RESEARCH

MEBS, a software platform to evaluate large (meta)genomic collections according to their metabolic machinery: unraveling the sulfur cycle

Authors

Valerie De Anda[1*], Icoquih Zapata-Peñasco[2], Augusto Cesar Poot-Hernandez[3] Luis E. Eguiarte[1], Bruno Contreras-Moreira[4,5*] and Valeria Souza[1*]

Affiliations:

[1]*Departamento de Ecología Evolutiva, Instituto de Ecología, Universidad Nacional Autónoma de México, 70-275, Coyoacán 04510México D.F.*
[2]*Dirección de Investigación en Transformación de Hidrocarburos. Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas, Norte 152, Col. San Bartolo Atepehuacan, 07730, México*
[3]*Departamento de Ingeniería de Sistemas Computacionales y Automatización. Sección de Ingeniería de Sistemas Computacionales. Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas.*
[4]*Estación Experimental de Aula Dei, Consejo Superior de Investigaciones Científicas (EEAD-CSIC), Avda. Montañana, 1005, Zaragoza 50059, Spain*
[5]*Fundación ARAID, calle María de Luna 11, 50018 Zaragoza, Spain.*

Correspondence address: Valerie de Anda. Instituto de Ecología, Universidad Nacional Autónoma de México, 70-275, Coyoacán 04510; México D.F; Tel: (+52)5556229006; E-mail valdeanda@ciencias.unam.mx; Bruno Contreras Moreira. Estación Experimental de Aula Dei, Avda. Montañana, 1005, Zaragoza 50059; Spain; Tel:(+34) 976716089 ; E-mail bcontreras@eead.csic.es; Valeria Souza . Instituto de Ecología, Universidad Nacional Autónoma de México, 70-275, Coyoacán 04510; México D.F; Tel: (+52)5556229006; E-mail souza@unam.mx

BACKGROUND: The increasing number of metagenomic and genomic sequences has dramatically improved our understanding of microbial diversity, yet our ability to infer metabolic capabilities in such datasets remains challenging. **FINDINGS:** We describe the Multigenomic Entropy Based Score pipeline (MEBS), a software platform designed to evaluate, compare and infer complex metabolic pathways in large 'omic' datasets, including entire biogeochemical cycles. MEBS is open source and available through https://github.com/eead-csic-compbio/metagenome_Pfam_score. To demonstrate its use we modeled the sulfur cycle by exhaustively curating the molecular and ecological elements involved (compounds, genes, metabolic pathways and microbial taxa). This information was reduced to a collection of 112 characteristic Pfam protein domains and a list of complete-sequenced sulfur genomes. Using the mathematical framework of relative entropy (*H'*), we quantitatively measured the enrichment of these domains among sulfur genomes. The entropy of each domain was used to both: build up a final score that indicates whether a (meta)genomic sample contains the metabolic machinery of interest and to propose marker domains in metagenomic sequences such as DsrC (PF04358). MEBS was benchmarked with a dataset of 2,107 non-redundant microbial genomes from RefSeq and 935 metagenomes from MG-RAST. Its performance, reproducibility, and robustness were evaluated using several approaches, including random sampling, linear regression models, Receiver Operator Characteristic plots and the Area Under the Curve metric (AUC). Our results support the broad applicability of this algorithm to accurately classify (AUC=0.985) hard to culture genomes (e.g., *Candidatus Desulforudis audaxviator*), previously characterized ones and

metagenomic environments such as hydrothermal vents, or deep-sea sediment. **CONCLUSIONS:** Our benchmark indicates that an entropy-based score can capture the metabolic machinery of interest and be used to efficiently classify large genomic and metagenomic datasets, including uncultivated/unexplored taxa.

Keywords:

Metabolic machinery, metagenomics, omic-datasets, Pfam domains, Relative entropy, sulfur cycle, Multigenomic Entropy-based Score.

Background

Over the last 15 years, the enormous advances in high-throughput sequencing technologies have revolutionized the field of microbial ecology, dramatically improving our understanding of life's microbial diversity to an unprecedented level of detail [1–4].

Nowadays, accessing the total repertoire of genomes within complex communities by means of metagenomics is becoming a standard and routine procedure in order to attain the full insight of the diversity, ecology, evolution and functional makeup of the microbial world [5]. Furthermore, the accurate reconstruction of microbial genomes and draft-populations from environmental metagenomic studies has been shown to be a powerful approach [6–10], providing clues about the potential metabolic strategies of hard-to-culture microbial lineages by linking the functional mechanisms that support specific metabolisms with taxonomic, systematic, and ecological contexts of that lineage [8].

Despite the accelerated accumulation of large collections of metagenomic and genomic sequences, our ability to analyze, evaluate and compare complex metabolic capabilities in large-scale 'omic' datasets remains biologically and computationally challenging [11]. Predicting the metabolic potential is a key step in describing the relationship between a microbial community and its ecosystem function. This is largely performed by mapping the protein coding genes of 'omic' data onto reference pathway databases such as MetaCyc [12] or KEGG [13] based on their homology to previously characterized genes [14]. The current available methods for metabolic pathway prediction or reconstruction rely on the use of several metrics to infer the overall repertoire of metabolic pathways present in a given metagenomic dataset (e.g., MinPath [14], HUMAnN[15], PRMT [16], MetaPathways [17]).

However, due to the challenges involved in testing meaningful biological hypotheses with complex data, only a small proportion of the metabolic information derived from these datasets is eventually used to draw ecologically relevant conclusions. In this regard, most of the microbial ecology-derived 'omic' studies have been mainly focused on either: i) developing broad description of the metabolic pathways within a certain environment e.g., [18,19]; ii) analyzing the relative abundance of marker genes involved in several metabolic processes and in certain ecosystems (e.g., primary productivity, decomposition, biogeochemical cycling [20–24]; or iii) discovering differentially abundant, shared or unique functional units (genes, proteins or metabolic pathways) across several environmental metagenomic samples [25–27].

Therefore, in order to address some of the limitations of these methods, we propose a novel approach to reduce the complexity of targeted metabolic pathways involved in several integral ecosystem processes -- such as entire biogeochemical cycles -- into a single informative score, called Multigenomic Entropy-Based Score (MEBS). This approach is based on the mathematical rationalization of Kullback-Leibler divergence, also known as relative entropy *H'* [28]. Relative entropy has been widely applied in physics, communication theory, and statistical inference, and it is interpreted as a measure of disorder, information and uncertainty, respectively [29]. Here we use the communication theory concept of *H'* to summarize the information derived from the metabolic machinery encoded by the protein coding genes of 'omic' datasets. The application of this metric in biology was originally developed by Stormo and colleagues identifying binding sites that regulate gene transcription sites [30].

In order to evaluate the performance of our approach, we selected the sulfur cycle (from now on S-cycle) because this is one of the most metabolically- and ecologically complex biogeochemical cycles, but there are few studies analyzing the complete repertoire (genes, proteins, or metabolic pathways) involved in the mobilization of inorganic-organic sulfur compounds through microbial-catalyzed reactions at a planetary scale [20,31–35].

MEBS description

MEBS (Multigenomic Entropy-Based Score, RRID: 015708) runs in Linux systems and is available at [36]. For practical purposes, the MEBS algorithm was divided into four stages summarized in Figure 1 and explained below.

STAGE 1: Manual curation of Sulfur cycle and 'omic' datasets

Sulfur taxonomic representatives. A data set comprehensively covering the currently known representatives of the S-cycle was obtained from primary literature and the MetaCyc database [12]. Each taxonomic representative (at genus or species level) was selected under the criteria of having evidence suggesting their physiological and biochemical involvement in the degradation, reduction, oxidation, or disproportionation of sulfur compounds. Then, each taxonomic representative was scanned against our Genomic dataset (see further details below), in order to obtain a list containing the completely sequenced and non-redundant genomes of the S-cycle. The resulting Sulfur list (or 'Suli') currently contains 161-curated genomes, and was used as the first input of the pipeline. Both the manually curated taxonomic representatives and Suli can be found in Table S1.

Random taxonomic representatives (RList). As a negative control, we generated 1000 lists of genomes that are not particularly enriched on sulfur metabolic preferences. Each list contains 161 random genomes, the same number of microorganisms included in Suli. These lists were obtained by randomly subtracting from the Genomic dataset (see below) 161 Refseq accession numbers and their corresponding names.

Metabolic pathways and genes. We gathered and classified the metabolic pathways involved in the S-cycle from the primary literature and two experimentally validated curated databases: KEGG (KEGG, RRID:SCR_012773) [13] and MetaCyc (MetaCyc , RRID:SCR_007778) [12] . All the molecular information was then combined into a single database named Sucy (for Sulfur cycle). Sucy currently contains 152 genes and 48 enzyme classification numbers annotated in the Enzyme classification

[37] (Table S2). The 152 FASTA sequences of the proteins encoded by these genes were downloaded from UniProt [38] and used as the second input of the pipeline.

Genomic dataset (Gen). At the time of the analysis (December 21, 2016), a total of 4,158 genomes were available from RefSeq database [39]. For comparative genomic purposes, we removed redundancy in this large data set by using the Web interface [40] described in [41]. As phylogenomic distance measure, we used a modified version of the Genomic Similarity Score defined as GSSb in [41]; we selected the most tolerant threshold of 0.95 (so as not to drop many sequenced genomes) and default parameters, resulting in 2,107 clusters containing similar genomes, ordered by size (largest to smallest). Then, the largest genome representative for each group was searched in the NCBI genome assembly summary file [42] and downloaded from the NCBI FTP site [43].

Metagenomic dataset (Met). We used the Meta Genome Rapid Annotation using Sub- system Technology server (MG-RAST, RRID:SCR_004814) [44] to download metagenomes that: i) were publicly available; ii) contained associated metadata; and iii) had been isolated from well-defined environments (i.e., rivers, soil, biofilms), discarding host associated microbiome sequences (i.e., human, cow, chicken). In addition we also included 35 unpublished metagenomes derived from sediment, water and microbial mats from Cuatro Ciénegas, Coahuila (CCC), Mexico. The latter were also submitted and annotated in the MG-RAST server, and will be described in depth elsewhere. The resulting collection of 935 FASTA files (≈ 500 GB), containing gene-called protein sequences (MG-RAST stage 350), were downloaded from the RESTful MG-RAST API (http://api.metagenomics.anl.gov/api.html). While these metagenomes were evaluated and scored in STAGE 4, they were also analyzed to estimate their mean sequence length, considering that the fragmented nature of metagenomic sequences would have an impact on homology detection, depending on the length of the reads [45,46]. Therefore, we measured the Mean Size Length (MSL) of the peptide sequences of the 935 metagenomes in Met and the 152-curated proteins in Sucy, which are summarized in Figure S1. It was observed that the MSL of Met varies broadly, with a majority of metagenomic peptides with MSL ≤ 30 aa, and that Sucy proteins range from 49 to 1,020

aa, with MSL=349 aa. According to this distribution, the metagenomes in Met were grouped into seven well-defined categories: MSL≤30, ≤60, ≤100, ≤150, ≤200, ≤250, ≤300 aa.

Fragmented genomic dataset (GenF). In order to simulate the observed variability of MSL across metagenomes, protein sequences encoded in the genomic dataset (Gen, containing 2,107 genomes) were *in silico* sheared with Perl script *get_protein_fragments.pl* into the seven MSL categories defined above (30 to 300). This produced the GenF dataset, which currently requires up to 104GB of disk space.

STAGE 2: Domain composition of the input proteins

The annotation of protein domains in Sucy was conducted using Interproscan 5.21-60.0 [47] against databases Pfam-A v30 (Pfam, RRID:SCR_004726) [48], TIGRFAM v13 (JCVI TIGRFAMS , RRID:SCR_005493) [49] and Superfamily v1.75 (SUPERFAMILY , RRID:SCR_007952) [50]. Then, the Hidden Markov Models (HMMs) from matched Pfam domains (n=112) were extracted from Pfam-A using script *extract_hmms.pl*. These selected HMMs were subsequently scanned against the Genomic, Genomic Fragmented and Metagenomic datasets (from now on 'omic' datasets, see subsequent stages) using HMMER 3.0 *hmmsearch* --cut_ga option [51].

STAGE 3: Relative entropy and its use in detecting informative domains

In order to detect protein domains enriched among sulfur-based microorganisms (Suli), we used a derivative of the Kullback-Leibler divergence [28] — also known as relative entropy *H'(i)* — to measure the difference between probabilities *P* and *Q* (see Eq. 1 below). In this context, *P(i)* represents the frequency of protein domain *i* in the 161 Suli genomes (observed frequency), while *Q(i)* represents its frequency in the 2,107 genomes in Gen (expected frequency). The script to compute the entropy (*entropy.pl*) requires the list of the genomes of interest (Suli) and the tabular output file obtained in from the scanning of Gen and GenF against Pfam-Sucy database. The obtained values of *H'* (in bits) capture to what extent a given Pfam domain informs about the metabolism of interest. In this case, domains with *H'* values close or greater than one, correspond to the most informative Pfam domains (enriched among S-based genomes), whereas low *H'* values

(close to zero) indicate non-informative ones. Negative values correspond to those observed less than expected.

$$H' = P(i) log_2 \frac{P(i)}{Q(i)} \qquad \textbf{Eq. 1}$$

As a negative control, the *H'* of the 112 Pfam domains were recalculated in both Gen and GenF datasets, but replacing Suli with 1,000 equally sized lists of random-sampled genomes (Rlist).

We evaluated the impact of the MSL in the computed entropy values using Gen and GenF. First, we focused on detecting informative Pfam domains that could be used as possible molecular marker genes in variable length, metagenomic sequences. Specifically, we looked for domains displaying stable *H'* values across both Gen and GenF by using the script *plot_cluster_comparison.py*, which implements the following methods: K-Means, Affinity propagation, Mean-shift Spectral, Ward hierarchical, Agglomerative, DBSCAN and Birch. All of these are part of the scikit-learn Machine Learning Python module [52].

STAGE 4: Final score, interpretation, properties and benchmark

Peptide sequences from a given genome or metagenome of interest are evaluated by first scanning their Pfam domains and then producing a final score, defined as the sum of the precomputed entropies of matched S-related Pfam domains (see Equation 2). This score (Sulfur Score '*SS*' in our case) summarizes the information content of the metabolic machinery of interest. In this context, informative sulfur protein domains would contribute to higher *SS*, whereas non-informative ones would decrease it. This is an extension of procedures originally developed for the alignment of DNA and protein motifs, in which individual positions are independent and additive, and can be simply summed up to obtain the total weight or information content [30]. Instead of aligning sequences, in our context we added up the entropy values of the Pfam domains matched in a given 'omic' sample (resulting from scanning the sample of interest against Pfam-Sucy), from which a total weight (*SS*) is computed by using script *pfam_score.pl.*

$$SS = \sum_{i=1}^{112} H' \qquad \textbf{Eq. 2}$$

Datasets in which the majority of informative S-cycle protein domains are represented will yield a high *SS*; in contrast, low *SS* values should be expected if proteins involved in the S-cycle are not particularly enriched.

MSL. As the calculation of the *SS* depends on the MSL of the omic sample of interest, script *pfam_score.pl* supports option –size, in amino acid residues (aa). In this way, appropriate precomputed *H'* values for Pfam domains can be selected to produce the final score. Currently 30, 60, 100, 150, 200, 250, 300 and real sizes are supported.

Metabolic pathway completeness and KEGG visualization. The presence-absence patterns of Pfam domains belonging to particular pathways can be exploited to compute metabolic completeness. This optional task is invoked with parameter –keggmap and a TAB-separated file mapping Pfam identifiers to KEGG Orthology entries (KO numbers) and the corresponding pathway in Sucy (see Table S3). To compute completeness, the total number of domains involved in a given pathway (i.e., sulfate reduction, sulfide oxidation) must be retrieved from the Sucy database (See Table S2). Then, the protein domains currently present in any given sample are divided by the total number of domains in the pre-defined pathway. The script produces: i) a detailed report of the metabolic pathways of interest; and ii) a list of KO numbers with Hex color codes, corresponding to KO matches in the omic sample, which can be exported to the KEGG Mapper – Search & Color Pathway tool [53] (see Figure S2).

Properties and performance of SS. Since the outcome of the final score (*SS*) largely depends on the list of microorganisms involved in the metabolism of interest (in our case Suli) and the Pfam domains found in the input protein sequences (n=112), we evaluated its robustness and reproducibility with several approaches. First, we compared our results with a benchmark performed three years ago in which we used Pfam-A v27 (instead of version 30), a genomic dataset

containing 1,528 non-redundant genomes (579 less genomes than our current Genomic dataset), and an input list of 156 genomes of interest (five less that our current Suli). Second, *SS* estimates were compared with scores obtained by randomly selecting ≈50% of the 112 Pfam domains with both Gen and Met. This analysis was performed a thousand times with *pfam_score.pl* –random. Third, we benchmarked the predictive capacity of the *SS* in order to accurately classify genomes of S-related organisms (Suli, n=161, positive instances), in contrast with a larger set of non-redundant genomes (Gen - Suli, n=1.946, negative instances). Therefore, we computed the True Positive Rates (TPR), False Positive Rates (FPR), Receiver Operating Characteristic (ROC) plots and the resulting Area Under the Curve (AUC) using the scikit-learn module described in [52].

Results and discussion

We present MEBS a new open source software to evaluate, quantify, compare, and predict the metabolic machinery of interest in large 'omic' datasets. The pipeline includes four stages. The first one consists on the systematic and targeted acquisition of the molecular and ecological information describing the metabolism of interest, represented by a list of curated microorganisms and a FASTA file of proteins involved in that metabolic network. In the second stage, the domain composition of the curated proteins is evaluated. Then, the domains enriched among the microorganisms of interest are identified by using the mathematical framework of the relative entropy (*H'*, third stage). Finally, the summation of the entropy of individual Pfam domains in a given genome or metagenomic dataset yields the final score (see Figure 1).

To test the applicability of this approach, we evaluated the metabolic machinery of the S-cycle. Due to its multiple redox states and its consequences on microbiological and geochemical transformations, S-metabolism can be observed as a complex metabolic machinery, involving a myriad of genes, enzymes, organic substrates and electron carriers, which largely depend on the surrounding geochemical and ecological conditions. For these reasons, the complete repertory involved in the metabolic machinery of S-cycle has remained underexplored despite the massive data produced in 'omic' experiments. Here, we performed an integral curation effort to describe all the elements involved in the S-cycle and then used, as explained in the following sections, to score genomic and metagenomic datasets in terms of their Sulfur relevance.

Manual curation: the complex metabolic machinery of the Sulfur cycle

In order to integrate the complete biogeochemical S-cycle, we manually curated and modeled the major processes involved in the mobilization and use of S-compounds through Earth biosphere. This effort resulted in two comprehensive databases. The first one includes most of the known microorganisms (with and without complete genomes) described in the literature to be closely involved in the S-cycle (Table S1). In this database, we included representative taxa from the following metabolic sulfur guilds: i) chemolithotrophic, colorless sulfur bacteria (CLSB: 24 genera); ii) anaerobic phototrophs, purple sulfur bacteria (PSB:25 genera), and green sulfur bacteria (GSB:9 genera); iii) sulfate reducing bacteria (SRB: 40 genera); and iv) deep-branch sulfur hyperthermophilic microorganisms, such as elemental sulfur reducing (SRM:19 genera) and oxidizers (SO:4 genera). From all the microorganisms described to be involved in the S-cycle, at the time of the analysis, a total of 161 were found to be completely sequenced and non-redundant genomes, and were used as the first input of the pipeline (Suli).

The second database (Sucy) contains genes, proteins, and pathways with experimental evidence linking them to the S-cycle. To compile this database, we first gathered the most important S-compounds derived from biogeochemical processes and biological catalyzed reactions. Then we classified each S-compound according to their chemical and thermodynamic nature (Gibbs free energy of formation, GFEF). Finally, we classified weather each compound can be used as a source of carbon, nitrogen, energy or electron donor, fermentative substrate, or terminal electron acceptor in respiratory microbial processes. The schematic representation of the manual curated effort summarizing the complexity of the sulfur biogeochemical cycle in a global scale is shown in Figure 2.

Once we selected the microorganisms, genes, and biogeochemical processes involved, we systematically divided the metabolic machinery of the S-cycle into 28 major metabolic pathways described in Table 1. In general terms we included pathways involved in: i) the oxidation/reduction of inorganic S-compounds, used as source of energy, electron donor or acceptor (P1-P7, P11 and P20 and P21); ii) the degradation of organic S-compounds, such as aliphatic sulfonates, sulfur amino acids, and organosulfonates (P8-P10, P12-P19, P22,P23,P27); iii) the methanogenesis from

304 methylated thiols, such as dimethyl sulfide DMS (P24), metylthio-propanoate (P25) and
305 methanethiol(P26), which are generated in nature by different biogeochemical processes [12]; and
306 finally, iv) the biosynthesis of sulfolipids (SQDG) (P28), because it has been observed that some
307 bacteria living in S-rich and P-lacking environments are able to synthetize sulfolipids, instead of
308 phospholipids, in the membrane as an adaptation to the selective pressures of these particular
309 environments [54].The synthetic pathway P29 is explained in further detail in the next sections
310 (Table 1).
311 After the comprehensive metabolic inventory was compiled, we linked all the elements in a single
312 network representation of the S-metabolic machinery (Figure 3). To the best of our knowledge, this
313 is the first molecular reconstruction of the cycle that considers all the sulfur compounds, genes,
314 proteins and the corresponding enzymatic steps resulting into higher order molecular pathways.
315 The latter representation also highlights the interconnection of pathways in terms of energy flow
316 and the interplay of the redox gradient (organic/inorganic) of the intermediate compounds that act
317 as key axes of organic and inorganic reactions (e.g., sulfite).

Table 1. Metabolic pathways of global biogeochemical S-cycle

Pathway number	Metabolism[a]	Chemical process[b]	Sulfur compound	Type[c]	Chemical formula	Source[d]	Number of Pfam domais[e]
P1	DS	O	Sulfite	I	SO32-	E	9
P2	DS	O	Thiosulfate	I	$S_2O_3^{2-}$	E	10
P3	DS	O	Tetrathionate	I	$S_4O_6^{2-}$	E	2
P4	DS	R	Tetrathionate	I	$S_4O_6^{2-}$	E	17
P5	DS	R	Sulfate	I	SO42-	E	20
P6	DS	R	Elemental sulfur	I	Sº	E	20
P7	DS	D	Thiosulfate	I	$S_2O_3^{2-}$	E	9
P8	DS	O	Carbon disulfide	O	CS_2	E	1
P9	A	DE	Alkanesulfonate	O	CH_3O_3SR	S	5
P10	A	R	Sulfate	I	SO_4^{2-}	S	20
P11	DS	O	Sulfide	I	H_2S	E/S	29
P12	A	DE	L-cysteate	O	$C_3H_6NO_5S$	C/E	1
P13	A	DE	Dimethyl sulfone	O	$C_2H_6O_2S$	C/E	3
P14	A	DE	Sulfoacetate	O	$C_2H_2O_5S$	C/E	2
P15	A	DE	Sulfolactate	O	$C_3H_4O_6S$	C/S	14
P16	A	DE	Dimethyl sulfide	O	C_2H_6S	C/S	16
P17	A	DE	Dimethylsulfoniopropionate	O	$C_5H_{10}O_2S$	C/S/E	12
P18	A	DE	Methylthiopropanoate	O	$C_4H_7O_2S$	C/S	7
P19	A	DE	Sulfoacetaldehyde	O	$C_2H_3O_4S$	C/S	7
P20	DS	O	Elemental sulfur	I	S°	C/S/E	7
P21	DS	D	Elemental sulfur	I	S°	C/S/E	1
P22	A	DE	Methanesulfonate	O	CH_3O_3S	C/S/E	7
P23	A	DE	Taurine	O	$C_2H_7NO_3S$	C/S/E	11
P24	DS	M	Dimethyl sulfide	O	C_2H_6S	C	1
P25	DS	M	Metylthio-propanoate	O	$C_4H_7O_2S$	C	1
P26	DS	M	Methanethiol	O	CH_4S	C	1
P27	A	DE	Homotaurine	O	$C_3H_9NO_3S$	N	1
P28	A	B	Sulfolipid	O	SQDG		4
P29			Markers		Markers		12

[a] Metabolism: Assimilative (A) inorganic compounds are reduced during biosynthesis; Dissimilative (DS) inorganic compounds used as electron acceptors in energy metabolism. A large amount of electron acceptor is reduced and the reduced product is secreted.

[b] Chemical Process: Oxidation (O): Reduction (R), Degradation (DE), Biosynthesis (B), Methanogenesis (M), Disproportionation (D).

[c] Compound Type: Organic (O): sulfur atoms with covalent bonds to carbon atoms. Inorganic (I): sulfur compounds with non-carbon atoms.

[d] Source: sulfur compound used as source of energy (E), sulfur (S), carbon (C), nitrogen (N).

[e] Number of Pfam domains belonging to each metabolic pathway described in Sucy (Table S2)

Annotation of Pfam domains within Sulfur proteins

Our approach requires the detection of structural and evolutionary units, also known as domains, in the curated list of protein sequences involved in the metabolism of interest (S-cycle in this case). The annotation of protein domains against the Pfam-A database resulted in a total of 112 domains identified in 147 proteins (out of 152). These 112 domains constitute the Pfam-Sucy database and represent all the pathways listed in Table 1. Two other protein family databases were tested (TGRFAM and Superfamily), but the number of proteins with positive matches was lower than with Pfam (57 and 137, respectively) and thus were not further considered.

Preparation of omic datasets: Gen, GenF and Met

The genomic dataset required for computing domain entropies (Gen) was obtained from public databases, as explained above in MEBS Description. A fragmented version of Gen, called GenF, was generated by considering the Mean Size Length (MSL) distribution of metagenomic sequences (Figure S1).

In order to benchmark MEBS with real environmental metagenomic samples, a collection of 900 public metagenomes was obtained from MG-RAST, to which we added 35 metagenomes sampled from an ultra-oligotrophic shallow lake in México (CCC). Altogether, these 935 metagenomes set up the Met dataset.

Using the relative entropy to recognize S-cycle domains and candidate markers

The next stage consists on the quantitative detection of informative domains (enriched among organism in Suli), by computing its relative entropy (*H'*) using Equation 1. The occurrences of each of the 112 Pfam domains in Suli and the genomic datasets were taken as observed and expected frequencies, respectively. Figure 4A summarizes the computed *H'* values in real (Gen) and fragmented genomic sequences of increasing size (GenF). The results indicate that only a few Pfam domains are equally informative regardless of the length of sequences. When *H'* values inferred from real, full-length proteins are compared to those of fragmented sequences, it can be seen that shorter sequences (MSL 30 & 60 aa) yield larger entropy differences than sequences of length > 100 aa (see in Figure 4B). Therefore, in order to shortlist candidate marker genes we selected those

Pfam domains displaying constant, high mean *H'* values in Gen and GenF, low *H'* standard deviation (std) and a clear separation from the random distribution.

We tested several clustering methods, summarized in Figure S3, with Ward and Birch performing best in grouping together informative protein domains with low std. However, the Ward classification was eventually selected as Birch failed to include a few Pfam domains relevant in the S-cycle (see Figure S4). By using Ward method, three well-defined clusters of Pfam domains were generated, as observed in Figure 4C. Cluster 0 included 94 domains containing H' values ranging from [-0.4, 0.4] and overlapping with the values obtained in the negative control explained in the next section. Cluster 1 consistently grouped together 12 Pfam domains listed in Table 2 with high entropy and low std, and can therefore be proposed as molecular markers in metagenomic sequences of variable length. Among the proposed marker domains are APS-Reductase (PF12139: H'=1.2), ATP-sulfurilase (PF01747: H'=1.03) and DsrC (PF04358: H'=0.52), key protein families in metabolic pathways involved in both sulfur oxidation/reduction processes. Finally, cluster 2 includes Pfam domains displaying high entropy values and high std, such as the PUA-like domain (PF14306: H'=1). We presume that domains within this cluster are also key players in S-metabolism; however, their high std makes them unsuitable for markers, particularly with metagenomic sequences of variable MSL. We suggest that further analyses will be required to test the implication in S-energy conservation processes of proteins containing domains such as PF03916, PF02665 or PF14697 (see complete list in Table S4).

Table 2 Informative Pfam domains with high *H'* and low std. Novel proposed molecular marker domains in metagenomic data of variable MSL

Pfam ID (Suli ocurrences)	*H'* mean	*H'* std	Description
PF12139 58/161	1.2	0.01	**Adenosine-5'-phosphosulfate reductase beta subunit:** Key protein domain for both sulfur oxidation/reduction metabolic pathways. Has been widely studied in the dissimilatory sulfate reduction metabolism. In all recognized sulfate-reducing prokaryotes, the dissimilatory process is mediated by three key enzymes: Sat, Apr and Dsr. Homologous proteins are also present in the anoxygenic photolithotrophic and chemolithotrophic sulfur-oxidizing bacteria (CLSB, PSB, GSB), in different cluster organization [35].
PF00374 135/161	1.1	0.09	**Nickel-dependent hydrogenase:** Hydrogenases with S-cluster and selenium containing Cys-x-x-Cys motifs involved in the binding of nickel. Among the homologues of this hydrogenase domain, is the alpha subunit of the sulfhydrogenase I complex of *Pyrococcus furiosus*, that catalyzes the reduction

			of polysulfide to hydrogen sulfide with NADPH as the electron donor [55].
PF01747 103/161	1.03	0.06	**ATP-sulfurylase:** Key protein domain for both sulfur oxidation and reduction processes. The enzyme catalyzes the transfer of the adenylyl group from ATP to inorganic sulfate, producing adenosine 5-phosphosulfate (APS) and pyrophosphate, or the reverse reaction [56].
PF02662 62/161	0.82	0.03	**Methyl-viologen-reducing hydrogenase, delta subunit:** Is one of the enzymes involved in methanogenesis and encoded in the mth-flp-mvh-mrt cluster of methane genes in *Methanothermobacter thermautotrophicus*. No specific functions have been assigned to the delta subunit [48].
PF10418 122/161	0.78	0.06	**Iron-sulfur cluster binding domain of dihydroorotate dehydrogenase B:** Among the homologous genes in this family are *asrA* and *asrB* from *Salmonella enterica enterica serovar Typhimurium*, which encode 1) a dissimilatory sulfite reductase, 2) a gamma subunit of the sulfhydrogenase I complex of *Pyrococcus furiosus* and, 3) a gamma subunit of the sulfhydrogenase II complex of the same organism [12].
PF13247 149/161	0.66	0.06	**4Fe-4S dicluster domain:** Homologues of this family include: 1) DsrO, a ferredoxin-like protein, related to the electron transfer subunits of respiratory enzymes, 2) dimethylsulfide dehydrogenase β subunit (ddhB), involved in dimethyl sulfide degradation in *Rhodovulum sulfidophilum* and 3) sulfur reductase FeS subunit (sreB) of *Acidianus ambivalens*, involved in the sulfur reduction using H_2 or organic substrates as electron donors [12].
PF04358 73/161	0.52	0	**DsrC like protein:** DsrC is present in all organisms encoding a dsrAB sulfite reductase (sulfate/sulfite reducers or sulfur oxidizers). The physiological studies suggest that sulfate reduction rates are determined by cellular levels of this protein. The dissimilatory sulfate reduction couples the four-electron reduction of the DsrC trisulfide to energy conservation [57]. DsrC was initially described as a subunit of DsrAB, forming a tight complex; however, it is not a subunit, but rather a protein with which DsrAB interacts. DsrC is involved in sulfur-transfer reactions; there is a disulfide bond between the two DsrC cysteines as a redox-active center in the sulfite reduction pathway. Moreover, DsrC is among the most highly expressed sulfur energy metabolism genes in isolated organisms and meta- transcriptomes (Santos et al., 2015).
PF01058 158/161	0.45	0.01	**NADH ubiquinone oxidoreductase, 20 Kd subunit:** Homologous genes are found in the delta subunits of both sulfhydrogenase complexes of *Pyrococcus furiosus* [12].
PF01568 156/161	0.4	0.05	**Molydopterin dinucleotide binding domain:** This domain corresponds to the C-terminal domain IV in dimethyl sulfoxide (DMSO) reductase [48].
PF09242 39/161	0.38	0.04	**Flavocytochrome c sulphide dehydrogenase, flavin-binding:** Enzymes found in S-oxidizing bacteria such as the purple phototrophic bacteria *Chromatium vinosum* [48].
PF04879 151/161	0.37	0.05	**Molybdopterin oxidoreductase Fe4S4 domain:** Is found in a number of reductase/dehydrogenase families, which include the periplasmic nitrate reductase precursor and the formate dehydrogenase alpha chain, *i.e., Wolinella succinogenes* polysulfide reductase chain. *Salmonella typhimurium* thiosulfate reductase (gene phsA).
PF08770 45/161	0.35	0.03	**Sulphur oxidation protein SoxZ:** SoxZ sulfur compound chelating protein, part of the complex known as the Sox enzyme system (for sulfur oxidation) that is able to oxidize thiosulfate to sulfate with no intermediates in *Paracoccus parantropus* [12] .

381

382

Is the entropy affected by the input list of microorganisms? Negative control test

In order to evaluate to what extent the *H'* values depend on the curated list of microorganisms, we performed a negative control by replacing Suli in 1,000 lists of randomly-sampled genomes and used them to compute the observed frequencies (see Equation 1). As expected, there was a clear difference between both *H'* estimates (see Figure S5). In particular, entropy values derived from the random test were found to be approximately symmetric and consistently low among the GenF size categories (compared with the real values), yielding values of -0.09, and 0.1 as 5% and 95% percentiles, respectively (Table S5).

Sulfur Score and its predictive capacity to detect S-microbial players in a large genomic dataset.

To test whether Pfam entropies can be combined to capture the S-metabolic machinery in 'omic'-samples, we calculated the final MEBS score, called in this case Sulfur Score (*SS*). We computed the *SS* on each of the 2,107 non-redundant genomes in Gen with script *score_genomes.sh.* The individual genomes along with their corresponding *SS* values and taxonomy according to NCBI are found in Table S6.

For evaluation purposes, we classified and manually annotated all the genomes in Gen according to their metabolic capabilities. First, we identified the 161-curated genomes belonging to Suli. Then, we focused on the remaining genomes. A set of 192 genomes with SS>4 were labeled as Sulfur unconsidered or related microorganisms (Sur). Finally, the rest of genomes in Gen were classified as NS (Non-Sulfur = Gen – (Suli + Sur)), including 1,754 genomes. The boxplots in Figure 5A summarize the scores obtained in these three subsets.

To double-check whether the Sur genomes -- selected due to their *SS* -- might be involved in the S-cycle, we manually annotated all of them focusing on relevant genomic, biochemical, physiological and environmental information that we might have missed since Suli was first curated (Table S7). Out of 192 genomes, 68 are reported to metabolize S-compounds under culture conditions in the literature. For instance, *Sideroxydans lithotrophicus ES*-1, a microaerophilic Fe-oxidizing bacterium, has been observed to also grow in thiosulfate as an energy source [58]. Another 59 Sur organisms have been isolated from Sulfur-rich environments, such as hot springs or solfataric muds.

412 Remarkably, some of this species include hard-to culture genomes reconstructed from
413 metagenomic sequences such as *Candidatus Desulforudis audaxviator MP104C* isolated from
414 basalt-hosted fluids of the deep subseafloor [6]; an unnamed endosymbiont of a scaly snail from a
415 black smoker chimney [59] and archaeon *Geoglobus ahangari,* sampled from a 2,000m depth
416 hydrothermal vent [60]. Furthermore, we also confirmed within Sur the implication of S-cycle of 20
417 species of the genus *Campylobacter*. These results are consistent with the ecological role of the
418 involved taxa, that along with SRB and methanogens inhabiting host-gastrointestinal and low
419 oxygen environments, where several inorganic (e.g., sulfates, sulfites) or organic (e.g., dietary
420 amino acids and host mucins) are highly metabolized by these metabolic guilds [61]. The
421 implication of *Campylobacter* species in the S-cycle is also supported by the fact that some of them
422 have been isolated from deep sea hydrothermal vents [62]. The remaining species in Sur were
423 classified in different categories, including bioremediation (7), Fe-environment (2), marine (2), peat
424 lands (2) and other environments (32, see Figure 5B).
425 When the *SS* values of genomes in Sur are compared to the S-metabolic guilds represented in Suli
426 (e.g PSB, SRB, GSB), it can be seen that they are indeed similar and clearly separated from the rest
427 of NS genomes (Figure 5C). This strongly suggests that high scoring genomes are indeed
428 ecologically and metabolically implicated in the S-cycle.
429 Finally, in order to quantify the capacity of the *SS* to accurately classify S-related microorganisms,
430 we computed a Receiver Operator Characteristic (ROC) curve (for a detailed description of ROC
431 curves see [63]). We thus defined genomes annotated in Suli as positive instances, and the rest as
432 negative ones. The results are shown in Figure 5D, with an estimated Area Under the Curve (AUC)
433 of 0.985, and the corresponding cut-off values of *SS* for several False Positive Rates (FPR).
434 According to this test, a *SS value of* 8.705 is required to rule out all false positives in Gen, while
435 *SS*=5.231 is sufficient to achieve a FPR < 0.05.
436 Overall, these results indicate that MEBS is a powerful and broadly applicable approach to predict,
437 and classify microorganisms closely involved in the sulfur cycle even in hard-to culture microbial
438 lineages.
439

Sulfur Score and its predictive capacity to detect S-related environments in a large metagenomic collection.

The SS was also computed for each metagenome in Met, using their corresponding MSL to choose the appropriate entropies previously calculated in dataset GenF (Table S8). In order to test whether *SS* values can be used to identify S-related environments, we performed the following analyses. First, we use the geographical metadata associated with each metagenome to map the global distribution of *SS*. In Figure 6A, *SS* values are colored from yellow to red. The most informative S-environments (displaying *SS* values equal or greater than the 95th percentile of each MSL category) are shown in blue.

Then, we sorted the metagenomes according to their environmental features as proposed by the Genomic Standards Consortium [GSC] and implemented in MG-RAST. Each feature corresponds to one of 13 environmental packages (EP) that standardize metadata describing particular habitats that are applicable across all GSC checklists and beyond [64]. Therefore, each EP represents a broad and general classification containing particular features. For example, the "water" EP includes 330 metagenomes from our dataset, belonging to several features such as freshwater, lakes, estuarine, marine, hydrothermal vents, etc. Since each of these features has different ecological capabilities in terms of biogeochemical cycles, we can expect different behaviors among *SS* values, as shown in Figure 6B. In general, all the metagenomes derived from hydrothermal vents (2), marine benthic (6), intertidal (8) and our unpublished CCC microbial mats had *SS* values above the 95th percentile, highlighting the importance of the S-cycle in these environments. In contrast, the metagenomes belonging to features such as sub-terrestrial habitat (7), saline evaporation pond (24) or organisms associated habitat (7) displayed consistently low or even negative *SS* values, indicating a negligible presence of S-metabolic pathways in those environments. The remaining features have intermediate median *SS* values and contain occasionally individual metagenomes with *SS* values above the 95th percentile, such as freshwater, marine, ocean or biofilm environments.

To validate the list of 50 high-scoring metagenomes (above the 95th percentiles), we double-checked their annotations. According to the literature and associated metadata, all these environments are closely involved in mineralization, uptake, and recycling processes of S-

compounds. For example, environmental sequences derived from costal Oligochaete worm *Olavius algarvensis,* hydrothermal vents and marine deep-sea surface sediments around the Deep-Water Horizon spill in the Gulf of Mexico. The complete list of annotated metagenomes, along with their ecological capabilities, is found in Table S9.

Evaluating the robustness of the Sulfur Score

To test the reproducibility and robustness of MEBS final score (*SS*), we conducted two further analyses. In the first one we compared *SS* estimates derived from Met dataset, computed with Pfam entropies obtained in the first MEBS benchmark performed three years ago (2014) with the current data described in this article (2017). Despite the changes of both databases (Pfam database version and the Suli list), we found a strong correlation (r^2=0.912) between the *SS* outcomes (Figure S6 A). A kernel density analysis of the latter comparison suggests a different behavior of low and high *SS* scores, with the latter being more reproducible (see Figure S6B).

In the second analysis, we quantitatively tested to what extent the entropy estimates of the 112 Pfam domains directly affect the outcome of the *SS* in Gen and Met. We randomly subsampled ≈50% of those domains to compute the *SS* a thousand times for each genome and metagenome in Gen and Met, respectively. The results, summarized in Table S10, confirm that *SS* values computed with random subsets of Pfam domains are generally lower than *SS* derived from the full list (n=112) of Sucy-Pfam domains. To further inspect the distribution of *SS* values produced with random subsets of domains (random *SS*), we focused on the particular case of the metagenomes belonging to the category MSL=60. As expected, the distribution of random *SS* oscillates between negative and positive values. Interestingly, metagenomes exhibiting only positive random *SS* are ranked above the 95th percentile according to their real *SS* values (See Figure S7A). The latter indicates that even a random subset of Pfam domains are used to compute the score, is more likely to high-rank metagenomes containing the sulfur metabolic machinery (large number of high-entropy Pfam domains), than those lacking the sulfur metabolism or displaying a large number of non-informative Pfam domains. Furthermore, by comparing the median of random *SS* with the real scores, we observe a clear separation between those distributions (see Figure S7B and Table S10).

Completeness of S-metabolic pathways

As we described above, the MEBS pipeline models a metabolic network as an array of S-related protein domains (Sucy-Pfam), to ultimately use their entropies to produce the final score (*SS*). For a closer look, we also dissected the total contribution of independent domains at the network level, in order to assess whether *SS* depends on the partial or complete detection of S-pathways. Consequently, we evaluated the pathway completeness in both genomic (Gen) and metagenomic (Met) datasets (see Tables S11 and S12, respectively). Since the number of Pfam domains per pathway goes from one to 29 (see Table 1 and Table S2), we suspect that pathways represented by a single domain might not reflect their complete metabolic function. For example, the pathways involved in the methanogenesis of compounds such as dimethylsulfide (DMS, P24), methyl-thiolpropanoate (MTPA, P25), and methanethiol (MeSH, P26) are represented by the same protein (MtsA, PF01208) in our Sucy database, as well as in Metacyc [12]. Therefore, we expect that pathways P24-26 will have identical presence-absence patterns in Gen and Met.

The boxplots in Figure 7A and 7B summarize the distribution of completeness for each S-metabolic pathway including the synthetic pathway (P29) composed by 12 candidate markers as described in Table 2. As expected, the observed completeness per pathway was higher in Met than in Gen, since microbial communities harbor a wider repertory of metabolic functions than single genomes. In the case of genomes, we noted that a few pathways were complete in most genomes, being the majority involved in the usage of organic sulfur compounds such as alkanesulfonates (P9), sulfoacetate (P14) and biosynthesis of sulfolipids (SQDG) and the single domain pathways P24-26. Remarkably, we also detected a few organisms displaying the highest levels of metabolic completeness in some S-energy based pathways. For example, we found that *Desulfosporosinus acidiphilus SJ4* (SS=8,91) was the only genome harboring the complete repertory of Pfam domains described in Sucy for the sulfite oxidation (P1), strongly suggesting that it may oxidize sulfite. However, this activity remains to be tested in culture [65]. In the case of thiosulfate oxidation (P3), we detected three genomes displaying the highest levels of completeness, in agreement with their ecological features: *Hydrogenobaculum sp. Y04AAS1* (SS=9,319) [66] and the CLSB: *Acidithiobacillus caldus ATCC 51756* (SS=6,525) [67] and *Acidithiobacillus ferrivorans* (SS=7,436) [68]. For the sulfate reduction dissimilative pathway (P5), out of 55 genomes displaying the higher completeness levels,

67% are actually SRB, 12% are Sur genomes, and the rest are sulfur oxidation microorganisms. Furthermore, the PSB *Thioflavicoccus mobilis 8321* (SS= 9,756), isolated from a microbial mat [69], was the genome displaying the most complete sulfide oxidation pathway (P11). Elemental sulfur disproportionation (P21) is represented by a single non-informative domain (PF07682, *H'*=0.172) that remarkably is found in 14 sulfur respiring or related genomes such as *Sulfolobus tokodaii str. 7* (SS= 5,341) and *Acidianus hospitalis W1* (SS= 3,88). Finally, we identified six genomes encoding all 12 proposed markers. Among them, three were GSB (*Pelodictyon phaeoclathratiforme BU-1,* SS=11,836, *Chlorobium chlorochromatii CaD3,* SS=11,625 and *Chlorobium tepidum TLS,* SS= 11,354), one CLSB (Thiobacillus denitrificans ATCC 25259 SS=11,61) ,another one PSB (*Thiocystis violascens DSM 198,* SS=10,633) and finally one Sur (*Sedimenticola thiotaurini* SS=10,109). For a complete description, see Table S13.

A global view of metabolic completeness was obtained by bulking the data from all pathways. Linear regression models between mean completeness and *SS* were computed confirm the , yielding r^2 values of 0.003 and 0.627 for Gen and Met, respectively (See Figures 7C and 7D). Moreover, we also assessed the relationship between the mean completeness of the synthetic pathway of candidate markers (P29) and the *SS*. As expected, significant correlations were obtained in both datasets (r^2= 0.645 and r^2=0.881 for Gen and Met, respectively; see Figure S8).

To get a more detailed insight of the completeness, we selected a few genomes and metagenomes displaying high and low *SS* values. Specifically, from the Gen dataset we selected one representative from the main S-guilds, one Sur genome and two genomes with low *SS* values (NS). As observed in Figure 7, the low-scoring genomes *Enterococcus durans* (SS=-0,194), *Micrococcus luteus NCTC_2665* (SS=-3,588), *and Ruegeria pomeroyi DSS-3* (SS=2,707) display unrelated patterns of sulfur metabolic completeness, compared with the rest of genomes and therefore are separated. In contrast, high-scoring S-respiring microorganisms *Desulfovibrio vulgaris DP4* (SS= 11,442), *Sulfolobus acidocaldarius DSM 639* (SS=5,457) and *Ammonifex degensii KC4* (SS=12.508) are clustered together. We also observed that mat-isolated cyanobacteria *Synechococcus sp. JA-2-3Ba 2-13*, classified as NS with SS=3,704, was clustered together with other high-scoring genomes, in agreement with the lack of correlation reported above.

In the case of metagenomes (see Figure 7E), we observed a clear correlation between *SS* and completeness. For example, metagenomes 4440320.3 and 4489656.3, with the lowest scores (SS=0.1 and SS=-2.649, respectively), also exhibit the largest number of incomplete pathways. Similarly, high-scoring metagenomes derived from black smoker or marine sediment are grouped together in terms of completeness.

Conclusions

Our study represents the first exploration of the Sulfur biogeochemical cycle in a large collection of genomes and metagenomes. The manually curated effort resulted in an inventory of the compounds, genes, proteins, molecular pathways, and microorganisms involved. This complex universe of articulated data was reduced into a list of microorganisms and Pfam domains encoded in the proteins that take part in that network. These domains were first ranked in terms of relative entropy, and then summed to produce a single S-score representing the relevance of a given genomic or metagenomic sample in terms of sulfur metabolic machinery. We took advantage of the mathematical framework of information theory, which has been widely used in computational biology.

The performance of the Multigenomic Entropy Based Score pipeline (MEBS) (designed for the above mentioned tasks) was benchmarked on large genomic and metagenomic sets. Our results support the broad applicability of this algorithm in order to classify annotated genomes as well as newly sequenced environmental samples without prior culture. We also assessed to what extent the final score depended on the partial or complete detection of pathways and observed a higher completeness per pathway in metagenomic sequences than in individual genomes.

We demonstrated that a measurable score can be applied to evaluate any given metabolic machinery or biogeochemical cycle in large (meta)genomic scale, holding the potential to dramatically change the current view of inferring metabolic capabilities in the present 'omic'-era.

Availability and requirements

Project name: MEBS

Project home page: https://github.com/eead-csic-compbio/metagenome_Pfam_score

584 Operating system(s): Linux

585 Programming language: Python 3, Perl5, Bash,

586 Other requirements: HMMER

587 License: GNU General Public License (GPL)

588 Availability of supporting data

589 The datasets supporting the results of this article are available in the GigaDB repository [REF#]

590

591 Abbreviations

592 MEBS: Multigenomic Entropy Based Score ; S: Sulfur ; S-cycle: Sulfur cycle; SS: Sulfur Score; Suli:
593 Sulfur list ; Sucy: Sulfur cycle database; Rlist: Random list of taxonomic representatives ; MSL:
594 Mean Size Length, H': Relative Entropy ; Sur: Sulfur unconsidered ; NS: Non sulfur related
595 genomes; Gen: Genomic dataset; Met: Metagenomic dataset; GenF : Genomic Fragmented
596 dataset; CLSB: Color-less Sulfur Bacteria; SOM: Sulfur Oxidizing Microorganims; GSB: Green Sulfur
597 Bacteria; PSB: Purple Sulfur Bacteria; SRB: Sulfate Reducing Bacteria; ESR: Elemental-Sulfur
598 Reducing microorganisms; CCC: Cuatro Cienegas, Coahuila; HMM: Hidden Markov Models (HMMs);
599 ROC: Receiver-operating characteristic; AUC: Area Under the Curve; TPR True Positive Rates; FPR
600 False Positive Rates; GSC: Genomic Standards Consortium; EP: environmental packages.

601

602

603 Acknowledgments

604 The authors gratefully acknowledge Emilio Morella for their valuable support, feedback and comments
605 throughout the development of this project. We really appreciate the reviewers Daan Speth and Thulani
606 Makhalanyane whose valuable comments and suggestions critically improve the manuscript and sofware.
607 Our special thanks to acknowledge Carlos P Cantalapiedra, Seth Barribeau, Will Levitt and anonymous
608 reviewers for their comments on earlier versions of the manuscript who immensely improved the algorithm
609 and the final version of the article. The authors also thank the Laboratory of Computational and Structural
610 Biology (EEAD-CSIC) and Laboratorio de Evolución Molecular y Experimental, Instituto de Ecología, UNAM
611 for providing the computational resources described in the article. The paper was written during a
612 sabbatical leave of LEE and VSS in the University of Minnesota in Peter Tiffin and Michael Travisano
613 laboratories.

614 Funding

615 VDA is a doctoral student from Programa de Doctorado en Ciencias Biomédicas, Universidad Nacional
616 Autónoma de México (UNAM) and received fellowship 356832 from CONACYT. This research was also
617 supported by funding from WWF-Alianza Carlos Slim, Sep-Ciencia Básica Conacyt grant 238245 to both VS
618 and LEE and Spanish MINECO grant CSIC13-4E-2490. BCM was funded by Fundación ARAID. The sabbatical

leave of LEE and VSS at the University of Minnesota were supported by scholarships from PASPA, DGAPA, UNAM.

Competing interest

The authors declare that they have no competing interest.

Author contribution

VDA, BCM and IZP wrote the paper. BCM, VDA and ACPH developed and wrote the software and performed all the bioinformatics analyses. VDA produced all the figures and wrote the documentation of the software. VDA and IZP conceived the manual curation of the Sulfur cycle inventory and the microbiological, biogeochemical, and ecological interpretation. LE and VS provided the intellectual framework, expertise and resources to develop and supervise the project. All the authors read and approved the final manuscript.

Endnotes

We are currently finishing the analyses to demonstrate the applicability of this approach to other biogeochemical cycles (C, N, O, Fe, P). Thereby, we hope that the pipeline MEBS will facilitate analysis of biogeochemical cycles or complex metabolic networks carried out by specific prokaryotic guilds, such as bioremediation processes (i.e., degradation of hydrocarbons, toxic aromatic compounds, heavy metals etc.). We look forward to collaborate and help other researchers by integrating comprehensive databases that might be helpful to the scientific community. Furthermore, we are currently working to improve the algorithm by using only a list of sequenced genomes involved in the metabolism of interest, in order to reduce the manual curation effort. We are also considering taking *k-mers* instead of peptide Hidden Markov Models to increase the speed of the pipeline. We anticipate that our platform will stimulate interest and involvement among the scientific community to explore uncultured genomes derived from large metagenomic sequences.

References

1. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb. Inform. Exp. [Internet]. BioMed Central Ltd; 2012;2:3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3351745%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract

2. Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, et al. Metagenomics: Tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinform. Biol. Insights. 2015;9:75–88.

3. Morales SE, Holben WE. Linking bacterial identities and ecosystem processes: Can “omic” analyses be more than the sum of their parts? FEMS Microbiol. Ecol. 2011;75:2–16.

4. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat. Microbiol. [Internet]. 2016;1:16048. Available from:

http://www.nature.com/articles/nmicrobiol201648

5. Marco D. Metagenomics : Current Innovations and Future Trends. Caister Academic Press; 2011.

6. Jungbluth SP, Glavina del Rio T, Tringe SG, Stepanauskas R, Rappé MS. Genomic comparisons of a bacterial lineage that inhabits both marine and terrestrial deep subsurface systems. PeerJ [Internet]. 2017;5:e3134. Available from: https://peerj.com/articles/3134

7. Iverson V, Morris RM, Frazar CD, Berthiaume CT, Morales RL, Armbrust E V. Untangling Genomes from Metagenomes: Revealing an Uncultured Class of Marine Euryarchaeota. Science (80-.). 2012;335:587–90.

8. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome [Internet]. Microbiome; 2016;4:8. Available from: http://microbiomejournal.biomedcentral.com/articles/10.1186/s40168-016-0154-5

9. Mehrshad M, Amoozegar MA, Ghai R, Shahzadeh Fazeli SA, Rodriguez-Valera F. Genome reconstruction from metagenomic data sets reveals novel microbes in the brackish waters of the Caspian Sea. Appl. Environ. Microbiol. 2016;82:1599–612.

10. Sharon I, Banfield JF. Genomes from Metagenomics. Science [Internet]. 2013;342:1057–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24288324

11. Hiraoka S, Yang C, Iwasaki W. Metagenomics and Bioinformatics in Microbial Ecology: Current Status and Beyond. Microbes Environ. [Internet]. 2016;31:204–12. Available from: https://www.jstage.jst.go.jp/article/jsme2/31/3/31_ME16024/_article

12. Caspi R, Altman T, Dreher K, Fulcher C a, Subhraveti P, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. [Internet]. 2012 [cited 2013 May 27];40:D742-53. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245006&tool=pmcentrez&rendertype=abstract

13. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. [Internet]. 2000;28:27–30. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=102409%7B&%7Dtool=pmcentrez%7B&%7Drendertype=abstract

14. Ye Y, Doak TG. A Parsimony Approach to Biological Pathway Reconstruction/Inference for Metagenomes. Handb. Mol. Microb. Ecol. I Metagenomics Complement. Approaches. 2011;5:453–60.

15. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. PLoS Comput. Biol. [Internet]. 2012 [cited 2014 Jan 23];8:e1002358. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3374609&tool=pmcentrez&rendertype=abstract

16. Larsen PE, Collart FR, Field D, Meyer F, Keegan KP, Henry CS, et al. Predicted Relative

694 Metabolomic Turnover (PRMT): determining metabolic turnover from a coastal marine
695 metagenomic dataset. Microb. Inform. Exp. [Internet]. BioMed Central Ltd; 2011;1:4. Available
696 from: http://www.microbialinformaticsj.com/content/1/1/4

697 17. Hanson NW, Konwar KM, Hawley AK, Altman T, Karp PD, Hallam SJ. Metabolic pathways for the
698 whole community. BMC Genomics [Internet]. 2014;15:619. Available from:
699 http://www.biomedcentral.com/1471-2164/15/619

700 18. Castañeda LE, Barbosa O. Metagenomic analysis exploring taxonomic and functional diversity
701 of soil microbial communities in Chilean vineyards and surrounding native forests. PeerJ [Internet].
702 2017;5:e3098. Available from: https://peerj.com/articles/3098

703 19. Fierer N, Leff JW, Adams BJ, Nielsen UN, Bates ST, Lauber CL, et al. Cross-biome metagenomic
704 analyses of soil microbial communities and their functional attributes. Proc. Natl. Acad. Sci.
705 [Internet]. 2012;109:21390–5. Available from:
706 http://www.pnas.org/cgi/doi/10.1073/pnas.1215210110

707 20. Llorens-Marès T, Yooseph S, Goll J, Hoffman J, Vila-Costa M, Borrego CM, et al. Connecting
708 biodiversity and potential functional role in modern euxinic environments by microbial
709 metagenomics. ISME J. [Internet]. 2015;1–14. Available from:
710 http://www.nature.com/doifinder/10.1038/ismej.2014.254

711 21. Quaiser A, Zivanovic Y, Moreira D, López-García P. Comparative metagenomics of bathypelagic
712 plankton and bottom sediment from the Sea of Marmara. ISME J. [Internet]. 2011 [cited 2014 Jul
713 16];5:285–304. Available from:
714 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3105693&tool=pmcentrez&rendertyp
715 e=abstract

716 22. Xie W, Wang F, Guo L, Chen Z, Sievert SM, Meng J, et al. Comparative metagenomics of
717 microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting
718 chemistries. ISME J. [Internet]. Nature Publishing Group; 2011;5:414–26. Available from:
719 http://dx.doi.org/10.1038/ismej.2010.144

720 23. Delmont TO, Malandain C, Prestat E, Larose C, Monier J-M, Simonet P, et al. Metagenomic
721 mining for microbiologists. ISME J. [Internet]. 2011 [cited 2014 Apr 30];5:1837–43. Available from:
722 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3223302&tool=pmcentrez&rendertyp
723 e=abstract

724 24. Ganesh S, Parris DJ, DeLong EF, Stewart FJ. Metagenomic analysis of size-fractionated
725 picoplankton in a marine oxygen minimum zone. ISME J. [Internet]. Nature Publishing Group;
726 2014;8:187–211. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24030599

727 25. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, et al. Metagenomic biomarker
728 discovery and explanation. Genome Biol. [Internet]. BioMed Central Ltd; 2011;12:R60. Available
729 from: http://genomebiology.com/2011/11/6/R60

730 26. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: Statistical analysis of taxonomic and
731 functional profiles. Bioinformatics. 2014;30:3123–4.

732 27. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in
733 metagenomic samples. Bioinformatics. 2015;31:2269–75.

734 28. Kullback S, Leibler RA. On Information and Sufficiency. Ann. Math. Stat. 1951;22:79–86.

735 29. Commenges D. Information Theory and Statistics: an overview. 2015;1–22. Available from:
736 http://arxiv.org/abs/1511.00860

737 30. Hertz GZ, Stormo GD. Identifying DNA and protein patterns with statistically significant
738 alignments of multiple sequences. Bioinformatics [Internet]. 1999;15:563–77. Available from:
739 http://www.ncbi.nlm.nih.gov/pubmed/10487864

740 31. Dar S a, Yao L, van Dongen U, Kuenen JG, Muyzer G. Analysis of diversity and activity of sulfate-
741 reducing bacterial communities in sulfidogenic bioreactors using 16S rRNA and dsrB genes as
742 molecular markers. Appl. Environ. Microbiol. [Internet]. 2007;73:594–604. Available from:
743 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1796976%7B&%7Dtool=pmcentrez%
744 7B&%7Drendertype=abstract

745 32. Kerkhof LJ. Phylogeography of Sulfate-Reducing Bacteria among Disturbed Sediments ,
746 Disclosed by Analysis of the Dissimilatory Sulfite Reductase Genes (dsrAB) Phylogeography of
747 Sulfate-Reducing Bacteria among Disturbed Sediments , Disclosed by Analysis of the Dissi. 2005;

748 33. Loy A, Duller S, Baranyi C, Mussmann M, Ott J, Sharon I, et al. Reverse dissimilatory sulfite
749 reductase as phylogenetic marker for a subgroup of sulfur-oxidizing prokaryotes. Environ.
750 Microbiol. [Internet]. 2009;11:289–99. Available from:
751 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2702494%7B&%7Dtool=pmcentrez%
752 7B&%7Drendertype=abstract

753 34. Hügler M, Gärtner A, Imhoff JF. Functional genes as markers for sulfur cycling and CO2 fixation
754 in microbial communities of hydrothermal vents of the Logatchev field. FEMS Microbiol. Ecol.
755 [Internet]. 2010 [cited 2013 Nov 3];73:526–37. Available from:
756 http://www.ncbi.nlm.nih.gov/pubmed/20597983

757 35. Meyer B, Kuever J. Molecular analysis of the diversity of sulfate-reducing and sulfur-oxidizing
758 prokaryotes in the environment, using aprA as functional marker gene. Appl. Environ. Microbiol.
759 [Internet]. 2007 [cited 2013 Jun 23];73:7664–79. Available from:
760 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2168068&tool=pmcentrez&rendertyp
761 e=abstract

762 36. MEBS [Internet]. Available from: https://github.com/eead-csic-
763 compbio/metagenome_Pfam_score

764 37. Enzyme Nomenclature [Internet]. Available from: http://enzyme.expasy.org/. Accesed 05 May
765 2016.

766 38. Magrane M, Consortium UP. UniProt Knowledgebase: A hub of integrated protein data.
767 Database. 2011;2011:1–13.

768 39. Reference and Representative Genomes [Internet]. Available from:

769 https://www.ncbi.nlm.nih.gov/genome/browse/reference/. Accesed 21 Dec. 2016

770 40. Genome clusters [Internet]. Available from:
771 http://microbiome.wlu.ca/research/redundancy/redundancy.cgi. Accesed 21 Dec. 2016

772 41. Moreno-Hagelsieb G, Wang Z, Walsh S, ElSherbiny A. Phylogenomic clustering for selecting
773 non-redundant genomes for comparative genomics. Bioinformatics [Internet]. 2013 [cited 2014
774 Sep 18];29:947–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23396122

775 42. NCBI genome assembly summary file [Internet]. Available from:
776 ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/assembly_summary_refseq.txt. Accesed 21 Dec. 2016.

777 43. NCBI. NCBI FTP site [Internet]. Available from: ftp://ftp.ncbi.nlm.nih.gov/genomes/all/. Accesed
778 21 Dec. 2016.

779 44. MG-RAST [Internet]. Available from: http://metagenomics.anl.gov/. Accesed 10 Nov. 2016.

780 45. Prakash T, Taylor TD. Functional assignment of metagenomic data: challenges and applications.
781 Brief. Bioinform. [Internet]. 2012;13:711–27. Available from:
782 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3504928%7B&%7Dtool=pmcentrez%
783 7B&%7Drendertype=abstract

784 46. Zhong C, Yang Y, Yooseph S. GRASP: Guided reference-based assembly of short peptides.
785 Nucleic Acids Res. 2015;43:e18.

786 47. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, et al. InterProScan 5: Genome-scale
787 protein function classification. Bioinformatics. 2014;30:1236–40.

788 48. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H-R, et al. The Pfam protein families
789 database. Nucleic Acids Res. [Internet]. 2008;36:D281--8. Available from:
790 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238907%7B&%7Dtool=pmcentrez%
791 7B&%7Drendertype=abstract

792 49. Haft DH. The TIGRFAMs database of protein families. Nucleic Acids Res. [Internet]. 2003 [cited
793 2014 Aug 11];31:371–3. Available from:
794 http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkg128

795 50. Gough J, Chothia C. SUPERFAMILY: HMMs representing all proteins of known structure. SCOP
796 sequence searches, alignments and genome assignments. Nucleic Acids Res. [Internet].
797 2002;30:268–72. Available from:
798 http://www.ncbi.nlm.nih.gov/pmc/articles/PMC99153/%5Cnhttp://www.ncbi.nlm.nih.gov/pmc/ar
799 ticles/PMC99153/pdf/gkf097.pdf

800 51. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching.
801 Nucleic Acids Res. [Internet]. 2011;39:W29--37. Available from:
802 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3125773%7B&%7Dtool=pmcentrez%
803 7B&%7Drendertype=abstract

804 52. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine
805 Learning in {P}ython. J. Mach. Learn. Res. 2011;12:2825–30.

806 53. KEGG Mapper [Internet]. Available from: http://www.genome.jp/kegg/mapper.html. Accesed 4
807 Sept. 2017.

808 54. Alcaraz LD, Olmedo G, Bonilla G, Cerritos R, Hernández G, Cruz A, et al. The genome of Bacillus
809 coahuilensis reveals adaptations essential for survival in the relic of an ancient marine
810 environment. Proc. Natl. Acad. Sci. U. S. A. [Internet]. 2008;105:5803–8. Available from:
811 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2311347%7B&%7Dtool=pmcentrez%
812 7B&%7Drendertype=abstract

813 55. Pedroni P, Volpe AD, Galli G, Mura GM, Pratesi C, Grandi G. Characterization of the locus
814 encoding the [Ni-Fe] sulfhydrogenase from the archaeon Pyrococcus furiosus: Evidence for a
815 relationship to bacterial sulfite reductases. Microbiology. 1995;141:449–58.

816 56. Taguchi Y, Sugishima M, Fukuyama K. Crystal Structure of a Novel Zinc-Binding ATP Sulfurylase
817 from Thermus. 2004;4111–8.

818 57. Santos AA, Venceslau SS, Grein F, Leavitt WD, Dahl C, Johnston DT, et al. A protein trisulfide
819 couples dissimilatory sulfate reduction to energy conservation. Science (80-.). 2015;350:1541–5.

820 58. Emerson D, Field EK, Chertkov O, Davenport KW, Goodwin L, Munk C, et al. Comparative
821 genomics of freshwater Fe-oxidizing bacteria: implications for physiology, ecology, and
822 systematics. Front. Microbiol. [Internet]. 2013 [cited 2014 Jun 11];4:254. Available from:
823 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3770913&tool=pmcentrez&rendertyp
824 e=abstract

825 59. Nakagawa S, Shimamura S, Takaki Y, Suzuki Y, Murakami S, Watanabe T, et al. Allying with
826 armored snails: the complete genome of gammaproteobacterial endosymbiont. ISME J. [Internet].
827 Nature Publishing Group; 2014;8:40–51. Available from:
828 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3869010&tool=pmcentrez&rendertyp
829 e=abstract

830 60. Manzella MP, Holmes DE, Rocheleau JM, Chung A, Reguera G, Kashefi K. The complete genome
831 sequence and emendation of the hyperthermophilic, obligate iron-reducing archaeon “Geoglobus
832 ahangari” strain 234T. Stand. Genomic Sci. [Internet]. Standards in Genomic Sciences; 2015;10:77.
833 Available from:
834 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4600277%7B&%7Dtool=pmcentrez%
835 7B&%7Drendertype=abstract

836 61. Carbonero F, Benefiel AC, Alizadeh-Ghamsari AH, Gaskins HR. Microbial pathways in colonic
837 sulfur metabolism and links with health and disease. Front. Physiol. 2012;3 NOV:1–11.

838 62. Nakagawa S, Takaki Y, Shimamura S, Reysenbach A-L, Takai K, Horikoshi K. Deep-sea vent
839 epsilon-proteobacterial genomes provide insights into emergence of pathogens. Proc. Natl. Acad.
840 Sci. U. S. A. 2007;104:12146–50.

841 63. Fawcett T. An introduction to ROC analysis. Pattern Recognit. Lett. 2006;27:861–74.

842 64. Field D, Sterk P, Kottmann R, Smet JW De, Amaral-zettler L, Cole JR, et al. Genomic Standards
843 Consortium Projects The Genomic Standards Consortium Initiating and Maintaining a Project

within the GSC The GSC Project Description template provides a References: 2014;599–601.

65. Alazard D, Joseph M, Battaglia-Brunet F, Cayol JL, Ollivier B. Desulfosporosinus acidiphilus sp. nov.: A moderately acidophilic sulfate-reducing bacterium isolated from acid mining drainage sediments. Extremophiles. 2010;14:305–12.

66. Romano C, D'Imperio S, Woyke T, Mavromatis K, Lasken R, Shock EL, et al. Comparative genomic analysis of phylogenetically closely related Hydrogenobaculum sp. isolates from yellowstone national park. Appl. Environ. Microbiol. 2013;79:2932–43.

67. Chen L, Ren Y, Lin J, Liu X, Pang X, Lin J. Acidithiobacillus caldus sulfur oxidation model based on transcriptome analysis between the wild type and sulfur oxygenase reductase defective mutant. PLoS One [Internet]. 2012 [cited 2013 Apr 24];7:e39470. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3440390&tool=pmcentrez&rendertype=abstract

68. Liljeqvist M, Valdes J, Holmes DS, Dopson M. Draft genome of the psychrotolerant acidophile Acidithiobacillus ferrivorans SS3. J. Bacteriol. 2011;193:4304–5.

69. Imhoff JF, Pfennig N. Thioflavicoccus mobilis gen. nov., sp. nov., a novel purple sulfur bacterium with bacteriochlorophyll b. Int. J. Syst. Evol. Microbiol. 2001;51:105–10.

63. GigaDB DOI Citation

Figure Legends

Figure 1. Schematic representation of the four stages of the MEBS algorithm focusing on the S-cycle. The first step consists on the systematic curation of a database containing the metabolic information of S-cycle, which is reduced to a FASTA-file of proteins involved (Sucy) and a list of 161 related microorganisms (Suli). A thousand lists of 161 random-sampled genomes were used as negative control (Rlist). The training dataset comprises 2,107 genomes (Gen), which were fragmented in different sizes by considering the Mean Size Length (MSL) of 935 metagenomes (Met). In the second stage the domain composition of Sucy proteins is obtained by scanning Pfam-A, resulting in the Pfam-Sucy database. Then, the relative entropy (*H'*) of each Sucy-Pfam domain is obtained in the third stage. Finally, the precomputed entropies in Gen and GenF are used to evaluate full-length genomic sequences (real) and metagenomic sequences of variable MSL (in this example A, B and C).

Figure 2. Sulfur cycle at global scale. The most important organic and inorganic S-compounds derived from biogeochemical processes are arranged according to the Standard Gibbs free energy of formation described in Caspi et al., (2012). The left column indicates whether specific microorganisms are able to use those S-compounds, as a source of Carbon (C), Nitrogen (N), Energy (E) or Electron donors (°). Double asterisks indicate whether the S-compound is used as sole source, of C, N, or E. The corresponding electron acceptors in redox-coupled reactions using the S-compound as electron donor are not shown. The right column

881 indicates whether the S-compound is used as fermentative substrate (F) or terminal electron acceptor in
882 respiratory processes (R). Colored boxes summarize the metabolic guilds involved in the metabolism of S-
883 compounds, in oxidation (i.e., CLSB, SOM, PSB, and GSB) or reduction (SR, SRB) processes. The complete list
884 of S-based microorganisms (Suli) is found in Table S1. Figure based on annotations from MetaCyc [12].

885 **Figure 3.** Comprehensive network representation of the machinery of the biogeochemical S-cycle in a single
886 cell. The 28 molecular pathways involved in the metabolism of sulfur compounds described in Table 1 are
887 included. The enzymatic steps are depicted as rectangles followed by arrows indicating the direction of the
888 reaction. Green hexagons represent metabolic links to other metabolisms. Bold dashed arrows indicate
889 bidirectional reactions. Inorganic S-compounds have been arranged according to their reduction potential,
890 from the most oxidized (yellow) to the most reduced (red) compounds. Grey rectangles indicate enzymes
891 acting in disproportionation processes in which a reactant is both oxidized and reduced in the same
892 chemical reaction, forming two separate compounds. Input biogeochemical S-compounds are shown
893 outside and connected with bold arrows. Dashed arrows indicate S-compounds excreted out of the cell. The
894 upper half of the modeled cell depicts the processes involved in the use of organic S-compounds (orange
895 circles) found in natural environments and used as source of carbon, sulfur and/or energy in several
896 aerobic/anaerobic strains described in Figure 2.

897 **Figure 4**. Entropy values of Sulfur-derived protein domains. A) Heatmap showing the entropy values (*H')* of
898 the 112 Pfam domains identified in proteins curated in SuCy. B) Difference between entropies estimated
899 from sizes categories of growing peptide size (GenF) and the real values measured within complete
900 genomes (Gen). Error bars show standard deviations. Both graphs were obtained with script
901 *plot_entropy.py*. Clustering of the Pfam relative entropies obtained in Gen and GenF produced with the
902 Ward method. Log frequency of the entropy values computed in the random test is colored in purple (see
903 scale bar). Cluster 0 (blue) groups protein domains with low relative entropy that overlap with the random
904 distribution. Cluster 1 (green) includes the Pfam domains that fulfill the requirements to be used as
905 molecular markers (high H' and low standard deviation, std). Red dots (cluster 2) correspond to Pfam
906 domains with high *H'* and std. The cluster was produced with script *F_meanVSstd.py*

907 **Figure 5.** Distribution of Sulfur Score (*SS*) in 2,107 non-redundant genomes (Gen). A) Subsets of genomes
908 annotated in Suli (n=161); ii) Sur, genomes not listed in Suli with SS > 4 and candidates to be S-related
909 microorganisms (n=192); iii) rest of the genomes in Gen (NS, n=1,754). According to the curated species,
910 True Positives can be defined as genomes with SS > max (SS_{NS}) distribution, whereas True Negatives are
911 those with SS < min(SS_{Suli}). B) Assignment of the 192 genomes in Sur to ecological categories based on
912 literature reports. C) Distribution of SS for different S-metabolic guilds, and the genomes in Sur. D) ROC
913 curve with Area Under the Curve (AUC) indicated together with thresholds for some False Positive Rates
914 (FPR).

915 **Figure 6.** Distribution of Sulfur Score (SS) in the metagenomic dataset Met. A) Geo-localized metagenomes
916 sampled around the globe are colored according to their *SS* values. The following cut-off values correspond
917 to the 95th percentiles of seven Mean Size Length classes (30, 60, 100, 150, 200, 250 and 300 aa): 7.66,
918 9.70, 8.81, 8.51, 8.18, 8.98 and 7.61, respectively. Circles with thick blue border indicate metagenomes with
919 SS ≥ the 95th percentile. B) Distribution of SS values observed in 935 metagenomes classified in terms of
920 features (X-axis) and colored according to their particular habitats Features are sorted according to their
921 median SS values. ccc: metagenomes from Cuatro Cienegas, Coahuila, Mexico. Green lines indicate the
922 lowest and largest 95th percentiles observed across MSL classes.

Figure 7. Metabolic completeness of the metabolic pathways described in Table 1. A) Boxplot distribution of the pathway completeness in genomic and B) metagenomic datasets. C) Linear regression models of the Sulfur Score (*SS*) and the mean completeness in Gen and D) Met dataset. E) Heatmap showing the metabolic completeness of the following genomes: *Desulfovibrio vulgaris DP4* (*SS*=11,442), *Ammonifex degensii KC4* (SS=12.508); *Pelodictyon phaeoclathratiforme BU-1* (*SS*=11,836); *Thiobacillus denitrificans ATCC 25259* (SS=11,61); PSB: *Allochromatium vinosum DSM 180* (SS=10. 737); Sur Methanosarcina barkeri MS(SS= 5,93); *Sulfolobus acidocaldarius DSM 639* (SS=5,457); *Synechococcus sp. JA-2-3Ba 2-13* (SS=3,704); *Hyphomicrobium denitrificans 1NES1* (SS= 3,236); *Ruegeria pomeroyi DSS-3 (SS=2,707)*; *Enterococcus durans* (SS=-0,194); *Micrococcus luteus NCTC_2665* (SS=-3,588). F) Heatmap showing the metabolic completeness of the metagenomes with the following MG-RAST ids and corresponding scores: 4489656.3 (SS=-2,649); 4440320.3(SS=0,1); 4441663.3(SS=9,986); 4510168.3 (SS=7,781) ; 4493725.3 (SS=9,547) ; 4461840.3 (SS=8,813); 4441599.3(SS=9,274); 4451035.3(SS=9,918); 4525341.3(SS=9,287); 4489328.3(SS=4,958); 4478222.3(SS=4,88).The color codes at the top of the heatmap correspond to different environments. For a more detailed description of each metagenome see Table S8.

Additional files -Supplementary Information

The supplementary pdf file contains the following information:

Supplementary figure S1. Histogram distribution of the Mean Size Length of metagenomes in Met and the input sulfur proteins.

Supplementary figure S2. Visualization of the Pfam domains mapped onto KEGG metabolic pathways

Supplementary figure S3. Comparison of clustering methods of the 112 Pfam entropies using script *plot_cluster_comparison.py*

Supplementary figure S4. Clustering comparison between Birch and Ward clustering methods to stand out the Pfam entropies with high *H'* and low std using the script

Supplementary figure S5. **.** Distribution of entropy values of 112 Pfam domains inferred from random-sampled and Suli genomes.

Supplementary figure S6. Comparison of Sulfur Scores (*SS*) with data obtained three years ago (2014), with the current data described in the article.

Supplementary table S4. Informative Pfam's with high *H'* and high std (not used as molecular marker genes) in metagenomic fragmented data.

Supplementary table S5. Percentile distribution of the 112 Pfam entropies in the random test

Supplementary table S10. Statistics of *SS* computed on genomic (Gen, real sequences) and metagenomic (Met, with increasing Mean Size Length, from 30 to 300aa) datasets

In separated excel files the following Supplementary tables are also provided:

Supplementary table S1: Table S1. Comprehensive list of the taxonomic representatives of sulfur cycle including Sulfur list or 'Suli' containing 161-curated genomes used as input for the pipeline

960 **Supplementary table S2**. Sucy database containing the identifiers of the Sulfur proteins and their
961 corresponding annotations derived from Interproscan and manual curation.

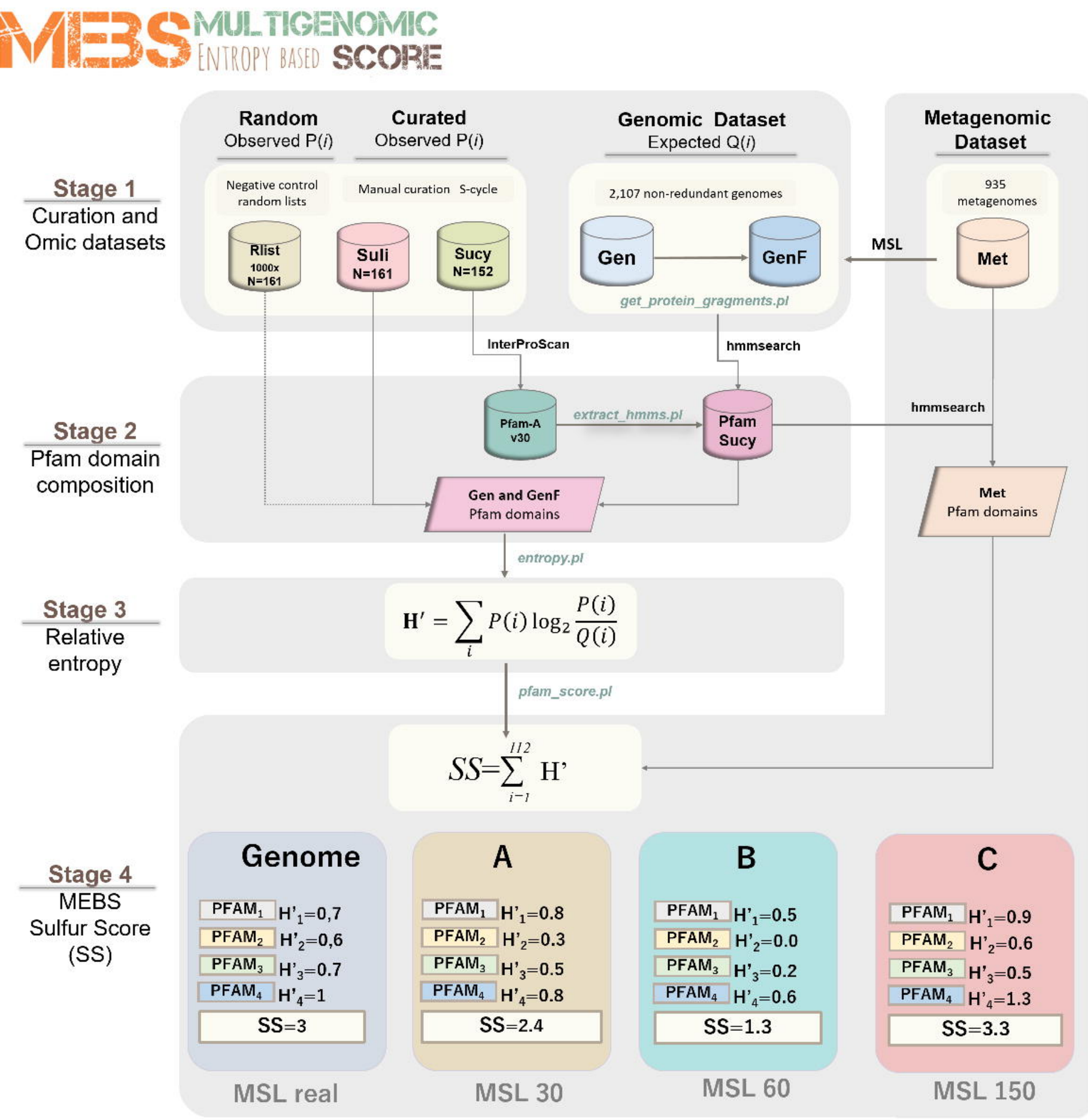
962 **Supplementary table S3.** Sulfur Pfam domains (Pfam-Sucy), and their corresponding mapping into KEGG (KO
963 number), and the manual assignation into sulfur metabolic pathways

964 **Supplementary table S6**. Gen dataset containing their corresponding SS and taxonomy assignment.

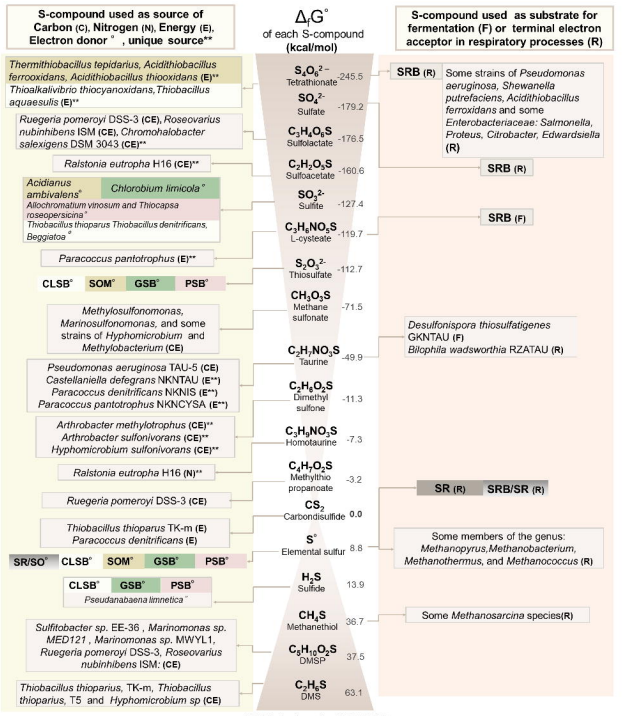
965 **Supplementary table S7.** Manual annotation of Sulfur unconsidered or related microorganisms (Sur) with
966 SS>4 Supplementary table S8

967 **Supplementary table S8.** Met dataset with their corresponding *SS* values and metatada.

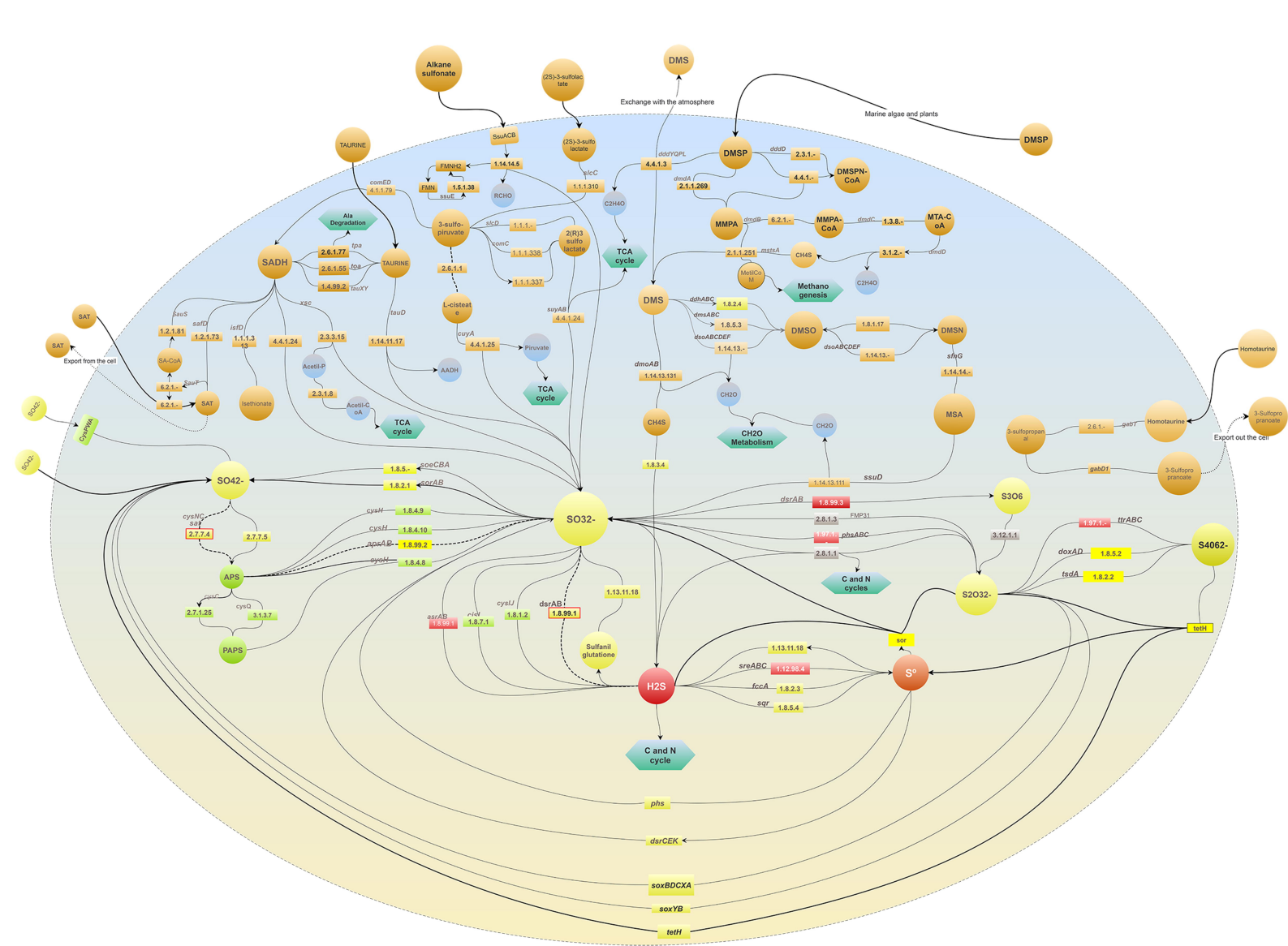
968 **Supplementary table S9**. Manually annotated high scoring metagenomes along with their ecological
969 capabilities in terms of sulfur cycle
970
971 **Supplementary table S11.** Metabolic completeness in Gen dataset for each of the 28 metabolic pathways of the S-
972 cycle described in Table 1. (Pathway 29 contains the proposed marker genes)
973
974 **Supplementary table S12.** Metabolic completeness in Men dataset for each of the 28 metabolic pathways of the S-
975 cycle described in Table 1. (Pathway 29 contains the proposed marker genes)
976
977 **Supplementary table S13**. Frequency and description of the most complete genomes in terms of S-cycle
978 metabolic pathways
979
980
981

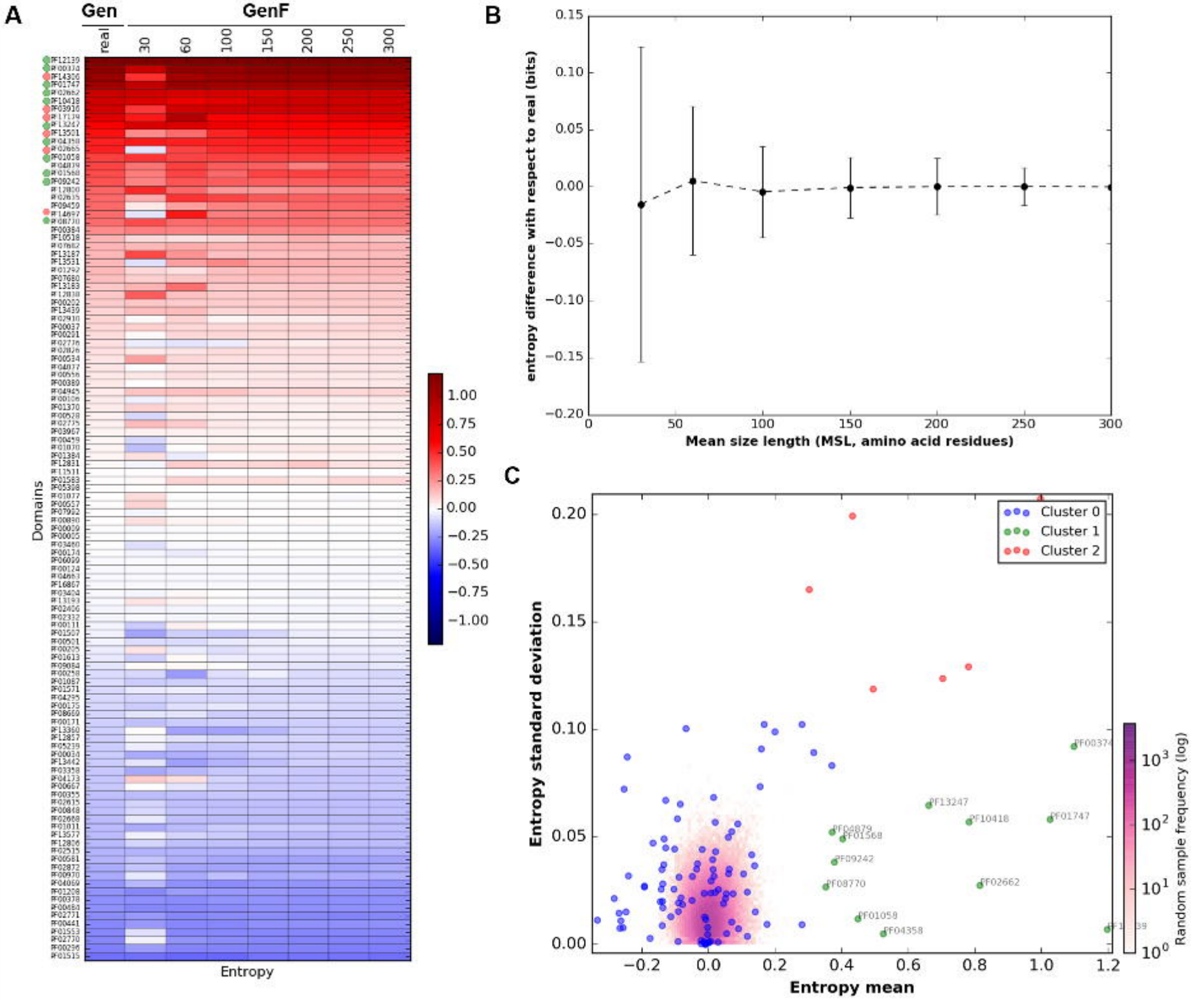

MEBS
MULTIGENOMIC
ENTROPY BASED SCORE
Stage 1
Curation and Omic datasets
Random
Observed P(i)
Curated
Observed P(i)
Genomic Dataset
Expected Q(i)
Metagenomic Dataset
Negative control random lists
Manual curation S-cycle
2,107 non-redundant genomes
935 metagenomes
Rlist 1000x N=161
Suli N=161
Sucy N=152
Gen
GenF
Met
MSL
get_protein_gragments.pl
InterProScan
hmmsearch
Stage 2
Pfam domain composition
Pfam-A v30
extract_hmms.pl
Pfam Sucy
hmmsearch
Gen and GenF Pfam domains
Met Pfam domains
entropy.pl
Stage 3
Relative entropy
$\mathbf{H}' = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$
pfam_score.pl
$SS = \sum_{i=1}^{112} \mathrm{H}'$
Stage 4
MEBS Sulfur Score (SS)
Genome
PFAM1 H'1=0,7
PFAM2 H'2=0,6
PFAM3 H'3=0.7
PFAM4 H'4=1
SS=3
MSL real
A
PFAM1 H'1=0.8
PFAM2 H'2=0.3
PFAM3 H'3=0.5
PFAM4 H'4=0.8
SS=2.4
MSL 30
B
PFAM1 H'1=0.5
PFAM2 H'2=0.0
PFAM3 H'3=0.2
PFAM4 H'4=0.6
SS=1.3
MSL 60
C
PFAM1 H'1=0.9
PFAM2 H'2=0.6
PFAM3 H'3=0.5
PFAM4 H'4=1.3
SS=3.3
MSL 150

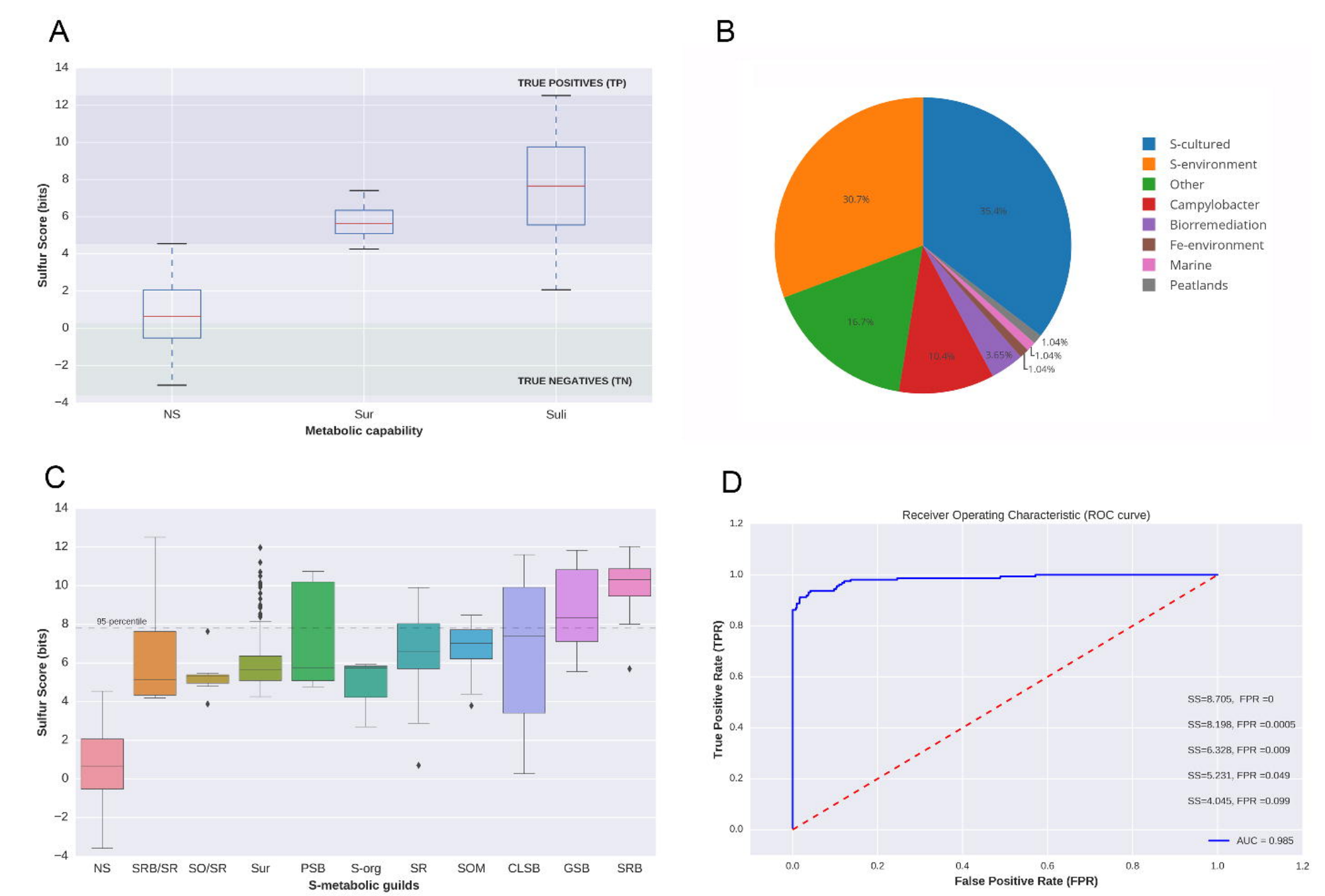
Biogeochemical Sulfur cycle at global scale

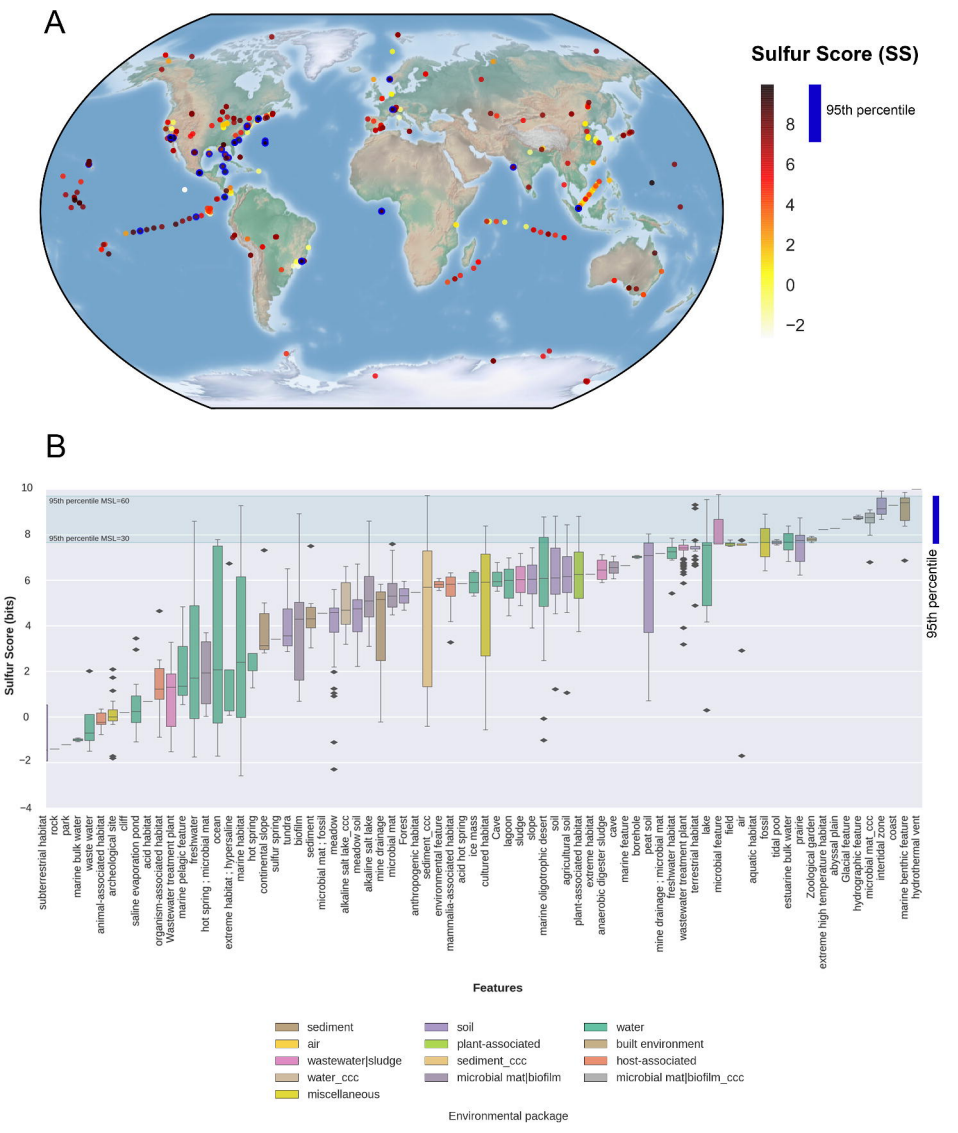


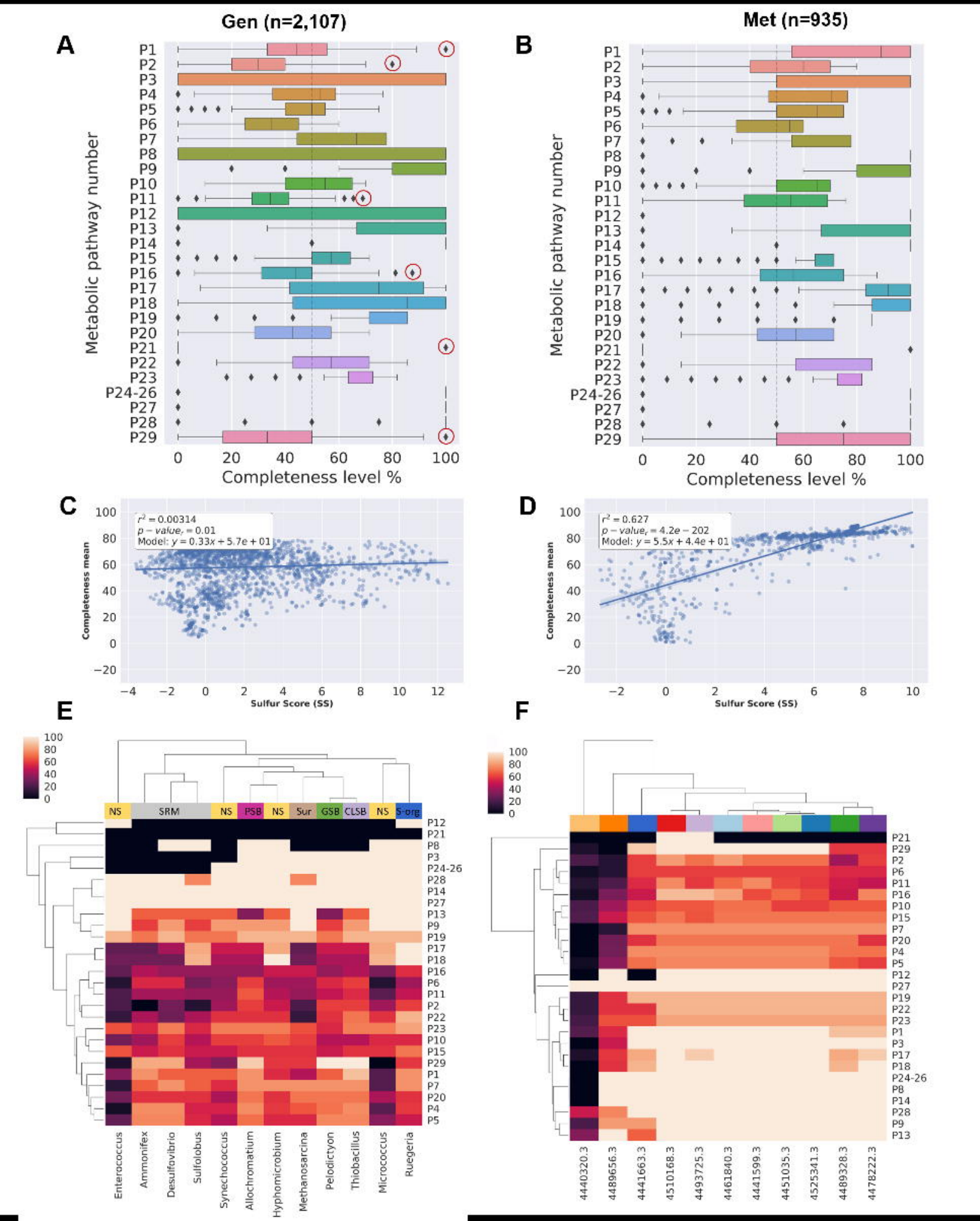
CLSB	**C**olor-less **S**ulfur **B**acteria: 24 genera (*i.e Beggiatoa, Thiomargarita, Thiobacillus*)
SOM	**S**ulfur **O**xidizing **M**icroorganims: 12 genera (*i.e Thermithiobacillus, Acidithiobacillus*)
GSB	**G**reen **S**ulfur **B**acteria: 9 genera (*i.e Clorobaculum, Chloroflexus, Chlorobium*)
PSB	**P**urple **S**ulfur **B**acteria: 25 genera (*i.e Ectothiorhodospira, Cromatium*)
SRB	**S**ulfate **R**educing **B**acteria: 40 genera (*i.e Desulfovibrio, Desulfotomaculum, Desulfotignum*)
SR	Elemental-**S**ulfur **R**educing microorganisms: 19 genera (*i.e Sulfospirillum, Desulfurella, Thermoproteus*)
ESR/SO	Elemental-**S**ulfur **R**educing and **S**ulfur **O**xidizing microorganisms: includes 4 genera: *Sulfolobus, Acidianus, Aquifex, Thermoplasma*
SRB/ESR	**S**ulfate **R**educing **B**acteria and Elemental **S**ulfur **R**educing microorganism : 42 genera (i. e *Thermovirga, Pyrolobus*)


Alkane sulfonate
(2S)-3-sulfolactate
DMS
Exchange with the atmosphere
Marine algae and plants
DMSP
TAURINE
SsuACB
1.14.14.5
FMNH2
FMN
1.5.1.38
ssuE
RCHO
(2S)-3-sulfolactate
slcC
1.1.1.310
C2H4O
TCA cycle
dddYQPL
4.4.1.3
DMSP
dddD
2.3.1.-
4.4.1.-
DMSPN-CoA
dmdA
2.1.1.269
MMPA
dmdB
6.2.1.-
MMPA-CoA
dmdC
1.3.8.-
MTA-C oA
comED
4.1.1.79
Ala Degradation
3-sulfo-piruvate
slcD
1.1.1.-
comC
1.1.1.338
2(R)3 sulfo lactate
1.1.1.337
2.6.1.1
L-cisteate
cuyA
4.4.1.25
Piruvate
TCA cycle
suyAB
4.4.1.24
SADH
tpa
2.6.1.77
2.6.1.55
hpa
1.4.99.2
tauXY
TAURINE
xsc
tauD
1.14.11.17
AADH
2.3.3.15
Acetil-P
2.3.1.8
Acetil-C oA
TCA cycle
sauS
1.2.1.81
safD
1.2.1.73
isfD
1.1.1.313
4.4.1.24
SA-CoA
6.2.1.-
sauT
SAT
Isethionate
SAT
SAT
Export from the cell
SO42-
CysPWA
SO42-
SO42-
1.8.5.-
soeCBA
1.8.2.1
sorAB
cysNC
sat
2.7.7.4
2.7.7.5
APS
cysH
1.8.4.9
cysH
1.8.4.10
aprAB
1.8.99.2
cysH
1.8.4.8
cysH
2.7.1.25
cysQ
3.1.3.7
PAPS
SO32-
dsrAB
1.8.99.3
2.8.1.3
FMP31
1.97.1.-
phsABC
2.8.1.1
C and N cycles
1.14.13.111
ssuD
MSA
dmoAB
1.14.13.131
CH2O
CH2O Metabolism
CH2O
CH4S
1.8.3.4
ddhABC
1.8.2.4
dmsABC
1.8.5.3
dsoABCDEF
1.14.13.-
DMSO
1.8.1.17
dsoABCDEF
1.14.13.-
DMSN
sfnG
1.14.14.-
DMS
Methano genesis
MetiCoM
C2H4O
2.1.1.251
mtsteA
CH4S
3.1.2.-
dmdD
3-sulfopropanal
2.6.1.-
gabT
Homotaurine
gabD1
3-Sulfopropranoate
Homotaurine
3-Sulfopropranoate
Export out the cell
S3O6
3.12.1.1
S2O32-
1.97.1.-
ttrABC
doxAD
1.8.5.2
tsdA
1.8.2.2
S4062-
tetH
sor
S°
1.13.11.18
sreABC
1.12.98.4
fccA
1.8.2.3
sqr
1.8.5.4
H2S
Sulfanil glutatione
asrAB
1.8.99.1
rlxA
1.8.7.1
cysIJ
1.8.1.2
dsrAB
1.8.99.1
1.13.11.18
C and N cycle
phs
dsrCEK
soxBDCXA
soxYB
tetH


A
Gen
GenF
real
30
60
100
150
200
250
300
Domains
Entropy
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
B
entropy difference with respect to real (bits)
Mean size length (MSL, amino acid residues)
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.20
0
50
100
150
200
250
300
C
Cluster 0
Cluster 1
Cluster 2
Entropy standard deviation
Entropy mean
Random sample frequency (log)
0.20
0.15
0.10
0.05
0.00
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
10^3
10^2
10^1
10^0
PF00374
PF13247
PF10418
PF01747
PF04879
PF01568
PF09242
PF08770
PF02662
PF01059
PF04358


A
TRUE POSITIVES (TP)
TRUE NEGATIVES (TN)
Sulfur Score (bits)
Metabolic capability
NS
Sur
Suli
B
S-cultured
S-environment
Other
Campylobacter
Biorremediation
Fe-environment
Marine
Peatlands
30.7%
35.4%
16.7%
10.4%
3.65%
1.04%
1.04%
1.04%
C
95-percentile
Sulfur Score (bits)
S-metabolic guilds
NS
SRB/SR
SO/SR
Sur
PSB
S-org
SR
SOM
CLSB
GSB
SRB
D
Receiver Operating Characteristic (ROC curve)
True Positive Rate (TPR)
False Positive Rate (FPR)
SS=8.705, FPR =0
SS=8.198, FPR =0.0005
SS=6.328, FPR =0.009
SS=5.231, FPR =0.049
SS=4.045, FPR =0.099
AUC = 0.985


A
Sulfur Score (SS)
95th percentile
8
6
4
2
0
−2
B
95th percentile MSL=60
95th percentile MSL=30
95th percentile
Sulfur Score (bits)
10
8
6
4
2
0
−2
−4
subterrestrial habitat
rock
park
marine bulk water
waste water
animal-associated habitat
archeological site
cliff
saline evaporation pond
acid habitat
organism-associated habitat
Wastewater treatment plant
marine pelagic feature
freshwater
hot spring : microbial mat
ocean
extreme habitat : hypersaline
marine habitat
hot spring
continental slope
sulfur spring
tundra
biofilm
sediment
microbial mat : biofilm
meadow
alkaline salt lake_ccc
meadow soil
alkaline salt lake
mine drainage
microbial mat
Forest
anthropogenic habitat
sediment_ccc
environmental feature
mammalia-associated habitat
acid hot spring
ice mass
cultured habitat
Cave
lagoon
sludge
Ocean
marine oligotrophic desert
soil
agricultural soil
plant-associated habitat
extreme habitat
anaerobic digester sludge
cave
marine feature
borehole
peat soil
mine drainage : microbial mat
microbial mat_ccc
wastewater treatment plant
terrestrial habitat
lake
microbial feature
field
air
aquatic habitat
fossil
tidal pool
estuarine bulk water
prairie
Zoological garden
extreme high temperature habitat
abyssal plain
Glacial feature
hydrographic feature
microbial mat_ccc
intertidal zone
coast
marine benthic feature
hydrothermal vent
Features
sediment
air
wastewater|sludge
water_ccc
miscellaneous
soil
plant-associated
sediment_ccc
microbial mat|biofilm
water
built environment
host-associated
microbial mat|biofilm_ccc
Environmental package


A
Gen (n=2,107)
Metabolic pathway number
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
P19
P20
P21
P22
P23
P24-26
P27
P28
P29
0
20
40
60
80
100
Completeness level %
B
Met (n=935)
Metabolic pathway number
Completeness level %
C
$r^2 = 0.00314$
$p-value_r = 0.01$
Model: $y = 0.33x + 5.7e+01$
Completeness mean
Sulfur Score (SS)
D
$r^2 = 0.627$
$p-value_r = 4.2e-202$
Model: $y = 5.5x + 4.4e+01$
Completeness mean
Sulfur Score (SS)
E
NS
SRM
NS
PSB
NS
Sur
GSB
CLSB
NS
S-org
Enterococcus
Ammonifex
Desulfovibrio
Sulfolobus
Synechococcus
Allochromatium
Hyphomicrobium
Methanosarcina
Pelodictyon
Thiobacillus
Micrococcus
Ruegeria
F
4440320.3
4489656.3
4441663.3
4510168.3
4493725.3
4461840.3
4441599.3
4451035.3
4525341.3
4489328.3
4478222.3