2022).
71. Giddings, L. A. *et al.* Characterization of an acid rock drainage microbiome and transcriptome at the Ely Copper Mine Superfund site. *PLoS One* **15**, e0237599 (2020).
72. Chen, L. X. *et al.* Shifts in microbial community composition and function in the acidification of a lead/zinc mine tailings. *Environ. Microbiol.* **15**, 2431–2444 (2013).
73. Krause, S., Bremges, A., Munch, P. C., McHardy, A. C. & Gescher, J. Characterisation of a stable laboratory co-culture of acidophilic nanoorganisms. *Sci. Rep.* **7**, 3289 (2017).
74. Muhling, M. *et al.* Reconstruction of the Metabolic Potential of Acidophilic Sideroxydans Strains from the Metagenome of an Microaerophilic Enrichment Culture of Acidophilic Iron-Oxidizing Bacteria from a Pilot Plant for the Treatment of Acid Mine Drainage Reveals Metabolic Versatility and Adaptation to Life at Low pH. *Front. Microbiol.* **7**, 2082 (2016).
75. Arif, S., Nacke, H. & Hoppert, M. Metagenome-Assembled Genome Sequences of a Biofilm Derived from Marsberg Copper Mine. *Microbiol. Resour. Announc.* **10**, e01253–01220 (2021).
76. Liljeqvist, M. *et al.* Metagenomic analysis reveals adaptations to a cold-adapted lifestyle in a low-temperature acid mine drainage stream. *FEMS Microbiol. Ecol.* **91** (2015).
77. Blin, K. *et al.* antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. *Nucleic Acids Res.* **41**, W204–212 (2013).
78. Wei, B. *et al.* An atlas of bacterial secondary metabolite biosynthesis gene clusters. *Environ. Microbiol.* **23**, 6981–6992 (2021).

Acknowledgements

This work was supported by the Second Tibetan Plateau Scientific Expedition and Research Program (STEP) (2021QZKK0101), National Key R&D Program of China (2018YFA0900704), the International Partnership Program of Chinese Academy of Sciences (Grant NO. 153D31KYSB20170121), Training Program of the Major Research Plan of the National Natural Science Foundation of China (92051116), Biological Resources Programme of Chinese Academy of Sciences (Grant NO. KFJ-BRP-017-79), Biological Resources Service Network Initiative of Chinese Academy of Sciences (Grant NO. KFJ-BRP-009-001), and the Program on Platform and Talent Development of the Department of Science and Technology of Guizhou China ([2019] 5617). We appreciate the strong support from the joint program of Guian Center for Biomedical Big Data of SINH/SIBS of CAS. This publication was made possible by support from Wensheng Shu. We also would like to thank Liang Li for the development of the website for the data deposition, ShengnanYuan and Ruixin Zhu for providing the metagenomic analysis pipeline, guiding the analysis and editing the manuscript, Yufeng Zhang for being involved in the discussion of this study, and the integrated Microbiome Analysis Cloud platform (iMAC platform) for the analysis process and computing power support for this project. We thank the Editor and Reviewers for their constructive reviews that helped improve the original manuscript.

Author contributions

Ling Wang assembled and annotated the metagenome. Wan Liu executed the data analysis and wrote the manuscript. Jieliang Liang provided the newly collected and sequenced AMD datasets. Linna Zhao, Qiang Li, Chenfen Zhou, and Hui Cen were involved in the discussion of this study. Qingbei Weng designed this study. Guoqing Zhang conceived of the studies, provided material support, and edited the manuscript. All of the authors read, edited, and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01866-6.

Correspondence and requests for materials should be addressed to Q.W. or G.Z.