Title

Diversity at single nucleotide to pangenome scales among sulfur cycling bacteria in salt marshes

Authors

Sherlynette Pérez Castro*[1], Elena L. Peredo*[1], Olivia U. Mason[2], Joseph Vineis[3], Jennifer Bowen[3], Behzad Mortazavi[4], Anakha Ganesh[5], S. Emil Ruff[1,5], Blair G. Paul[5], Anne E. Giblin[1], Zoe G. Cardon[1]

* Corresponding authors. These authors contributed equally to leading this research.

Author Affiliations

[1]The Ecosystems Center, Marine Biological Laboratory, Woods Hole, MA, USA

[2]Department of Earth, Ocean and Atmospheric Science, Florida State University, Tallahassee, FL, USA

[3]Department of Marine and Environmental Sciences, Marine Science Center, Northeastern University, Nahant, MA, USA

[4]Department of Biological Sciences, University of Alabama, Tuscaloosa, AL, USA

[5]Bay Paul Center, Marine Biological Laboratory, Woods Hole, MA, USA

Corresponding authors contact information

sperezcastro@mbl.edu, elperedo@mbl.edu

Competing interests: The authors declare no competing interests.

Abstract

Sulfur-oxidizing and sulfate-reducing bacteria in salt marsh sediments are major controllers of ecosystem-scale carbon cycling. Cross-site comparisons of S-cycling communities are difficult given the rampant uncultured microbial diversity in sediment, yet comparisons are essential for revealing biogeographic, phylogenetic and functionally significant variation. Here, we use deep shotgun metagenomic sequencing data to construct and compare metagenome-assembled genomes (MAGs) of sulfur-cycling bacteria from Massachusetts and Alabama salt marshes that contrast in seasonality and sediment organic matter content. Samples were collected from sediments under *Sporobolus alterniflorus* and *Sporobolus pumilus* in separate MA vegetation zones, and under *Sporobolus alterniflorus* and *Juncus roemerianus* co-rooted in AL marsh. We grouped metagenomic data by plant species and site and identified 38 MAGs that included pathways for dissimilatory sulfate reduction or sulfide oxidation. Phylogenetic analyses indicated that 30 of the 38 were affiliated with uncultivated lineages. Read-mapping to MAGs showed significant differentiation of AL and MA samples, differentiation of samples taken in *S. alterniflorus* and *S. pumilus* vegetation zones in MA, but no differentiation of samples taken under *S. alterniflorus* and *J. roemerianus* that were rooted together in AL marsh. Pangenomic analyses of eight ubiquitous MAGs also detected site- and vegetation- specific genomic features, including varied sulfur-cycling operons, carbon fixation pathways, fixed single nucleotide variants, and active diversity-generating retroelements. This genetic diversity, detected at multiple scales even within uncultured groups, suggests evolutionary relationships affected by distance and local environment, and demonstrates differential microbial capacities for sulfur and carbon cycling in salt marsh sediments.

Importance

Salt marshes are known for their significant carbon storage capacity, and sulfur cycling is closely linked with the ecosystem-scale carbon cycling in these ecosystems. Sulfate reducers are the major decomposers in salt marsh systems, and sulfur-oxidizing bacteria remove sulfide, a toxic byproduct of sulfate reduction, supporting the productivity of marsh plants. To date, the complexity of coastal environments, heterogeneity of the rhizosphere, high microbial diversity and uncultured majority hindered our understanding of the genomic diversity of sulfur-cycling microbes in salt marshes. Here we use comparative genomics to overcome these challenges and provide an in-depth characterization of microbial diversity in salt marshes. We characterize sulfur-cycling communities across distinct sites and plant species and uncover extensive genomic diversity at the taxon level and specific genomic features present in MAGs affiliated with sulfur-cycling uncultivated lineages. Our work provides insights into the partnerships in salt marshes and a roadmap for multiscale analyses of diversity in complex biological systems.

Keywords: sulfur-oxidizing bacteria, sulfate-reducing bacteria, site-specific genetic diversity, DGR, SNV, pangenomics.

Introduction

Sulfur-cycling bacteria are central to the function of salt marsh ecosystems worldwide (1). While sulfate reducers catalyze the dominant decomposition pathway in salt marshes, sulfide is produced as a metabolic byproduct that can be toxic to plant roots (2–4). Microorganisms that oxidize sulfide (5), can detoxify rhizosphere sediments for plants. Thus, the metabolism of S-cycling microbes is a major contributor to high rates of marsh plant productivity (6–9). The tremendous diversity of sulfur-cycling bacteria has made cross-site comparisons of their broad

genomic content difficult (10). Studying the genomic diversity of S-cycling microbes requires deeply sequenced metagenomic data to reveal gene function and phylogenetic relationships (11).

Salt marshes distributed along the eastern coast of the United States are characterized by dominant vegetation regulated by climatic, tidal and edaphic factors. Salt marsh plants are usually distributed along an elevation gradient depending on their adaptability and tolerance to reduced sediments and saline conditions. Typical species located from low to high marsh include *Sporobolus alterniflorus* (formerly *Spartina alterniflora*), *Sporobolus pumilus* (formerly *Spartina patens*), and *Juncus roemerianus*. *S. alterniflorus* is most tolerant of the frequent flooding and high salinity characterizing the intertidal zone and low elevations (12, 13). In New England, *S. pumilus* is dominant in the less frequently flooded high marsh zone (14). In Alabama, *J. roemerianus*, though often in high marsh, can mix with *S. alterniflorus* (15).

Over the past several decades, examination of microbial communities by sequencing 16S rRNA genes has identified a wide range of sulfur-cycling bacterial lineages within marsh sediments. Sulfate reducers belong to bacterial phyla such as *Desulfobacterota* (formerly *Deltaproteobacteria*), *Acidobacteriota*, and *Bacteroidota* (16, 17), while sulfur oxidizers are often found in the phyla *Proteobacteria*, *Bacteroidota*, and *Campylobacterota* (5, 18, 19), including orders *Chlorobiales*, *Chromatiales*, *Rhizobiales*, and *Rhodobacterales* (5). In a biogeographic survey of sulfate reducers in multiple East Coast salt marsh sediments, (20) found that the distribution of 16S rRNA genes varied with the environment and geographic distance, but the sulfate reductase gene did not. New metagenomic approaches are now emerging as a powerful tool to detect heterogeneity and compare sequences in multiple genes, even in highly diverse microbial populations (21, 22).

Here, we focus on the sulfur-cycling microbial communities inhabiting two contrasting salt marshes differing in type of sediment (rich in organic matter/fine), latitude (North/South), and vegetation (*S. alterniflorus* and *S. pumilus*/ *S. alterniflorus*, *J. roemerianus*). We explored the occurence of diverse metagenome-assembled genomes (MAGs) encoding dissimilatory sulfate reduction or sulfide oxidation genes (1) across sites and under different plant taxa and characterized their genomic diversity from single nucleotide to pangenome scales (Fig. 1). We investigated shared features and site- and vegetation-specific genetic diversity, providing insights into biogeographic patterns and functional diversity.

Results

Dataset comparison and identification of MAGs

In our study, deep metagenomic datasets allowed us to characterize sulfur cycling microbial diversity across distinct environments and plant species. For the reconstruction of "primary MAGs" (see Methods Section), metagenomic samples were grouped according to geographical site, Alabama (AL) and Massachusetts (MA) and dominant vegetation on the collection site, *S. alterniflorus* (SA), *S. pumilus* (SP), and *J. roemerianus* (JR) (Table S1). The four co-assemblies yielded 251,034 (ALJR), 249,343 (ALSA), 70,380 (MASA), and 100,526 (MASP) contigs, with similar N50 values (7 Kb) (Table S2) and taxonomic profiles (Fig. S1). Binning resulted in 118 medium-quality MAGs (>90% completion, <5% contamination), 38 were identified by their gene content as sulfur cycling bacteria (Fig. 2, Supplemental Data 1-2), and they belong to bacterial groups traditionally associated with salt marshes including *Acidobacteriota*, *Bacteroidota*, *Desulfobacterota*, *Gemmatimonadota*, and *Proteobacteria* (Fig. 3, Table S3). Seventeen of the 38 S-cycling MAGs were sulfate reducers (including genes *sat*,

*apr*AB, and *dsr*AB) belonging to uncultured lineages (labeled by alphanumeric identifiers), with five identifiable only at the family level (Fig. 3, Table S3). All but one of these SRB MAGs were assembled from AL samples. The sulfide detoxifying enzyme *sqr* (sulfide:quinone oxidoreductase) was found in 6 of the SRB MAGs (Fig 2). Twenty-one of the 38 MAGs were sulfur oxidizers with genes encoding the truncated thiosulfate oxidation system (Fig. 2); approximately equal numbers of MAGs were assembled from AL and MA samples. All were in the phylum *Proteobacteria* (Table S3). Two MAGs were affiliated with the uncultivated lineage UBA6429, which was the only lineage assembled in both Massachusetts and Alabama samples. All but three SOX MAGs harbored *sat*, *apr*AB, and *dsr*AB, likely operating in reverse, oxidizing sulfide to sulfate (Fig. 2). The majority of MAGs encoded *sqr*, but three lacked *sqr* and encoded *soxCD* genes (Fig. 2). Genes for sulfite (*soeABC*) and sulfide (*fccAB*) dehydrogenase were variably represented (Fig. 2).

Distribution of sequences from the 38 MAGs across samples

To quantify and compare MAG sequence distribution across samples, we used read-mapping (Fig. 4A). On average, more reads mapped to contigs from sulfur-oxidizing "primary" MAGs (see Methods section) across all samples (average CPM 11 for SOX, vs. average CPM 3 for SRB). Also, we observed a geographical pattern related to the percentage of the MAG successfully recruiting reads. When AL metagenomic samples were mapped to AL-derived MAGs, up to 99% of contigs had CPM values different from 0. For samples collected in MA, this was reduced to 92% of contigs in samples from sediments dominated by *S. alterniflorus* and further reduced to ~75% samples collected by *S. patens*. We did not observe such a pronounced pattern in MA-derived MAGs as values ranged from 99% to 92-95% (Supplemental Data 3). For

each primary MAG, patterns of read recruitment to individual contigs can be found in the supplementary results (Fig. S3-S5).

Various site- and vegetation-specific patterns emerged. For example, the presence of MASA5 (GCA-1735895 sp.) was only detected in MA samples and was enriched in sediments inhabited by *S. alterniflorus* (Fig. 4A). Among sulfate reducers, ALSA176 (*Acidobacteriota* Fen-336 family) was the only MAG which recruited reads in all samples (average 6 CPM, Fig. 4A), and closely-related ALSA143 (also in the Fen-336 family) recruited reads from all but one sample (Fig. 4A). Sequences from BM002 MAGs (related to *Syntrophobacteria*) were poorly represented in MA samples vegetated with *S. pumilus* (Fig. 4A).

Such differences in MAG abundance were revealed in patterns by NMDS analysis based on the read recruitment values (CPMs). This analysis uncovered separation by site (MA or AL; Fig. 4A) and vegetation within site (Fig. 4B), by the month that the sample was collected (MA only, Fig. 4C) and by the depth of sediment (AL-only, Fig. 4D). PerMANOVA post-hoc comparisons detected significant differences between MASP and MASA samples when compared to each other and to ALJR ($p = 0.01$) and ALSA ($p = 0.01$, Fig. 4A), but no significant differences were found between ALJR and ALSA samples ($p = 0.15$, Fig. 4A). No significant patterns were detected by sampling month in MA (Fig. 4B). In AL (Fig. 4D), 0-2 cm depth samples were significantly different from 5-7 cm and 0-10 cm samples ($p = 0.045$).

The same read recruitment approach was applied to more sparsely sampled metagenomic data from unvegetated creek bed sediment samples, revealing for several primary MAGs clear variation in read recruitment as a function of salinity in two tidal creeks at PIE-LTER (Fig. S6, Supplemental Data 4).

Contig-based analysis of genomic variability

Genomic coverage and genetic diversity in eight primary MAGs (identified by arrows, Fig. 4A) that occurred in all sampling locations (Fig. S7) were analyzed further using the Anvi'o profiling pipeline (Supplemental Data 5). For all MAGs (Fig. 5A), we observed a relative uniformity in GC content (inner ring) and mean read coverage (colored rings) across the contigs forming each bin. Overall, read recruitment to contigs was influenced by, but was not solely governed by sample sequencing depth (see grey histograms Fig. 5A, Fig. S8-S15). While sequence quality, contig length, assembly method, genetic diversity, and experimental design can influence read recruitment, here we observe an effect on the geographical origin of the sample. Each MAG was better represented in the pool of samples originally used for its assembly and was similarly abundant in the samples of a given site (Fig. 5A). This was not the case for the sediments inhabited by *S. pumilus*, where a MAG could be highly abundant just in one sample. This seems to suggest that in this area microorganisms might have a greater spatial heterogeneity than those areas inhabited by *S. alterniflorus* and *J. roemerianus* (Table S4).

Patterns of recruitment varied across primary MAGs. For example, UBA6429 MASA10 contained an 80.4 Kbp region, encoding 86 genes, that was only detected in samples collected in Massachusetts under *S. alterniflorus* (Fig. 5A, arrow). This MASA-exclusive region encodes duplicated copies of genes involved in the biosynthesis of molybdenum cofactor (Moco), and it was rich in transport genes (Supplemental Data 6). Moco is essential for the activity of molybdoenzymes involved in the metabolism of sulfur compounds such as the conversion of toxic sulfite to sulfate. *Salaquimonas* sp. MASP16 included an 18.5 Kb region that was only present in the samples collected in Massachusetts, in either *S. alterniflorus* or *S. pumilus* vegetation zones (Fig. 5A, arrow). This region encoded 22 genes, including a second set of genes

of the phospholipid transporter system MlaFEDB and the genes *arcC* and *arcA* involved in arginine biosynthesis (Supplemental Data 6).

We investigated single nucleotide variants (SNVs) by quantifying the divergence of recruited reads from the primary MAGs (Fig. 5A, Supplemental Data 5). For each of the four AL-derived MAGs, and for UBA6429 MASA10, the number of SNVs (black histograms) was correlated with the number of mapped reads (grey histograms) across all 24 samples (Fig. 5A, Fig. S7, and Supplemental Data 5). But for primary MAG MASA4, AL samples had far more SNVs per mapped read than did MA samples (Fig. 5A). Similar global variation was observed in samples MASP15 and MASP16, all AL samples, and MA samples under *S. alterniflorus*, wherein all had far more SNVs per mapped read than the MA *S. pumilus* samples (Supplemental Data 5).

Gene-based analysis of sample-specific reassembled genomes

Using the same 8 primary MAGs as references, we reassembled sample-specific MAGs where possible (Table S5). Completeness varied widely and was highest for samples that had originally been coalesced to generate the primary MAG co-assembly. We used Anvi'o pangenomic analysis (Fig. 5B, Fig. S8-S15) to annotate the genomes and identify clusters of homologous genes. Our results indicate that the geographical origin of the primary MAG (AL or MA) determined the reassembly success. In ALJR15, ALSA124, and ALSA174, initially identified in co-assemblies of AL samples, we observed a high degree of completeness among Alabama-reassembled MAGs (55-92%) while we observed limited reassembly success among the MA samples. An extreme case was MAG ALSA174 where Massachusetts-derived MAGs completeness was calculated at 1-3%. For all the MA MAGs and a single AL one (ALJR36), we successfully reassembled sample-specific MAGs, whether derived from MA (55-65%) or AL

(75%) samples (Supplemental Data 7). For UBA6429 MASA10 and ALJR36, sample-specific reassemblies from all sites shared a very large proportion of the gene clusters found in the original primary MAG. Single-copy core genes were readily identified as 42% and 33%, respectively, of the total number of genes (Fig. 5B, Fig. S8 and S12). For the other 6 MAGs, SCGs were sparse or absent (Fig. 5B, Fig. S9-S15).

Hot spots of variability in sample-specific reassembled MAGs

The close phylogenetic relationship of UBA6429 ALJR36 and MASA10 (Fig. 3, ANI: 99.43%) and coexistence across sites (Fig. 4, 5B) provided a unique opportunity for the study of genetic diversity and microevolution that might be linked to metabolic capabilities. Previous analyses (described above) generated SNV counts for each of the 24 metagenomic samples (Fig. 5A, grey histograms). To determine the average number of SNVs per kilobase, we coupled the analysis of SNVs in the sample contigs with gene annotation data (black bars, Fig. 6A). We show that SNVs remained stable across the MAGs, with the exception of a handful of ORFs with outlier values (Fig. 6A, arrows). Notably high variability was observed within prophages and transposons (COG5361), rRNA, tetratricopeptide repeats responsible for protein-protein interactions or assembly of multiprotein complexes (COG0457), and the large subunit of a Rubisco-like protein (COG1850). As observed in the previous SNV analysis (Fig. 5A, Supplemental Data 5), SNVs were more abundant across the full genomes of Alabama-specific MAGs that had more mapped reads (Fig. 6A, dense black bars across full genomes.)

Pangenomic analysis

Sample-specific genomes reassembled using as a guide the closely related UBA6429 MASA10 and ALRJ36 primary MAGs were grouped by site (AL or MA) for pangenomic analysis within Anvi'o. Three MA *S. pumilus* samples were excluded because the completeness

of the reassembled genomes was <50% (Supplemental Data 7). All samples were compared in terms of presence/absence or orthologous genes, revealing a shared subset of the genomes (Fig. 6C). Of the 2,719 genes identified in the pangenomic analysis, 2,375 were present in the MASA10 group and 2,355 in the ALJR36 group (Fig. 6C). Functional characterization was relatively complete with 81.5% of the genes identified using COG20, and 82.5% with Pfam. Moreover, 58.3% of the genes were assigned Kegg numbers (grey outer rings, Fig. 6C). A repertoire of 503 SCG were identified across the two groups of metagenomes (purple ring, Fig. 6C). The order of ORFs around the rings in Fig. 6C (which is not related to the order in genomes) visually emphasizes genes found only in re-assembled MAGs from MA samples taken under *S. alterniflorus* (labeled MASA), and genes found only in the family of sample-specific MAGs reassembled with primary MAG MASA10 as a guide (labeled MASA10-specific genes).

Despite the large similarity among both groups of genomes, 284 genes were present only in ALJR36 reassembled metagenomes and 292 genes in the MASA10 group. Within the MASA10 group, 80 genes were exclusively identified in samples collected in Massachusetts in intertidal sediments inhabited by *S. alterniflorus* (Fig. 6C, sector at top of anviogram). Functional enrichment analysis indicated that 11 COG categories were statistically enriched in MAGs from the group ALJR36 and 8 in MAGs from MASA10 (Table S6, Supplemental Data 8). These COG categories include inorganic ion, carbohydrate, and amino acid transport. KEGG enrichment analysis indicated different metabolic capabilities between the MASA10 and ALJR36 groups of metagenomes. Notably, phosphate acetyltransferase-acetate kinase and the reductive citrate cycle (Arnon-Buchanan cycle) pathways were enriched in the ALJR36 group (Table S7, Fig. S17). This was caused by the presence of the genes encoding the enzymes acetate kinase (*AckA*) and aconitate hydratase (*AcnB*) in the ALJR36 group, and the lack of the same

genes in the MASA10 group. The MASA10 group was significantly enriched for sulfur-related KEGG modules. Particularly distinguishing are that the ALJR36 group MAGs are missing complete genes encoding subunit B of the sulfohydrogenase *soe*ABC complex, and several key genes (*sat*, *aprB*) in sulfur cycling pathways (Fig. 2, Fig. 6C). Although DRAM did not find *soeC* in ALJR36 (Fig. 2), we were able to identify an incomplete copy using Anvi'o (Fig. 6C).

Phylogenetic analysis indicates structure in genomic variability

In the analyses described above, nucleotide variation and entropy analyses captured the heterogeneity of the samples, while pangenomic analysis suggested metabolic differences between closely-related MAGs. Using reassembled MAGs, we conducted phylogenetic analyses of single-copy genes and protein-coding marker genes (Supplemental Data 9), capitalizing on fixed changes at the sequence level (Fig. 5, Fig. 7A) indicative of divergent populations.

For the four groups of sample-specific MAGs whose reassembly had been guided by MA primary MAGs (MASA10, MAS15, MASP16, MASA4), sample-specific MAGs largely fell in well-supported Massachusetts (red-brown) and Alabama (green-olive) clades (Fig. 7B). MAGs from samples from the two MA vegetation zones (*S. alterniflorus* and *S. pumilus*) also tended to fall in separate clades, possibly either because the vegetation was different, the inundation intensity was different, or both.

MAGs from Alabama samples in group ALJR15 clearly segregated by plant type (Fig. 7B), though the plants occurred together under the same inundation regime. Sample-specific MAGs in the other three AL primary MAG groups did not segregate by vegetation or site (Fig. 7B).

Diversity-Generating Retroelements (DGRs) are present and active

Finally, we investigated the possibility that DGRs contributed to the diversification of the sample-specific MAGs. DGRs use an error-prone reverse transcriptase to generate variability in specific target genes (Fig. 8A). Two sulfur-oxidizing MAGs (ALSA124 and MASP15, Fig. 2), both in the *Gammaproteobacteria*, encoded full DGRs. (Beyond the 38 MAGs otherwise targeted in this paper, DGRs were also found in two other *Gammaproteobacteria* in the original 118 MAGs – ALJR7 and MASP22.) ALSA124 and ALJR7 (*Chromatiales*) contain a single DGR and a single target gene. MASP15 and MASP22 (UBA4575*)* encoded two and three DGRs, respectively, one (in MASP22) with multiple target genes (Fig. 8B).

In all target genes, the variable region of the DGR was encoded in a C-type lectin-like domain. This domain is found in a diverse group of proteins, some of which are involved in protein-protein, protein-lipid, or protein-nucleic acid interactions (23). Energy-related domains such as NTP hydrolases were found in 5 out of the 8 target genes. The most obvious instance of active DGR activity was detected in MASP22. Adenines in the template region were highly conserved (Fig. 8B). In the DGR target gene, 15 variable codons correspond to template (TR) adenines, which have the potential for DGR-directed non-synonymous substitutions, while not directly recoding to a stop codon (Fig. 8A). These correspond to nine asparagine (AAY) residues, one isoleucine (ATC), one glutamine (CAA), one tyrosine (TAT), two histidines (CAY), and one aspartic acid (GAT) encoded in TR. Taken together, the DGR target region can accommodate up to ~1.97x10^14 amino acid variants. In the mapped reads to this MAG, we observed recoding to between 2-8 amino acids at each DGR-variable position. This observation suggests that a diversity of between ~378 million and ~$2x10^{14}$ variants could be randomly rewritten through adenine mutagenesis of this single gene.

Discussion

Sulfur-cycling organisms are taxonomically diverse and very abundant in salt marsh sediments. Deep metagenomic sampling in AL and MA marsh sediments and comparative genomic analysis enabled the characterization of microbial genetic and genomic patterns from single nucleotide to pangenome scales.

Genomic patterns reveal the potential contribution of taxa across environments

We found SRB MAGs hosting genes for dissimilatory sulfate reduction (*sat*, *aprAB*, *dsrAB*) within multiple phyla, with *Desulfobacterota* dominant (53% of MAGs) as previously observed in Massachusetts (16), Mississippi (19), and Georgia (24) salt marshes. Other SRB MAGs belonged to groups (*Acidobacteriota*, *Bacteroidota*, and *Gemmatimonadota*) well recognized to contain sulfate reducers (10, 19, 25–28). All but one SRB primary MAG were co-assembled from AL samples, perhaps because SRBs are so diverse and the sequencing depth at AL was double that in MA. All sulfate-reducing MAGs belonged to genera without cultivated representatives, underlining the capacity of metagenomic analyses to further expand the recognized diversity and genomic potential of sulfate reducers (17, 29).

The taxonomic distribution of SOX MAGs containing the truncated thiosulfate oxidation system (*soxAXYZB*) was simpler, falling only within the phylum *Proteobacteria*, with *Gammaproteobacteria* representing 86% of the genomes. In contrast to the SRBs, 38% of the identified sulfur oxidizers belonged to genera with cultivated representatives. Consistent with previous work (5, 30), *Chromatiales* were well represented overall, and read-mapping to ALSA27 (a *Thiocapsa* sp.) indicated significantly greater abundance under *S. alterniflorus* in AL compared to any other sample set, including AL *J. roemerianus* (30). Variations in gene content associated with distinct metabolic steps of sulfur oxidation are well documented in the literature

(1, 31, 32). For example, sulfur oxidizers with the truncated *sox* system (lacking *sox*CD) often have *soe*ABC (e.g., (31)), and this pattern was observed for 12 of 21 SOX MAGs.

Clear patterns emerged in read recruitment from the 24 samples to primary SOX and SRB MAGs. These patterns indicate far more sparse mapping to SRB MAGs no matter the group from which they were co-assembled, possibly reflecting, again, the greater diversity of SRBs in marsh sediment. NMDS analysis of the read-mapping metric CPM shows strong overall separation between MA and AL samples, no effect of sampling month (variable in MA only), and a distinctive surface (0-2 cm) community in AL. Using phospholipid fatty acid analysis, (33) similarly found that salt marsh microbial communities sampled at the Wadden Sea were influenced by sediment depth but not the month (April, July, October) of sampling. In a 16S rRNA-based comparison of sediment microbial communities from MA to South Carolina, (34) also detected strong north-south differentiation.

NMDS analysis also clearly shows separation within MA of samples taken under *S. pumilus* and *S. alterniflorus*, which may reflect the influence of the plants themselves, the tidal regime (inundation frequency and depth being low for *S. pumilus* and high for *S. alterniflorus*), or both. Data from (35) show that in the greenhouse, rhizosphere microbial communities vary even with different MA genotypes of *S. alterniflorus*, but once transplanted back into the marsh, plant-specific differences disappeared and microbial communities correlated instead with local sediment properties. (33) also demonstrated that salt marsh zones defined by flooding regimes had distinct microbial communities. The strong influence of the local sediment tidal regime is further suggested by the lack of difference between AL samples taken under *S. alterniflorus* and *J. roemerianus*, which were both rooted in the same area with the same tidal regime.

Biogeography and potential microdiversification

Local variants were detected among sample-specific MAGs reassembled with guidance from the primary MAGs. Already we had seen that reads mapping to some groups of contigs appeared only in samples from particular locations (e.g., MA samples in the MASP16 group) or even only in samples under one vegetation/tidal regime in MA (e.g., MASA10). Not only did AL and MA S-cycling MAGs separate cleanly on NMDS plots, but also the phylogenetic trees built from sample-specific MAGs showed consistent separation of MA and AL samples (except when orthologous genes in MA sample-specific MAGs were rare, as in ALJR15, ALSA124, and ALSA 174. These geographic patterns are consistent with the possibility of distance and climate differences working as a geographical barrier and contributing to allopatric differentiation (when a species separates into two or more species, due to geographic isolation and differences in environmental conditions) of microbial communities and/or the emergence of subspecies/strains (36, 37).

UBA6429 MAGs, ALJR36, and MASA10, had particularly high similarity to one another (ANI >99%), high read recruitment across all sample-specific reassembled MAGs, strong representation of SCG clusters, and yet notable differences in the presence/absence of orthologous genes. The contrasts in genomic content in these two sets of MAGs could point to metabolic differentiation, or even hint at adaptation to different environments or niches. Functional enrichment analysis identified that the ALJR36 group of sample-specific MAGs were enriched in key enzymes of the reductive citrate cycle (Arnon-Buchanan cycle (38). This CO_2 fixation cycle is found in anaerobic and microaerobic autotrophic bacteria (39) and anoxygenic phototrophs (40). When water-saturated, salt marsh sediments are anaerobic. Particularly conducive conditions for the reductive citrate cycle's operation would be provided by the low permeability, high clay content sediment in AL marsh near Dauphin Island airport (41). In

contrast, functional enrichment analysis identified that the MASA10 group of sample-specific MAGs were enriched in a diverse genetic repertoire of genes involved in assimilatory and dissimilatory sulfur cycling including the full set of *soeABC*, *sat*, and *aprAB* genes. ALJR36 MAGs lack *aprB*, *soeB*, and *sat* genes, potentially significantly altering transformations among sulfate, sulfite, and APS. Because these are not isolated organisms, it is not possible to test unique physiologies. There are no cultivated members of the UBA6429 family of sulfur oxidizers. But, if these MAGs do accurately reflect the genetic repertoire of microbes abundant enough to have supported MAG assembly, such differences in basic gene content could reflect genomic microdiversification contributing to the microbes' coexistence despite their sharing large portions of their genomes (42, 43).

Finally, maintaining genetic variation is important for species stability and survival (44, 45). We found DGRs encoded in the MAGs and evidence of their dynamism in the form of extraordinarily elevated non-synonymous substitutions, corresponding to adenine-specific variation of VR encoded in the target genes. Though the identity of the target genes remains unknown in these MAGs, in all target genes the variable region of the DGR was encoded in a C-type lectin-like domain, known to be involved in protein-protein, protein-lipid or protein-nucleic acid interactions (23).

These results, from single nucleotide to pangenome scales, demonstrate that a tremendous range of genetic information available in metagenomic datasets, even those from microbially diverse systems such as salt marsh sediments, can be harnessed for analysis of biogeographic and biotic patterns in targeted taxa.

Methods

Data set collection. We compared the sulfur-cycling bacteria in sediments of two salt marshes, representing Northern (Massachusetts, MA) and Southern sites (Alabama, AL). These sites also differed in the type of sediment with MA area being carbon rich and AL formed by fine, sandy sediments. Metagenomic datasets were retrieved in July 2020 from the Genomes Online Database (*https://gold.jgi.doe.gov/*) and had been sequenced by the Joint Genome Institute using Illumina NovaSeq as sequencing technology and all samples consisted of paired-end, non-overlapping, 150 nt read length. The 24 metagenomes analyzed included samples taken from rooted sediment under common dominant plant species in their native range, and all of them have a sequencing depth of > 30 million reads per sample (Table S1). The Alabama dataset (Gs0135940 P.I. Olivia Mason), included rhizosphere sediment samples collected under *J. roemerianus* (n=8, total reads ~ 578 M, abbreviated as ALJR) and *S. alterniflorus* (n=6, total reads ~ 514 M, abbreviated as ALSA) on Dauphin Island (15). Two to three samples were taken at various depths (all within 0-10 cm) in May during 2015, 2016, and 2017. The elevation of *J. roemerianus* and *S. alterniflorus* patches do not differ and the marsh is flooded on every high tide. The Massachusetts dataset (Gs0142363, P.I. Jennifer Bowen) included samples taken during May, August, and October from low marsh sediments (0-5 cm depth) under *S. alterniflorus* (n=5, total reads ~ 231 M, abbreviated as MASA) and from high marsh sediments under *S. pumilus* (n=5, total reads ~ 208 M, abbreviated as MASP) in the Plum Island Long Term Ecological Research (PIE-LTER) site.

Metagenomic assembly, binning and analyses. Samples were grouped by site and vegetation for increased sequencing depth during the initial reconstruction of "primary MAGs". Reads were assembled using MEGAHIT (version 1.2.9) (46) and the quality of the four resultant assemblies

was evaluated using MetaQUAST (47) (Table S2). Taxonomic classification of the contigs forming each co-assemblies was performed using Kraken2 (48). Contigs were assigned into bins using MetaWRAP (49) processing modules with initial extraction using MaxBin2 (50), metaBAT2 (51), and CONCOCT (52) and finalized using the bin refinement module. Completeness and contamination of each MAG were evaluated with CheckM (53). Quality was reported based on Bowers et al. (54). MAGs were annotated using Distilled and Refined Annotation of Metabolism (DRAM) (55). Sulfur-cycling pathways were manually curated by searching key sulfur genes in the raw (Supplemental Data 1) and metabolism (Supplemental Data 2) output files provided by DRAM. MAGs corresponding to sulfur-cycling bacteria were identified by the presence of dissimilatory reduction of sulfate to sulfite pathway (*sat*, *apr*AB, and *dsr*AB) or the truncated thiosulfate oxidation system (*sox*AX, *sox*YZ, *sox*B) (56). The presence of sulfide and sulfite oxidation genes including flavocytochrome c complex (*fcc*AB), sulfide:quinone oxidoreductase (*sqr*), and the sulfite:quinone oxidoreductase complex (*soe*ABC) was also investigated. Taxonomy and closest phylogenetic neighbors were assigned using GTDB-tk (v1.5.0) (57) and average nucleotide identity (ANI) computed by EZBioCloud OrthoANIu Calculator (58). A maximum likelihood phylogenetic tree of genomes was generated using the Bacterial and Viral Bioinformatic Resource Center (59, 60).

Distribution of 38 primary MAGs' sequences across samples. The metaWRAP Quant_bins module (49) with Salmon (61) was used to recruit reads from each sample to the contigs forming each of the 38 primary MAGs. Coverage values are standardized by library size and contig length, then expressed as counts per million (CPM). CPM metrics for entire MAGs were calculated as the average CPM across individual contigs. Results below 2 CPM were filtered out.

For each MAG, average CPM values were compared to contig-specific CPM. Additionally, to determine MAGs occurrence in less deeply sequenced data, metagenomic reads from unvegetated salt marsh sediments from where the MA samples originated (62) (PRJNA812947) were mapped to the 38 MAG contigs.

Finally, we quantitatively compared the distribution of sulfur cycling primary MAG sequences across sites and vegetation (in AL and MA), sample depths (AL only), and sampling month (MA only) using permutational multivariate analysis of variance (PerMANOVA) (63) and post-hoc comparisons (64), based on the average CPM values calculated for each MAG. Pairwise statistical analysis of metagenomic profiles (STAMP) was used to identify primary MAGs for which read-mapping was statistically more abundant in particular sites and/or vegetation zones (65). Non-metric multidimensional scaling (NMDS) analyses were generated using the function metaMDS (63).

Analysis of genomic diversity. We selected eight primary MAGs for further analyses, two from each site and vegetation, that were detected (by Quant_bin) at all sampling locations, and that varied notably in average CPM across samples. Six were sulfur-oxidizers (ALJR36, ALSA124, MASA10, MASA4, MASP15, MASP16) and two were sulfate-reducers (ALJR15, and ALSA174).

Genomic diversity was analyzed using Anvi'o v7.1 "hope" following the workflow outlined in (66). Open reading frames (ORFs) for each MAG were identified with Prodigal v2.6.3 (67) and HMMER (68) and annotated with COG20 (69) and GhostKoala (70). Anvi-init-bam was used to sort bam files generated with Bowtie2 (71) and calculate coverage and genetic variability metrics. Site-specific variability was analyzed in a subset of representative genes

(Supplemental Data 7) including ribosomal proteins, DNA-directed RNA polymerase subunits, translation initiation factor, representative genes of the broad metabolism (72), and sulfur cycling genes. For each gene, we calculated the average gene entropy using custom R scripts (Supplemental Material).

Pangenomic analysis of reference-guided, sample-specific, assembled MAGs. A reference-guided reassembly strategy allowed us to reconstruct sample-specific MAGs and capture the genomic variability of a given taxon across all the study sites. We generated new, sample-specific MAGs guided by the selected eight primary MAGs using BWA (73) and SPAdes (74) using the reassemble_bins module from metaWRAP (49). Primary MAG and the derived reassembled genomes were analyzed with the Anvi'o pangenomics pipeline [https://merenlab.org/2016/11/08/pangenomics-v2/]. Gene functions of the derived MAGs were assigned using NCBI COG20 (69), HMM hits from Kofam (75), and EBI's Pfam database (76). The predicted amino acid sequence for a set of core genes was extracted with anvi-get-sequences-for-gene-clusters (Supplemental Data 9) and used to build maximum-likelihood phylogenetic trees with FastTree (77). The module anvi-compute-functional-enrichment was used to identify functions and pathways differentially represented in pan-groups consisting of site-specific MAGs at the various sites and vegetation zones.

Identification of diversity-generating retroelements in sulfur cycling bacteria. Finally, to investigate possible sources of genetic dynamism, we explored the distribution and activity of diversity-generating retroelements (DGRs) (78). DGRs were identified in binned assemblies with the python package DGRpy (https://pypi.org/project/DGR-package/), which annotates the

essential DGR features (retrotranscriptase RT, variable region VR, and template region TR) within 20kbp windows (78, 79). Putative DGR hypermutation target genes were identified by manual curation of the contigs. The dynamism of the DGRs was detected by visually inspecting the variability of recruited reads against the VR.

Acknowledgements

This study was supported by a Gordon and Betty Moore Foundation grant #9437 to ZGC, ELP, BP, SER, and AEG, and by grants from the Simons Foundation (824763, to SER) and the National Science Foundation (2224608, to AEG). We thank Richard Fox and the MBL Bay Paul Center for supporting necessary computing resources.

Data availability

All sequence data is available at NCBI under BioProject IDs PRJNA620304, PRJNA539072, and PRJNA814317. Accession numbers for individual metagenomes can be found in Table S1 .

References

1. Wu B, Liu F, Fang W, Yang T, Chen G-H, He Z, Wang S. 2021. Microbial sulfur metabolism and environmental implications. Sci Total Environ 778:146085.
2. Lee RW, Kraus DW, Doeller JE. 1999. Oxidation of sulfide by *Spartina alterniflora* roots. Limnol Oceanogr 44:1155–1159.
3. Lamers LPM, Govers LL, Janssen ICJM, Geurts JJM, Van der Welle MEW, Van Katwijk MM, Van der Heide T, Roelofs JGM, Smolders AJP. 2013. Sulfide as a soil phytotoxin—a review. Front Plant Sci 4: 268.

4. Zhang P, Luo Q, Wang R, Xu J. 2017. Hydrogen sulfide toxicity inhibits primary root growth through the ROS-NO pathway. Sci Rep 7:868.

5. Thomas F, Giblin AE, Cardon ZG, Sievert SM. 2014. Rhizosphere heterogeneity shapes abundance and activity of sulfur-oxidizing bacteria in vegetated salt marsh sediments. Front Microbiol 5: 309.

6. Teal JM. 1962. Energy flow in the salt marsh ecosystem of Georgia. Ecology 43:614–624.

7. Rocha AV, Goulden ML. 2009. Why is marsh productivity so high? New insights from eddy covariance and biomass measurements in a *Typha* marsh. Agric For Meteorol 149:159–168.

8. Forbrich I, Giblin AE, Hopkinson CS. 2018. Constraining marsh carbon budgets using long-term C burial and contemporary atmospheric CO_2 fluxes. J Geophys Res Biogeosciences 123:867–878.

9. Valiela I, Lloret J, Bowyer T, Miner S, Remsen D, Elmstrom E, Cogswell C, Robert Thieler E. 2018. Transient coastal landscapes: Rising sea level threatens salt marshes. Sci Total Environ 640–641:1148–1156.

10. Hausmann B, Pelikan C, Herbold CW, Köstlbacher S, Albertsen M, Eichorst SA, Glavina del Rio T, Huemer M, Nielsen PH, Rattei T, Stingl U, Tringe SG, Trojan D, Wentrup C, Woebken D, Pester M, Loy A. 2018. Peatland *Acidobacteria* with a dissimilatory sulfur metabolism. ISME J 12:1729–1742.

11. Chase AB, Karaoz U, Brodie EL, Gomez-Lunar Z, Martiny AC, Martiny JBH. 2017. Microdiversity of an abundant terrestrial bacterium encompasses extensive variation in ecologically relevant traits. mBio 8:e01809-17.

12. Naidoo G, McKee K, Mendelssohn I. 1992. Anatomical and metabolic responses to

waterlogging and salinity in *Spartina alterniflora* and *S. patens* (Poaceae). Am J Bot 79:765–770.

13. Pennings SC, Grant M-B, Bertness MD. 2005. Plant zonation in low-latitude salt marshes: Disentangling the roles of flooding, salinity and competition. J Ecol 93:159–167.
14. Bertness MD. 1991. Zonation of *Spartina patens* and *Spartina slterniflora* in New England Salt Marsh. Ecology 72:138–148.
15. Mason OU, Chanton P, Knobbe LN, Zaugg J, Mortazavi B. 2021. New insights into the influence of plant and microbial diversity on denitrification rates in a salt marsh. Wetlands 41:33.
16. Bahr M, Crump BC, Klepac-Ceraj V, Teske A, Sogin ML, Hobbie JE. 2005. Molecular characterization of sulfate-reducing bacteria in a New England salt marsh. Environ Microbiol 7:1175–1185.
17. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC, Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF. 2016. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun 7:13219.
18. Cleary DFR, Coelho FJRC, Oliveira V, Gomes NCM, Polónia ARM. 2017. Sediment depth and habitat as predictors of the diversity and composition of sediment bacterial communities in an inter-tidal estuarine environment. Mar Ecol 38:e12411.
19. Mavrodi OV, Jung CM, Eberly JO, Hendry SV, Namjilsuren S, Biber PD, Indest KJ, Mavrodi DV. 2018. Rhizosphere microbial communities of *Spartina alterniflora* and *Juncus roemerianus* from restored and natural tidal marshes on Deer Island, Mississippi. Front Microbiol 9:1–13.

20. Angermeyer A, Crosby SC, Huber JA. 2016. Decoupled distance-decay patterns between *dsr*A and 16S rRNA genes among salt marsh sulfate-reducing bacteria. Environ Microbiol 18:75–86.

21. Polz MF, Rajora OP. 2019. Population genomics : Microorganisms. Springer.

22. Shaiber A, Willis AD, Delmont TO, Roux S, Chen L-X, Schmid AC, Yousef M, Watson AR, Lolans K, Esen ÖC, Lee STM, Downey N, Morrison HG, Dewhirst FE, Mark Welch JL, Eren AM. 2020. Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. Genome Biol 21:292.

23. Kilpatrick DC. 2002. Animal lectins: A historical introduction and overview. Biochim Biophys Acta BBA - Gen Subj 1572:187–197.

24. Kolton M, Rolando JL, Kostka JE. 2020. Elucidation of the rhizosphere microbiome linked to *Spartina alterniflora* phenotype in a salt marsh on Skidaway Island, Georgia, USA. FEMS Microbiol Ecol 96:1–19.

25. Müller AL, Kjeldsen KU, Rattei T, Pester M, Loy A. 2015. Phylogenetic and environmental diversity of DsrAB-type dissimilatory (bi)sulfite reductases. ISME J 9:1152–1165.

26. Thiel V, Garcia Costas AM, Fortney NW, Martinez JN, Tank M, Roden EE, Boyd ES, Ward DM, Hanada S, Bryant DA. 2019. "Candidatus *Thermonerobacter thiotrophicus*," A non-phototrophic member of the *Bacteroidetes*/*Chlorobi* with dissimilatory Sulfur metabolism in hot spring mat communities. Front Microbiol 9:3159.

27. Kalam S, Basu A, Ahmad I, Sayyed RZ, El-Enshasy HA, Dailin DJ, Suriani NL. 2020. Recent understanding of soil *Acidobacteria* and their ecological significance: A critical review. Front Microbiol 11:580024.

28. Langwig MV, De Anda V, Dombrowski N, Seitz KW, Rambo IM, Greening C, Teske AP, Baker BJ. 2022. Large-scale protein level comparison of *Deltaproteobacteria* reveals cohesive metabolic groups. 1. ISME J 16:307–320.

29. Anantharaman K, Hausmann B, Jungbluth SP, Kantor RS, Lavy A, Warren LA, Rappé MS, Pester M, Loy A, Thomas BC, Banfield JF. 2018. Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. 7. ISME J 12:1715–1728.

30. Rietl AJ, Overlander ME, Nyman AJ, Jackson CR. 2016. Microbial Community composition and extracellular enzyme activities associated with *Juncus roemerianus a*nd *Spartina alterniflora* vegetated sediments in Louisiana saltmarshes. Microb Ecol 71:290–303.

31. Watanabe T, Kojima H, Umezawa K, Hori C, Takasuka TE, Kato Y, Fukui M. 2019. Genomes of neutrophilic sulfur-oxidizing chemolithoautotrophs representing 9 proteobacterial species from 8 genera. Front Microbiol 10.

32. Rameez MJ, Pyne P, Mandal S, Chatterjee S, Alam M, Bhattacharya S, Mondal N, Sarkar J, Ghosh W. 2020. Two pathways for thiosulfate oxidation in the alphaproteobacterial chemolithotroph *Paracoccus thiocyanatus* SST. Microbiol Res 230:126345.

33. Rinke M, Maraun M, Scheu S. 2022. Spatial and temporal variations in salt marsh microorganisms of the Wadden Sea. Ecol Evol 12:e8767.

34. Angermeyer A, Crosby SC, Huber JA. 2018. Salt marsh sediment bacterial communities maintain original population structure after transplantation across a latitudinal gradient. PeerJ 2018:1–21.

35. Hanley TC, Bowen JL, Kearns PJ, Hughes AR. 2021. Short- and long-term effects of nutrient enrichment on salt marsh plant production and microbial community structure. J

Ecol 109:3779–3793.

36. Bobay L-M, Raymann K. 2019. Population genetics of host-associated microbiomes. Curr Mol Biol Rep 5:128–139.

37. Van Rossum T, Ferretti P, Maistrenko OM, Bork P. 2020. Diversity within species: interpreting strains in microbiomes. 9. Nat Rev Microbiol 18:491–506.

38. Buchanan BB, Sirevåg R, Fuchs G, Ivanovsky RN, Igarashi Y, Ishii M, Tabita FR, Berg IA. 2017. The Arnon–Buchanan cycle: A retrospective, 1966–2016. Photosynth Res 134:117–131.

39. Berg IA. 2011. Ecological aspects of the distribution of different autotrophic CO2 fixation pathways. Appl Environ Microbiol 77:1925–1936.

40. Bhatnagar S, Cowley ES, Kopf SH, Pérez Castro S, Kearney S, Dawson SC, Hanselmann K, Ruff SE. 2020. Microbial community dynamics and coexistence in a sulfide-driven phototrophic bloom. Environ Microbiome 15:3.

41. Ellis AM, Smith CG, Marot ME. 2018. The sedimentological characteristics and geochronology of the marshes of Dauphin Island, Alabama. 2017–1165The sedimentological characteristics and geochronology of the marshes of Dauphin Island, Alabama. USGS Numbered Series. U.S. Geological Survey, Reston, VA.

42. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. 2018. Microdiversification in genome-streamlined ubiquitous freshwater *Actinobacteria*. ISME J 12:185–198.

43. Chase AB, Weihe C, Martiny JBH. 2021. Adaptive differentiation and rapid evolution of a soil bacterium along a climate gradient. Proc Natl Acad Sci 118:e2101254118.

44. Sharon I, Morowitz MJ, Thomas BC, Costello EK, Relman DA, Banfield JF. 2013. Time series community genomics analysis reveals rapid shifts in bacterial species, strains, and

phage during infant gut colonization. Genome Res 23:111–120.

45. Paul BG, Eren AM. 2022. Eco-evolutionary significance of domesticated retroelements in microbial genomes. Mob DNA 13:6.

46. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. 2015. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31:1674–1676.

47. Mikheenko A, Saveliev V, Gurevich A. 2016. MetaQUAST: Evaluation of metagenome assemblies. Bioinformatics 32:1088–1090.

48. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257.

49. Uritskiy GV, DiRuggiero J, Taylor J. 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome 6:158.

50. Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32:605–607.

51. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ 2019:1–13.

52. Alneberg J, Bjarnason BS, De Bruijn I, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nat Methods 11:1144–1146.

53. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055.

54. Bowers RM, Lauber CL, Wiedinmyer C, Hamady M, Hallar AG, Fall R, Knight R, Fierer N. 2009. Characterization of airborne microbial communities at a high-elevation site and their potential to act as atmospheric ice nuclei. Appl Environ Microbiol 75:5121–5130.

55. Shaffer M, Borton MA, McGivern BB, Zayed AA, La Rosa SL, Solden LM, Liu P, Narrowe AB, Rodríguez-Ramos J, Bolduc B, Gazitúa MC, Daly RA, Smith GJ, Vik DR, Pope PB, Sullivan MB, Roux S, Wrighton KC. 2020. DRAM for distilling microbial metabolism to automate the curation of microbiome function. Nucleic Acids Res 48:8883–8900.

56. Berben T, Overmars L, Sorokin DY, Muyzer G. 2019. Diversity and distribution of sulfur oxidation-related genes in *Thioalkalivibrio*, a genus of chemolithoautotrophic and haloalkaliphilic sulfur-oxidizing bacteria. Front Microbiol 10.

57. Chaumeil PA, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: A toolkit to classify genomes with the genome taxonomy database. Bioinformatics 36:1925–1927.

58. Yoon S-H, Ha S, Lim J, Kwon S, Chun J. 2017. A large-scale evaluation of algorithms to calculate average nucleotide identity. Antonie Van Leeuwenhoek 110:1281–1286.

59. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, Conrad N, Dietrich EM, Disz T, Gabbard JL, Gerdes S, Henry CS, Kenyon RW, Machi D, Mao C, Nordberg EK, Olsen GJ, Murphy-Olson DE, Olson R, Overbeek R, Parrello B, Pusch GD, Shukla M, Vonstein V, Warren A, Xia F, Yoo H, Stevens RL. 2017. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res 45:D535–D542.

60. Olson RD, Assaf R, Brettin T, Conrad N, Cucinell C, Davis JJ, Dempsey DM, Dickerman A, Dietrich EM, Kenyon RW, Kuscuoglu M, Lefkowitz EJ, Lu J, Machi D, Macken C, Mao C, Niewiadomska A, Nguyen M, Olsen GJ, Overbeek JC, Parrello B, Parrello V, Porter JS,

Pusch GD, Shukla M, Singh I, Stewart L, Tan G, Thomas C, VanOeffelen M, Vonstein V, Wallace ZS, Warren AS, Wattam AR, Xia F, Yoo H, Zhang Y, Zmasek CM, Scheuermann RH, Stevens RL. 2022. Introducing the bacterial and viral bioinformatics resource center (BV-BRC): A resource combining PATRIC, IRD and ViPR. Nucleic Acids Res 51:D678–D689.

61. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. 4. Nat Methods 14:417–419.
62. Graves CJ, Makrides EJ, Schmidt VT, Giblin AE, Cardon ZG, Rand DM. 2016. Functional responses of salt marsh microbial communities to long-term nutrient enrichment. Appl Environ Microbiol 82:2862–2871.
63. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens M, Wagner H. 2013. Vegan: Community Ecology Package. R Package Version. 2.0-10. CRAN.
64. Hervé M. 2022. RVAideMemoire: Testing and plotting procedures for biostatistics (0.9-81-2).
65. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. 2014. STAMP: Statistical analysis of taxonomic and functional profiles. Bioinformatics 30:3123–3124.
66. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, Fink I, Pan JN, Yousef M, Fogarty EC, Trigodet F, Watson AR, Esen ÖC, Moore RM, Clayssen Q, Lee MD, Kivenson V, Graham ED, Merrill BD, Karkman A, Blankenberg D, Eppley JM, Sjödin A, Scott JJ, Vázquez-Campos X, McKay LJ, McDaniel EA, Stevens SLR, Anderson RE, Fuessel J, Fernandez-Guerra A, Maignien L, Delmont TO, Willis AD. 2021. Community-led, integrated, reproducible multi-omics with Anvi'o. Nat Microbiol 6:3–6.

67. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: Prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119.

68. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: Interactive sequence similarity searching. Nucleic Acids Res 39:29–37.

69. Galperin MY, Wolf YI, Makarova KS, Alvarez RV, Landsman D, Koonin EV. 2021. COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. Nucleic Acids Res 49:D274–D281.

70. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462.

71. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

72. Saad S, Bhatnagar S, Tegetmeyer HE, Geelhoed JS, Strous M, Ruff SE. 2017. Transient exposure to oxygen or nitrate reveals ecophysiology of fermentative and sulfate-reducing benthic microbial populations. Environ Microbiol 19:4866–4881.

73. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760.

74. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477.

75. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. 2020. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score

threshold. Bioinformatics 36:2251–2252.

76. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res 49:D412–D419.

77. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

78. Roux S, Paul BG, Bagby SC, Nayfach S, Allen MA, Attwood G, Cavicchioli R, Chistoserdova L, Gruninger RJ, Hallam SJ, Hernandez ME, Hess M, Liu W-T, McAllister TA, O'Malley MA, Peng X, Rich VI, Saleska SR, Eloe-Fadrosh EA. 2021. Ecology and molecular targets of hypermutation in the global microbiome. 1. Nat Commun 12:3076.

79. Paul BG, Burstein D, Castelle CJ, Handa S, Arambula D, Czornyj E, Thomas BC, Ghosh P, Miller JF, Banfield JF, Valentine DL. 2017. Retroelement-guided protein diversification abounds in vast lineages of *Bacteria* and *Archaea*. 6. Nat Microbiol 2:1–7.

Author contributions

S.P.C. and E.LP. designed the study, processed and analyzed the metagenomic data, and with Z.G.C. wrote the manuscript with input from all co-authors. O.M. and B.M. collected and sequenced Alabama samples. J.V. and J.B. collected and sequenced Massachusetts samples, A.G. and B.P. processed and analyzed the DGR surveys. S.P.C., E.LP., S.E.R., B.P., A.E.G., and Z.G.C. conceived and designed the study.

Figure Legends

Figure 1 Overview of study sites and metagenomic workflow. Rhizosphere sediment samples were collected in Alabama under *J. roemerianus* (ALJR) and *S. alterniflorus* (ALSA) and in Massachusetts under *S. alterniflorus* (MASA) and *S. pumilus* (MASP). Twenty-four metagenomes yielded 38 MAGs for S-cycling bacteria (17 SRB and 21 SOX). Ubiquitous lineages were used to investigate genomic diversity.

Figure 2 Core sulfur cycling genes detected in MAGs from Alabama and Massachusetts salt marsh samples. Key genes encoded dissimilatory sulfate reduction (purple) and sulfur oxidation (yellow). MAGs are organized by their GTDB-tk taxonomy, and named by site (AL or MA), vegetation, and bin identification number. Multiple copy genes are given as numbers.

Figure 3 Maximum likelihood phylogenetic trees of sulfate-reducing (left) and sulfur-oxidizing (right) MAGs from Alabama and Massachusetts salt marshes, and their closest relatives, based on analysis of single-copy genes. Phyla are shown on the left of each tree.

Figure 4 Distribution of sulfur cycling primary MAG sequences across sites and vegetation zones. **A)** Quantification of read-mapping from AL and MA samples to primary MAG contigs, with coverage standardized by library size and contig length and expressed as counts per million reads (CPM). Circle size reflects averaged CPM across all contigs for each MAG. **B-D)**Non-metric multidimensional scaling (NMDS) analyses comparing MAGs distribution as expressed as CPM in: **B)** Alabama and Massachusetts salt marshes, **C)** Massachusetts salt marshes by month, and **D)** Alabama salt marshes by depth.

Figure 5 A). Anvi'o coverage profiles of eight selected sulfur-cycling bacteria MAGs. Outer cycles represent the contig coverage (as mean coverage) of each of the 24 salt-marsh metagenomic samples included in this analysis. Inner rings represent GC content (dark gray) and contig length (gray). Contigs are clustered (inner tree) based on the sequence composition and differential coverage using Euclidean distance and Ward hierarchical clustering method. Sample order (rings) was determined using a clustering method based on the mean coverage and each ring is color-coded according to site and vegetation type. Bar plots represent (top to bottom) the total number of reads in each library, the total number of reads mapped to each respective MAG, the percentage of mapped reads, and the total number of Single Nucleotide Variants (SNV). **B)** Anvi'o pangenomic analysis of eight sulfur-cycling bacteria MAGs. In each ring, each vertical line represents a gene. Gene order was determined using Euclidean distance and Ward clustering method based on presence/absence of a gene cluster across MAGs. Annotation based on COG 2020 is indicated in gray and single-copy core genes in purple. Primary MAG is shown in dark teal blue. Each of the remaining circles represents a reference-guided reassembled MAG from each of the salt-mash metagenomic samples included in this analysis. Samples are color-coded according to geographical site and vegetation.

Figure 6 Comparative analysis of genomic and metabolic diversity of two closely-related sulfur-oxidizing MAGs (*Thiohalomonadales*) **A)** Distribution of genetic variability represented as SNV per kilobase (black bars) and calculated for each metagenomic sample. Hotspots of variability include COG5361(Mobilome), COG1850 (Rubisco-like protein), COG0457 (Tetratricopeptide (TPR) repeat), and rRNA. **B)** Diagram of sulfur metabolism pathways highlighting the

differences in gene content between sample-specific MAGs from the ALJ36 and MASA10 primary MAG groups (adapted from Kegg map00920) **C)** Pangenomic analysis of metagenomes from group ALJR36 and MASA10. Primary MAGs are indicated by the dark ring and sample-specific re-assembled MAGs are color-coded by site and vegetation. Genes clusters are ordered with forced synteny to ALJR36 to highlight the genes exclusively found in the MASA10 group. Annotations, single-copy core genes, and the combined homogeneity index are shown in outer rings.

Figure 7 Biogeographic and vegetation-specific distributions of genomic variability. **A)** Example aggregated view of metagenomic reads from different samples mapped to the MAG ALJR15. Samples collected in Alabama matched the reference while those collected in Massachusetts in sediments dominated by *S. pumilus* displayed numerous fixed mismatches. Mismatches are indicated with colors: green, T; yellow, G; blue, C; red, A. **B)** Trees depicting phylogenetic relationships among reassembled sample-specific metagenomes. Trees were built using single-copy genes, housekeeping genes, and sulfur-cycling genes. Tree branches are color-coded according to site and vegetation.

Figure 8 A) Structure of an example DGR identified in MASP22. Diversification of this DGR in this metagenomic sample is indicated by the accumulation of in-frame mutations in the variable region (VR) among the aligned reads. The template region (TR) is conserved both at the nucleotide and amino acid levels. In the bar graphs, the bar size is proportional to the degree of conservation of the nucleotide position. Adenine positions in the VR and TR regions are indicated in red. In the recruited reads, identical residues are indicated with dots, and changes are

color coded. **B)** Structure and domain composition of the target genes containing variable regions (in purple) of the DGRs.

Figure 1 Overview of study sites and metagenomic workflow. Rhizosphere sediment samples were collected in Alabama under J. roemerianus (ALJR) and S. alterniflorus (ALSA) and in Massachusetts under S. alterniflorus (MASA) and S. pumilus (MASP). Twenty-four metagenomes yielded 38 MAGs for S-cycling bacteria (17 SRB and 21 SOX). Ubiquitous lineages were used to investigate genomic diversity.

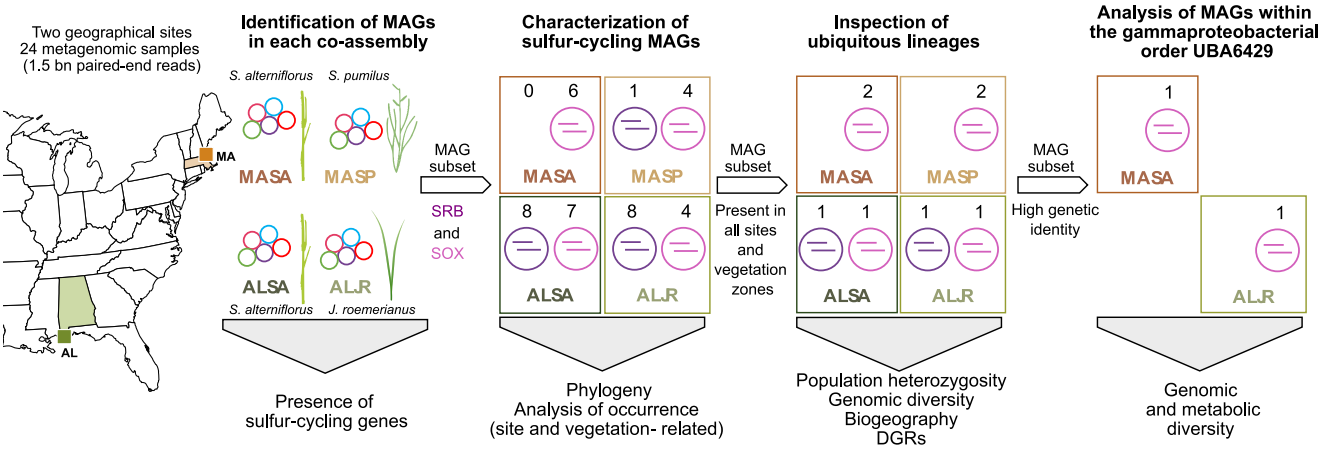


Figure 2 Core sulfur cycling genes detected in MAGs from Alabama and Massachusetts salt marsh samples. Key genes encoded dissimilatory sulfate reduction (purple) and sulfur oxidation (yellow). MAGs are organized by their GTDB-tk taxonomy, and named by site (AL or MA), vegetation, and bin identification number. Multiple copy genes are given as numbers.

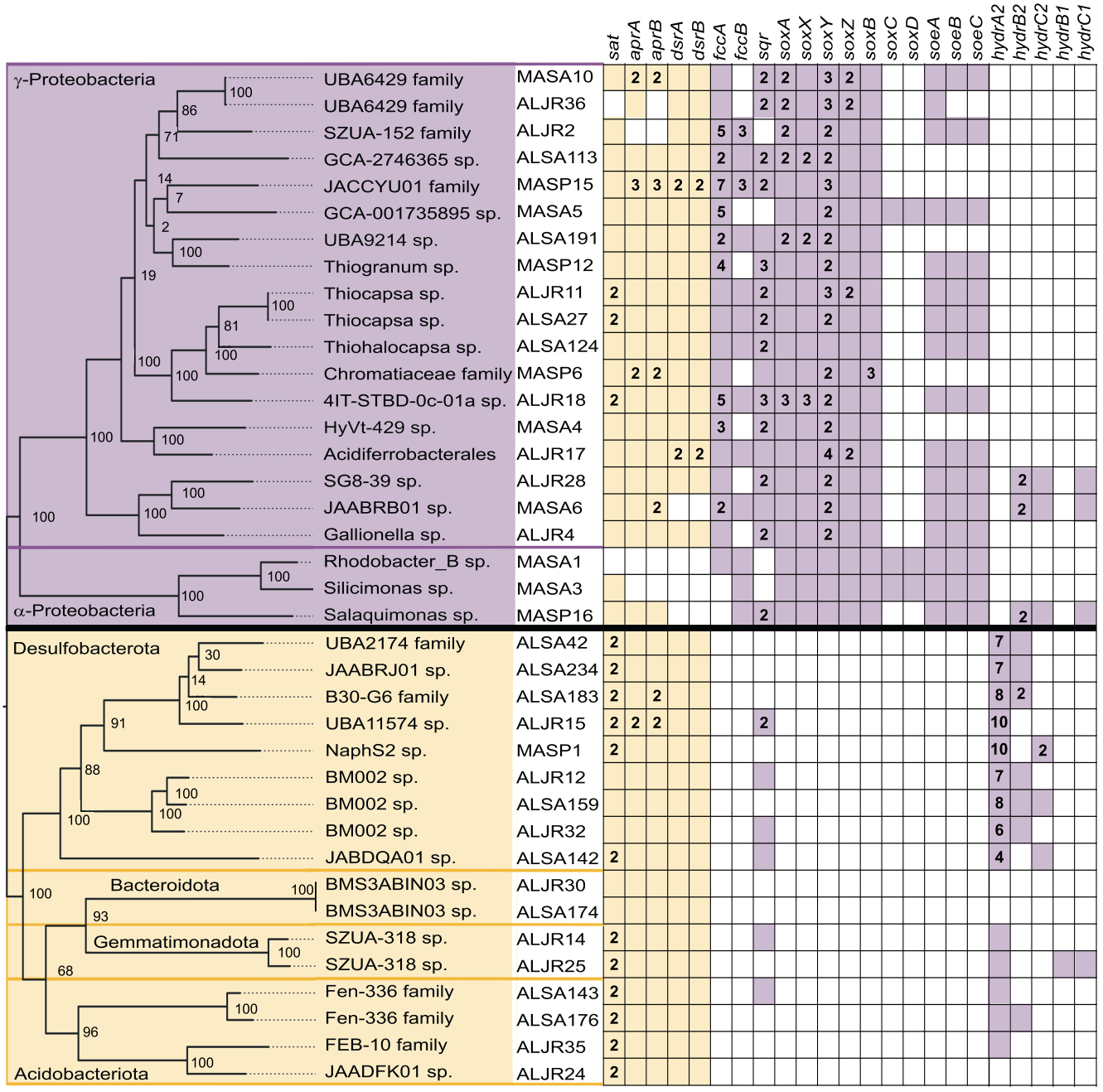


Figure 3 Maximum likelihood phylogenetic trees of sulfate-reducing (left) and sulfur-oxidizing (right) MAGs from Alabama and Massachusetts salt marshes, and their closest relatives, based on analysis of single-copy genes. Phyla are shown on the left of each tree.

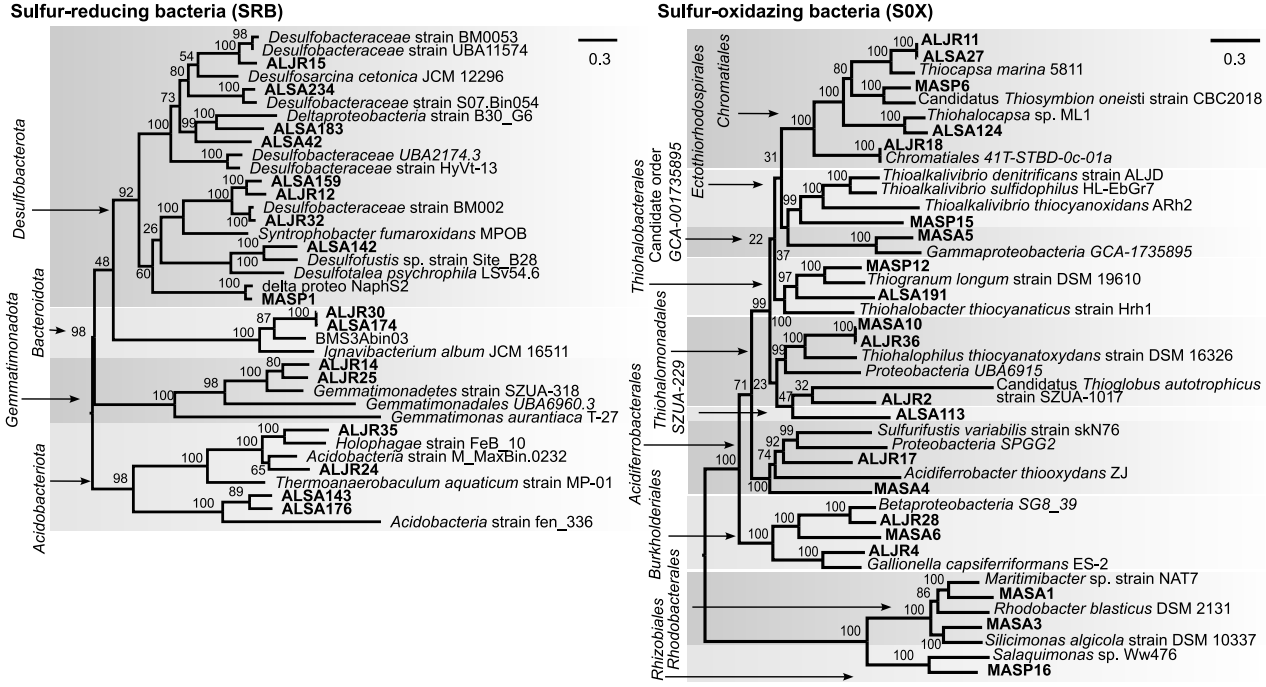


Figure 4 Distribution of sulfur cycling primary MAG sequences across sites and vegetation zones. A) Quantification of read-mapping from AL and MA samples to primary MAG contigs, with coverage standardized by library size and contig length and expressed as counts per million reads (CPM). Circle size reflects averaged CPM across all contigs for each MAG. B-D) Non-metric multidimensional scaling (NMDS) analyses comparing MAGs distribution as expressed as CPM in: B) Alabama and Massachusetts salt marshes, C) Massachusetts salt marshes by month, and D) Alabama salt marshes by depth (in cm).

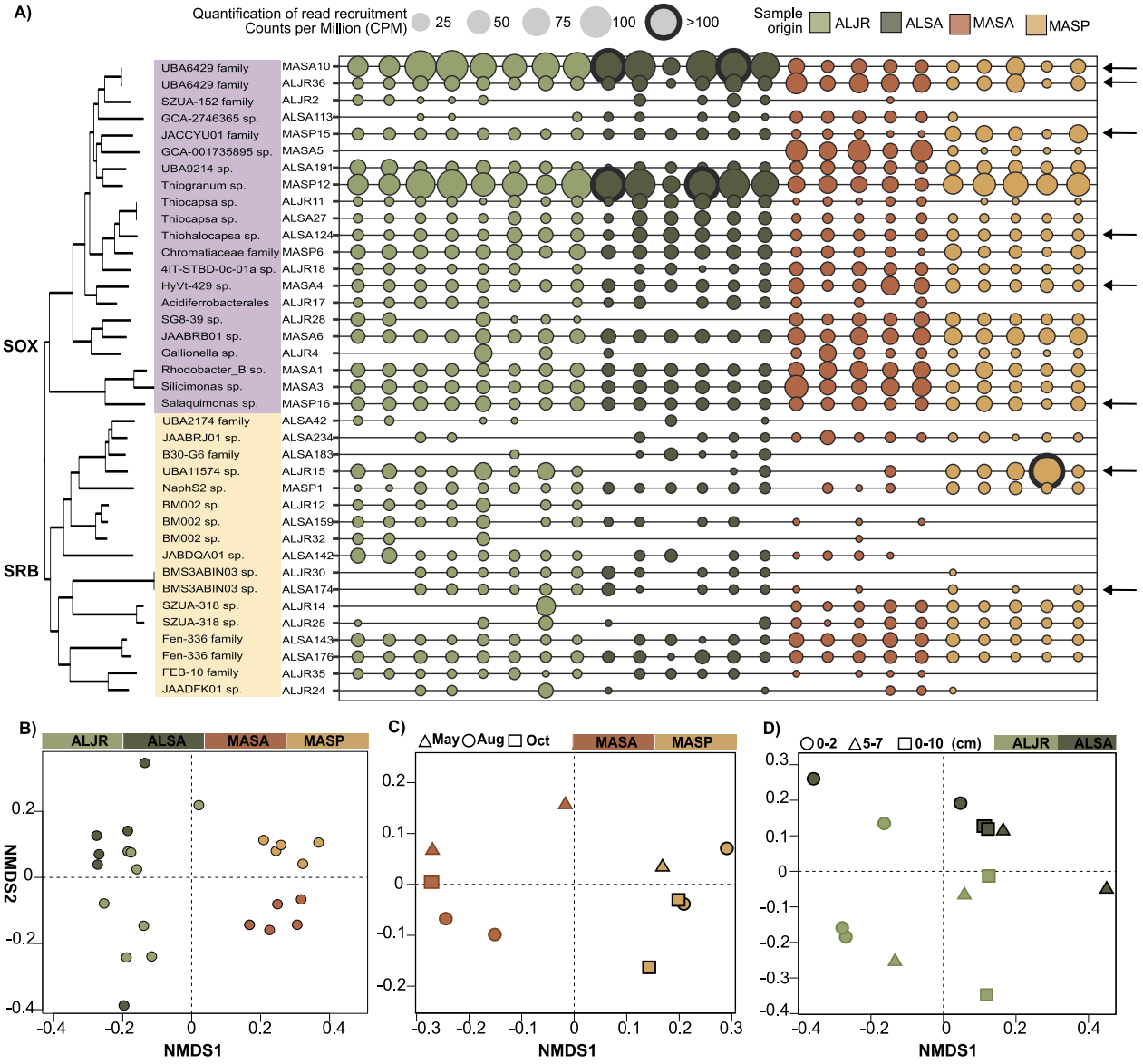


Figure 5 A) Anvi'o coverage profiles of eight selected sulfur-cycling bacteria MAGs. Outer cycles represent the contig coverage (as mean coverage) of each of the 24 salt-marsh metagenomic samples included in this analysis. Inner rings represent GC content (dark gray) and contig length (gray). Contigs are clustered (inner tree) based on the sequence composition and differential coverage using Euclidean distance and Ward hierarchical clustering method. Sample order (rings) was determined using a clustering method based on the mean coverage and each ring is color-coded according to site and vegetation type. Bar plots represent (top to bottom) the total number of reads in each library, the total number of reads mapped to each respective MAG, the percentage of mapped reads, and the total number of Single Nucleotide Variants (SNV).
B) Anvi'o pangenomic analysis of eight sulfur-cycling bacteria MAGs. In each ring, each vertical line represents a gene. Gene order was determined using Euclidean distance and Ward clustering method based on presence/absence of a gene cluster across MAGs. Annotation based on COG 2020 is indicated in gray and single-copy core genes in purple. Primary MAG is shown in dark teal blue. Each of the remaining circles represents a reference-guided reassembled MAG from each of the salt-mash metagenomic samples included in this analysis. Samples are color-coded according to geographical site and vegetation.

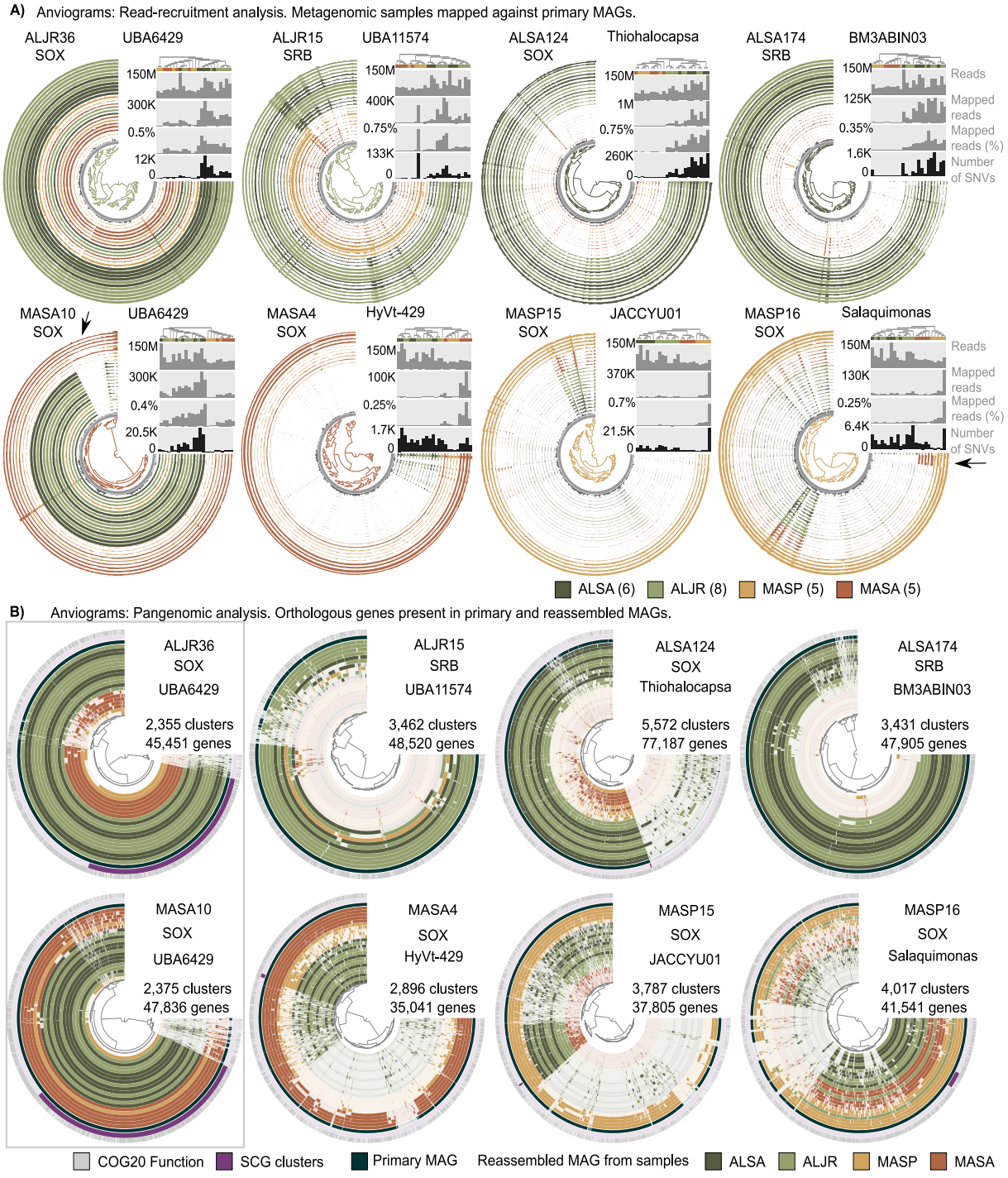
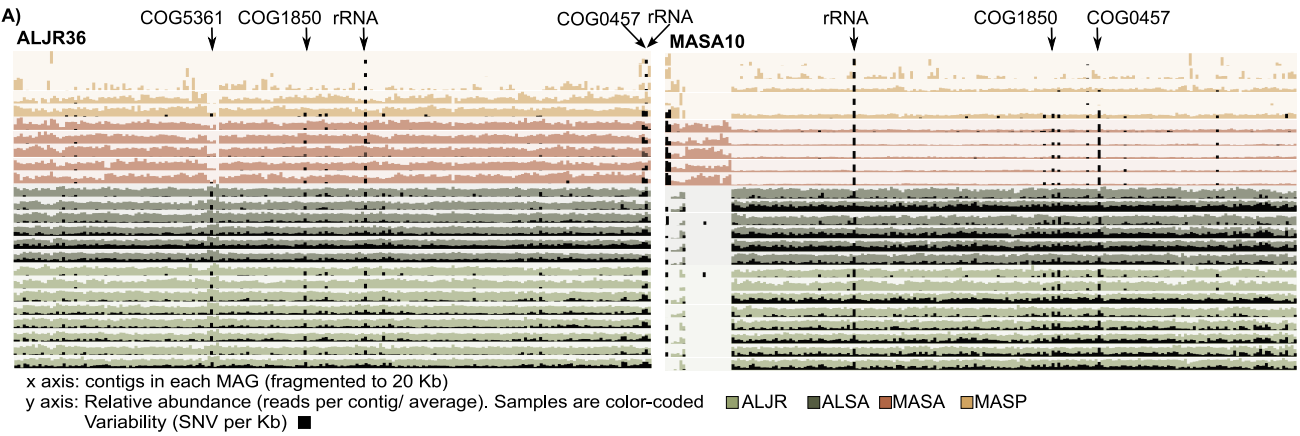
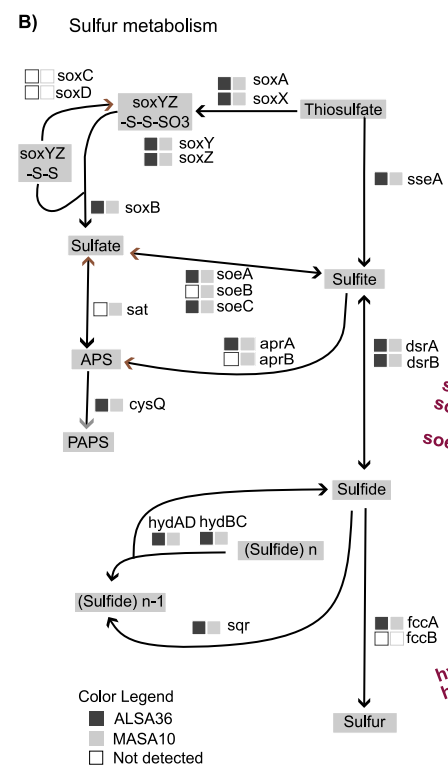
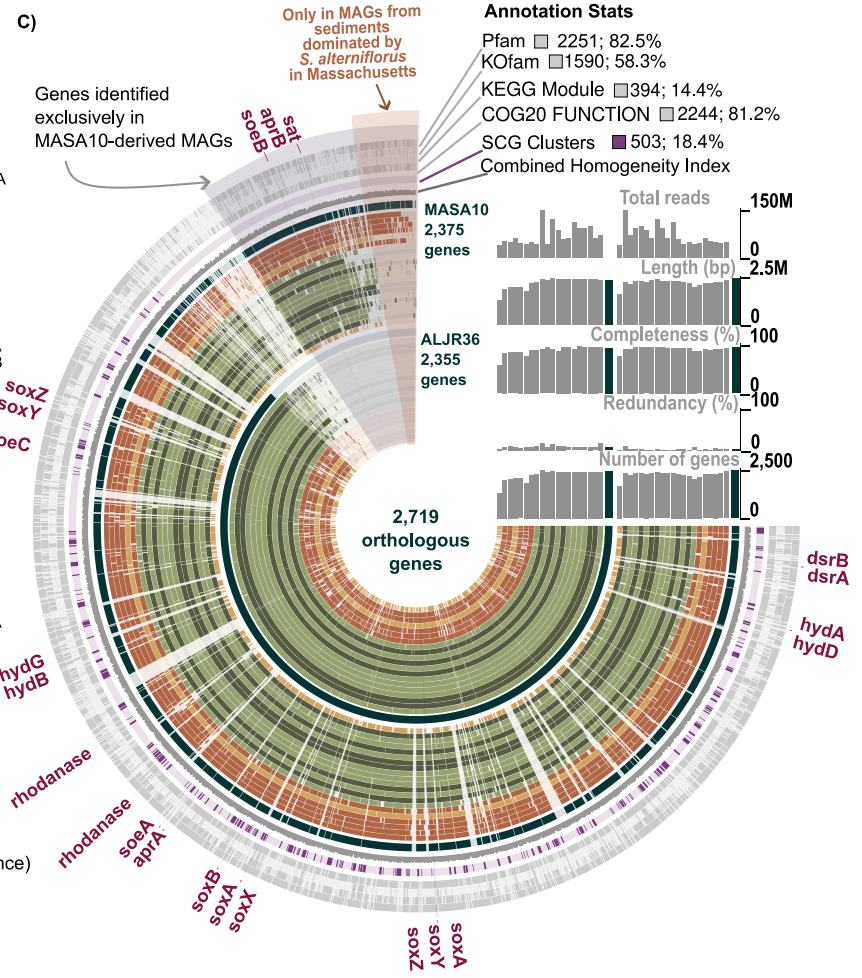


Figure 6 Comparative analysis of genomic and metabolic diversity of two closely-related sulfur-oxidizing MAGs (Thiohalomonadales) A) Distribution of genetic variability represented as SNV per kilobase (black bars) and calculated for each metagenomic sample. Hotspots of variability include COG5361(Mobilome), COG1850 (Rubisco-like protein), COG0457 (Tetratricopeptide (TPR) repeat), and rRNA. B) Diagram of sulfur metabolism pathways highlighting the differences in gene content between sample-specific MAGs from the ALJ36 and MASA10 primary MAG groups (adapted from Kegg map00920) C) Pangenomic analysis of metagenomes from group ALJR36 and MASA10. Primary MAGs are indicated by the dark ring and sample-specific re-assembled MAGs are color-coded by site and vegetation. Genes clusters are ordered with forced synteny to ALJR36 to highlight the genes exclusively found in the MASA10 group. Annotations, single-copy core genes, and the combined homogeneity index are shown in outer rings.







Pangenomics analysis (panel c, right)
View: clusters of orthologous genes (presence absence)
Order: Forced synteny (ALJR36)

Color Legend
Sulfur-cycling genes
Primary MAGs
ALJR
ALSA
MASA
MASP

Figure 7 Biogeographic and vegetation-specific distributions of genomic variability. A) Example aggregated view of metagenomic reads from different samples mapped to the MAG ALJR15. Samples collected in Alabama matched the reference while those collected in Massachusetts in sediments dominated by *S. pumilus* displayed numerous fixed mismatches. Mismatches are indicated with colors: green, T; yellow, G; blue, C; red, A. B) Trees depicting phylogenetic relationships among reassembled sample-specific metagenomes. Trees were built using single-copy genes, housekeeping genes, and sulfur-cycling genes. Tree branches are color-coded according to site and vegetation.

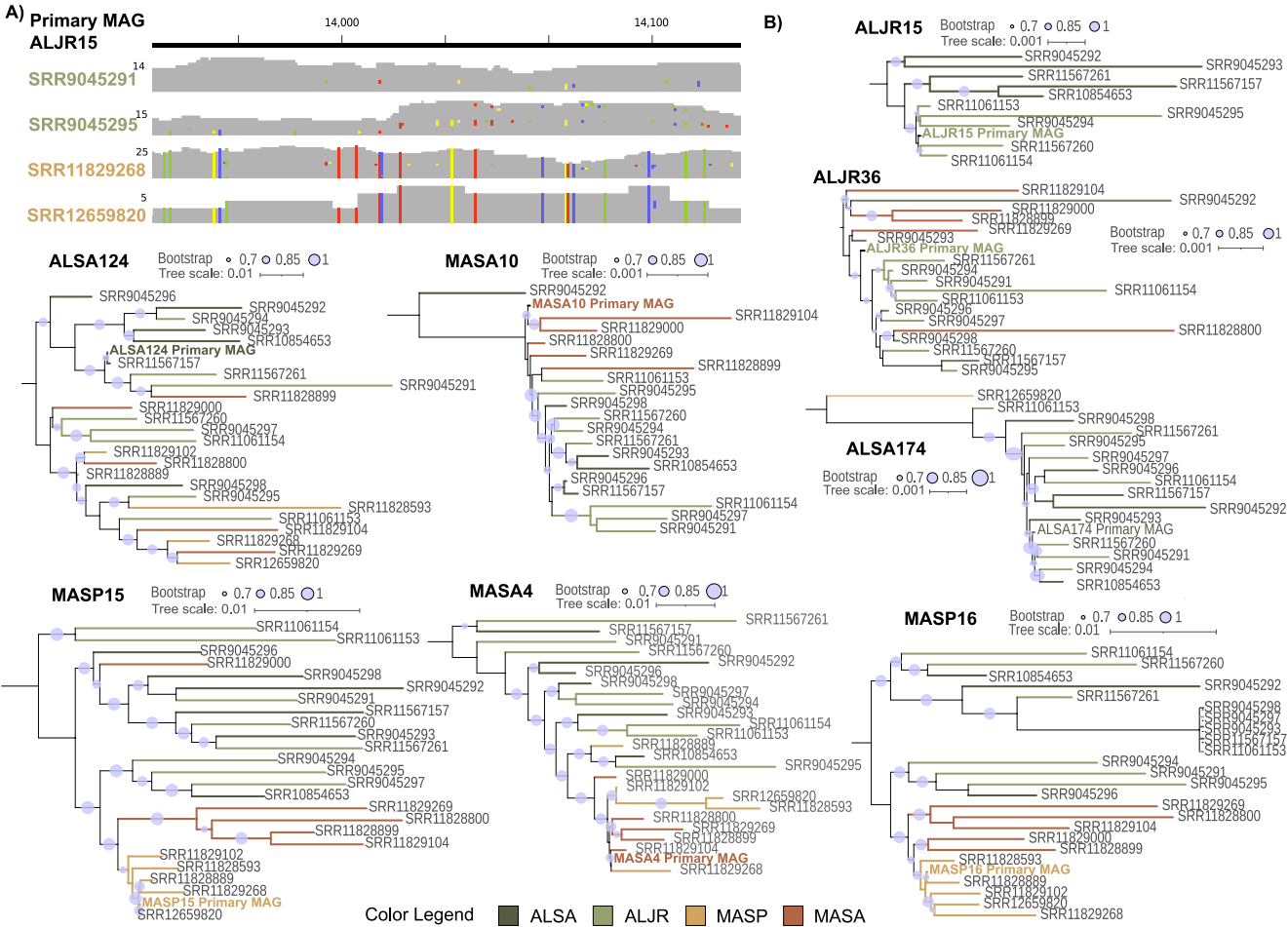
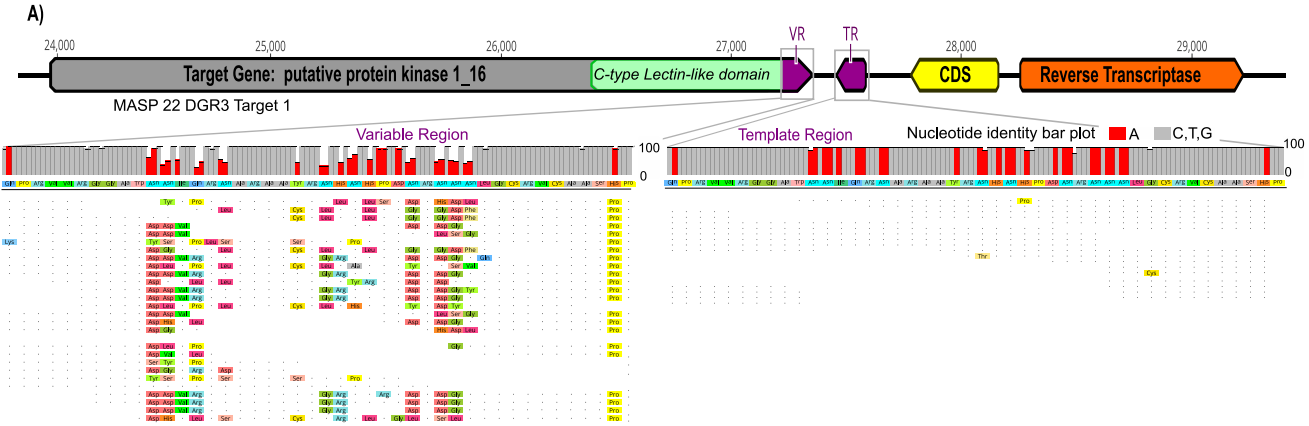
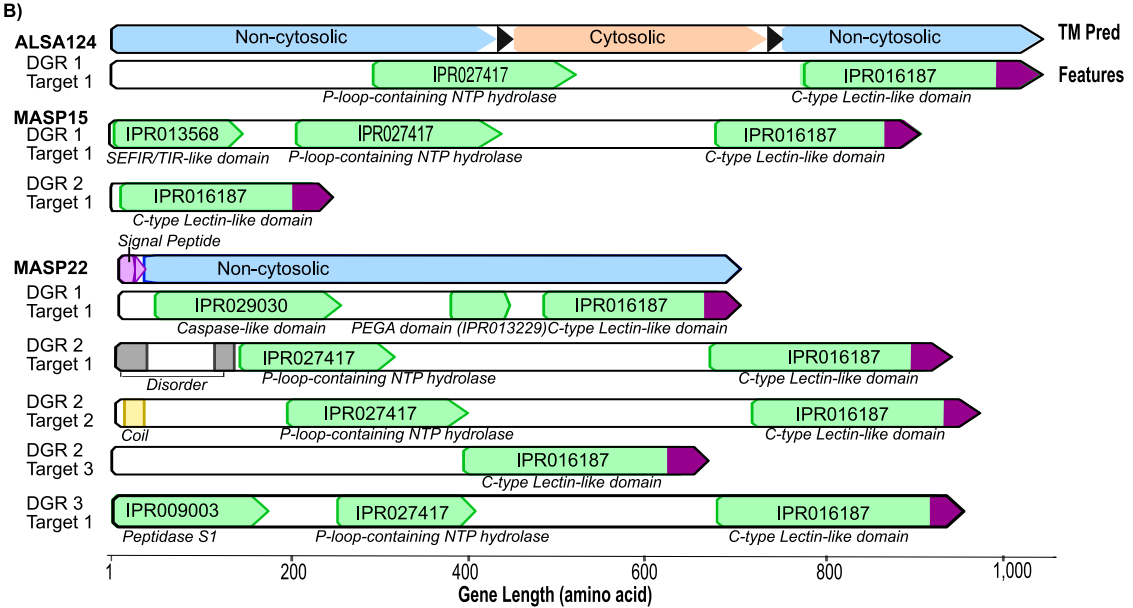


Figure 8 A) Structure of an example DGR identified in MASP22. Diversification of this DGR in this metagenomic sample is indicated by the accumulation of in-frame mutations in the variable region (VR) among the aligned reads. The template region (TR) is conserved both at the nucleotide and amino acid levels. In the bar graphs, the bar size is proportional to the degree of conservation of the nucleotide position. Adenine positions in the VR and TR regions are indicated in red. In the recruited reads, identical residues are indicated with dots, and changes are color coded.
B) Structure and domain composition of the target genes containing variable regions (in purple) of the DGRs.



Top: Translated amino acid sequence (aa) in reference MAG. Bottom. Variability detected in metagenomic population (aa). Identical aa are indicated with dots.