

# UC Berkeley

## UC Berkeley Previously Published Works

### Title

Diverse fates of ancient horizontal gene transfers in extremophilic red algae

### Permalink

<https://escholarship.org/uc/item/37w24455>

### Journal

Environmental Microbiology, 26(5)

### ISSN

1462-2912

### Authors

Van Etten, Julia  
Stephens, Timothy G  
Chille, Erin  
[et al.](#)

### Publication Date

2024-05-01

### DOI

10.1111/1462-2920.16629

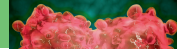
### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>


Peer reviewed

## BRIEF REPORT

## ENVIRONMENTAL MICROBIOLOGY



# Diverse fates of ancient horizontal gene transfers in extremophilic red algae

Julia Van Etten<sup>1</sup> | Timothy G. Stephens<sup>2</sup> | Erin Chille<sup>1</sup> | Anna Lipzen<sup>3</sup> | Daniel Peterson<sup>3</sup> | Kerrie Barry<sup>3</sup> | Igor V. Grigoriev<sup>3,4</sup> | Debashish Bhattacharya<sup>2</sup> 

<sup>1</sup>Graduate Program in Ecology and Evolution, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

<sup>2</sup>Department of Biochemistry and Microbiology, Rutgers, The State University of New Jersey, New Brunswick, New Jersey, USA

<sup>3</sup>U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, California, USA

<sup>4</sup>Department of Plant and Microbial Biology, University of California, Berkeley, Berkeley, California, USA

## Correspondence

Julia Van Etten, Graduate Program in Ecology and Evolution, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA.

Email: [julia.vanetten@rutgers.edu](mailto:julia.vanetten@rutgers.edu)

Debashish Bhattacharya, Department of Biochemistry and Microbiology, Rutgers, The State University of New Jersey, New Brunswick, NJ 08901, USA.

Email: [dbhattac@rutgers.edu](mailto:dbhattac@rutgers.edu)

## Funding information

NIFA-USDA, Grant/Award Number: NJ01180; National Aeronautics and Space Administration, Grant/Award Numbers: 80NSSC19K0462, 80NSSC19K1542; U.S. Department of Energy, Grant/Award Number: DE-AC02-05CH11231

## Abstract

Horizontal genetic transfer (HGT) is a common phenomenon in eukaryotic genomes. However, the mechanisms by which HGT-derived genes persist and integrate into other pathways remain unclear. This topic is of significant interest because, over time, the stressors that initially favoured the fixation of HGT may diminish or disappear. Despite this, the foreign genes may continue to exist if they become part of a broader stress response or other pathways. The conventional model suggests that the acquisition of HGT equates to adaptation. However, this model may evolve into more complex interactions between gene products, a concept we refer to as the 'Integrated HGT Model' (IHM). To explore this concept further, we studied specialized HGT-derived genes that encode heavy metal detoxification functions. The recruitment of these genes into other pathways could provide clear examples of IHM. In our study, we exposed two anciently diverged species of polyextremophilic red algae from the *Galdieria* genus to arsenic and mercury stress in laboratory cultures. We then analysed the transcriptome data using differential and coexpression analysis. Our findings revealed that mercury detoxification follows a 'one gene-one function' model, resulting in an indivisible response. In contrast, the *arsH* gene in the arsenite response pathway demonstrated a complex pattern of duplication, divergence and potential neofunctionalization, consistent with the IHM. Our research sheds light on the fate and integration of ancient HGTs, providing a novel perspective on the ecology of extremophiles.

## INTRODUCTION

Horizontal genetic transfer (HGT), once deemed highly unlikely in eukaryotes due to the presence of a nuclear envelope and in many cases, a sequestered germline, is now accepted as a common feature across the tree of life (Van Etten & Bhattacharya, 2020). Yet, the mechanisms that drive HGT persistence and integration into host pathways are poorly understood. The

Cyanidiophyceae is a group of polyextremophilic red algae that dominate geothermal environments, often comprising >90% of the biomass in these locations that are characterized by high temperature, extremely low pH, fluctuating light, high salt and high toxic heavy metal content (Reeb et al., n.d.; Castenholz & McDermott, 2010; Doemel & Brock, 1971; Seckbach, 1972; Van Etten, Cho, et al., 2023). These algae have become models for studying eukaryotic HGT due to

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Environmental Microbiology* published by John Wiley & Sons Ltd.



their simple genome structure (e.g., highly reduced genome size, limited non-coding DNA and few introns) and the presence of many adaptive HGTs acquired from prokaryotes, including those that encode heavy metal detoxification (Qiu et al., 2013; Rossoni et al., 2019; Schönknecht et al., 2013; Schönknecht et al., 2014). Most of these genes are of ancient provenance, with many acquired over a billion years ago in the ancestor of this lineage. Whereas HGT in the prokaryotic domains is understood to be widespread and continual (Jain et al., 1999; Koonin, 2016; Rivera et al., 1998), only in recent studies has eukaryotic HGT also been shown to be a common and continuous process, for example, in grasses (Pereira et al., 2022), and various microbial eukaryotes (Alsmark et al., 2013; Huang, 2013; Nowack et al., 2016). Therefore, the principles underlying the process of prokaryote HGT and gene retention may also apply to eukaryotes. Analysis of the transcription patterns, regulation, and integration of HGTs into eukaryotic metabolic networks is needed to address this issue.

In prokaryotes, genes for mercury and arsenic detoxification are encoded in operons (Ben Fekih et al., 2018; Boyd & Barkay, 2012; Yang et al., 2012), the latter controlled by the *ArsR* transcriptional repressor (Francisco et al., 1990). Many of the genes in these operons were transferred via HGT into different eukaryotic lineages (Chen et al., 2017; Marcet-Houben & Gabaldón, 2010; Palmgren et al., 2017; Ribeiro & Lahr, 2022). However, there is no evidence in any of these cases that a whole operon was transferred or retained; rather, single genes were likely to have been acquired: that is, *arsR* is missing from these species, the HGTs show no clear pattern of colocalization within the genome, and the putative donor lineage of each gene often differs, suggesting multiple, independent origins (see Discussion). The Cyanidiophyceae live in association with many extremophilic prokaryotes, sometimes in biofilms where HGT is hypothesized to occur more frequently due to the proximity of organisms (Alsmark et al., 2013; Lehr et al., 2007; Soucy et al., 2015). Levels of heavy metals in geothermal habitats have fluctuated greatly over time and pose a threat to cells not able to detoxify these poisons (Chen et al., 2017; Christakis et al., 2021; Fru et al., 2019; Zhu et al., 2014). Across the phylogeny of Cyanidiophyceae, a single protein from the bacterial *mer* operon, MerA (mercury(II) reductase), is implicated in the mercury stress response and is present in every sequenced genome, although it is the result of two independent transfers among these taxa (Rossoni et al., 2019). Likewise, there exist several genes in Cyanidiophyceae that originated from the prokaryotic *ars* operon that deal directly with arsenic: *ArsA* (arsenical pump-driving ATPase), *ArsB* (arsenite efflux transporter), *ArsC* (arsenate reductase), *ArsM* (arsenite

methyltransferase) and *ArsH* (arsenical resistance protein). Previous phylogenetic work has demonstrated the sporadic distribution of these prokaryote-derived genes in the orders Galdieriales and Cyanidiales within Cyanidiophyceae, suggesting they result from independent acquisitions (Rossoni et al., 2019; Schönknecht et al., 2013; see Figure 1).

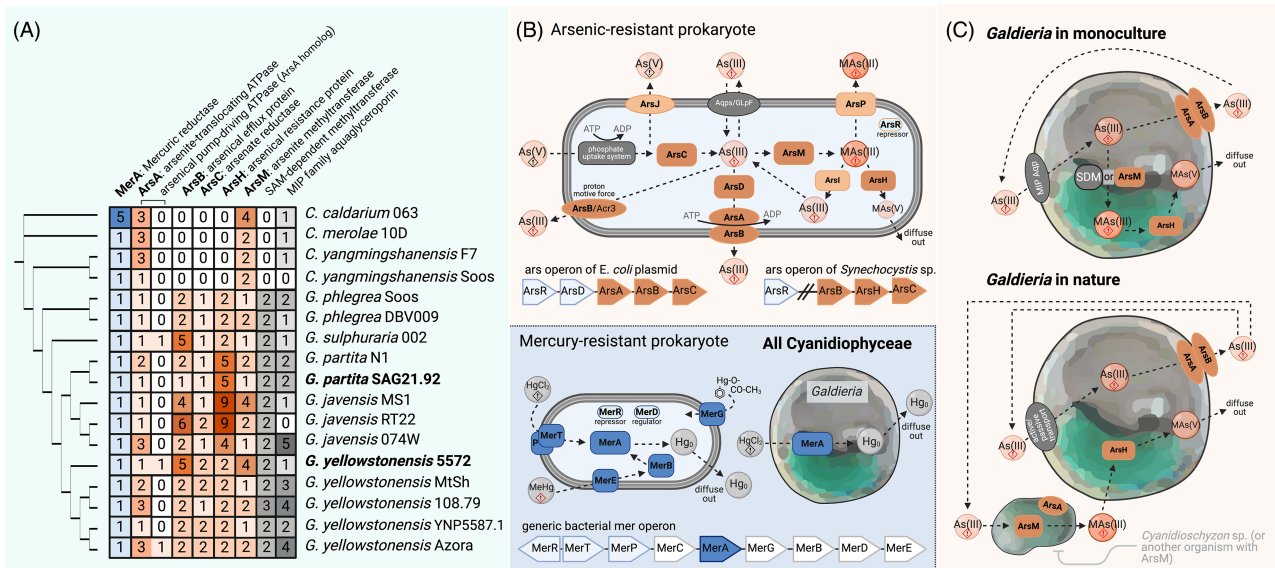
In this study, we investigate the effects of mercuric chloride ( $\text{HgCl}_2$ ) and sodium arsenite ( $\text{NaAsO}_2$ ) on the transcriptional dynamics of HGT-derived mercury and arsenic detoxification genes in two anciently diverged *Galdieria* species. These results were used to explore the evolutionary history of heavy metal detoxification in Cyanidiophyceae, and the divergent fates of HGT-derived genes in two distinct resistance pathways.

Our results show that *merA*, which is part of a relatively simple detoxification pathway (requiring a single gene), has a putative stress-driven transcriptome response. In contrast, the *ars* genes, which have distinct but related efflux and detoxification functions, have complex, and often divergent stress-driven transcriptomic responses, and are integrated into metabolic networks associated with diverse functions. For foreign sequences to persist for millions or billions of years, they must provide a long-term adaptive advantage at (and after) acquisition. As environmental conditions change over time, the selective pressures that drive the acquisition of HGT-derived genes will likely fluctuate or disappear, yet these genes may survive if they have become integrated into other systems that are critical to cell survival (Burch et al., 2023; Huang, 2013; Jones et al., 2022). In other words, the standard model, whereby HGT acquisition equals adaptation, which has been frequently validated, may evolve into a more complex set of gene interactions which we refer to here as the ‘integrated HGT model’ (IHM). This latter outcome is challenging to study in the absence of HGT-derived genes involved in highly specialized functions (e.g., metal detoxification), whereby their recruitment into other pathways is easier to interpret and may be explained by the IHM. We explore the predictions of this model by studying the fate and integration of ancient HGTs. These data lead us to propose a hypothetical framework to explain how the arsenic pathway functions among Cyanidiophyceae, thereby providing a novel perspective on the ecology of extremophiles.

## EXPERIMENTAL PROCEDURES

### RNA-seq experiments

Based on preliminary growth experiments summarized in Appendix (see [Preliminary growth experiments](#)



**FIGURE 1** Distribution and expression of HGT-derived genes in the Cyanidiphyceae that are involved in arsenic and mercury detoxification. (A) Cladogram of the 17 available genomes and the presence (coloured boxes) and copy number of relevant genes (number in box). (B) Hypothetical prokaryotic cells showing schematic arsenic (top) and mercury (bottom) detoxification pathways. Each image includes all enzymes; however, it is important to note that not all prokaryotes contain these genes. Below each hypothetical cell are examples of real operons for each gene cluster. For arsenic, we have included the *arsRABCD* operon in the *Escherichia coli* plasmid and the *arsRBCH* operon in *Synechocystis* sp. (Kalia & Joshi, 2009). For mercury, a hypothetical operon with all possible mercury genes is based on Boyd and Barkay (2012), with genes found in all *mer* operons shown in light blue (and dark blue, MerA) and genes found in some operons in white. On the right of these hypothetical cells (B and C) are schematic *Galdieria* depicting their mercury and arsenic resistance pathways. (C) *Galdieria* cells represent the two possible detoxification pathways discussed in the paper for cultured cells (top) and in nature, based on community dynamics (bottom). Note that the copy number of the genes encoding these enzymes varies between *G. partita* SAG21 and *G. yellowstonensis* 5572 (see left panel). Image made in Biorender.com. HGT, horizontal genetic transfer.

methods and results and Figures A1–A5), 5 mM NaAsO<sub>2</sub> and 3 μM HgCl<sub>2</sub> were chosen as treatment concentrations for the RNA-seq experiments. *Galdieria yellowstonensis* 5572 and *G. partita* SAG21 (hereafter 5572 and SAG21; both of these species were formerly classified as strains of *G. sulphuraria*; Park et al., 2023) were acclimated in batch cultures to the control media conditions for 1 month prior to the experiments. Twelve 500-mL flasks were prepared with either 2× modified Allen medium with 25 mM glucose at pH 2 (control) or this medium plus 5 mM NaAsO<sub>2</sub> or 3 μM HgCl<sub>2</sub> (treatments) (Allen, 1959). At the start of the experiment, an equal amount of algal biomass was added to each flask (Figure A6) which was incubated at 42°C, 40 rpm, in 90 μE m<sup>-2</sup> s<sup>-1</sup> continuous white light. Sampling took place at four-time points in 1 week: 1, 24, 72 and 168 h (TP1–TP4), which was determined based on changes in OD recorded during the preliminary experiments; see Spreadsheet S1: <https://zenodo.org/doi/10.5281/zenodo.8377091> and Figures A2–A5. During sampling, flasks were removed from the incubator and briefly placed in a biosafety cabinet where 50 mL aliquots from each flask were added to 50 mL Falcon tubes, centrifuged at 4000 rpm for 5 min, and then the supernatant was removed. The remaining cell pellets were flash-frozen in liquid nitrogen and stored at –80°C.

## Sample preparation and sequencing

Following completion of the experiments, samples were homogenized via bead-beating (vortexed at the highest speed for 10 min) with 0.5 mm silica beads. RNA was extracted using the Qiagen (Hilden, Germany) RNeasy Plant Mini Kit and cleaned using the Zymo (Orange, California, USA) RNA Clean and Concentrator-25 kit. Samples were checked for quality, purity, and concentration using the Nanodrop 3000C and Qubit 2.0. They were stored at –80°C and shipped on dry ice to the Joint Genome Institute (JGI) in California, USA for sequencing. There, plate-based RNA sample prep was performed on the PerkinElmer Sciclone NGS robotic liquid handling system using Illumina's TruSeq Stranded mRNA HT sample prep kit utilizing a poly-A selection of mRNA following the protocol outlined by Illumina in their user guide: [https://support.illumina.com/sequencing/sequencing\\_kits/truseq-stranded-mrna.html](https://support.illumina.com/sequencing/sequencing_kits/truseq-stranded-mrna.html), and with the following conditions: total RNA starting material was 1000 ng per sample and eight cycles of PCR was used for library amplification. The prepared libraries were then quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the flowcell was performed on the

Illumina NovaSeq sequencer using NovaSeq XP V1.5 reagent kits, S4 flowcell, following a 2 × 151 indexed run recipe.

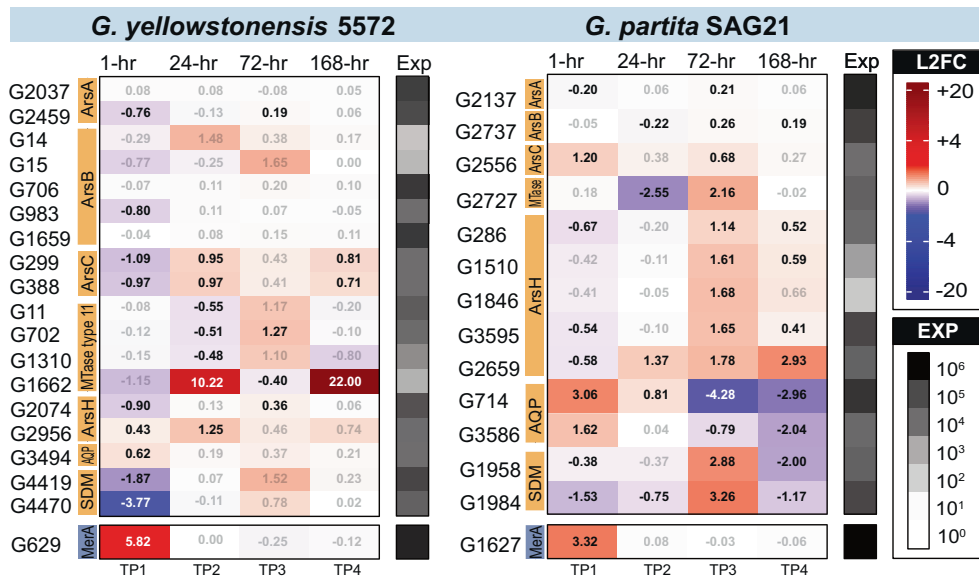
## Differential expression analysis

Raw sequencing reads were filtered and trimmed using the JGI QC pipeline (see Appendix [JGI QC pipeline](#)). Interleaved quality trimmed reads were separated using BMap (Bushnell, 2014). The reference genomes for 5572 and SAG21 (Rossoni et al., 2019) were indexed, reads were mapped to them using HISAT2 (Kim et al., 2019), and transcripts were built using existing gene models for SAG21 and 5572 with StringTie2 v2.2.1 (Kovaka et al., 2019). The performance of the assemblies was assessed using gffcompare (Pertea & Pertea, 2020). Compilation of gtf files into gene and transcript count matrices was done using the prepDE.py script available with StringTie2. Next, count matrices were imported into RStudio 2022.02.3+492 using R version 4.2.3 and differential expression (DE) analysis was carried out using DESeq2 v1.38.3 (Love et al., 2014; R code in Code1, PCA plot of treatment-time point clustering in Figure A7). Count matrices were separately normalized by two methods, transcripts per million (TPM) and ‘ratio of means’ (via DESeq2), so that expression profiles for genes of interest could be generated and compared (see Figure 2 and Spreadsheet S2: <https://zenodo.org/doi/10.5281/zenodo.8377091>).

8377091). The parameters used to call a gene DE between conditions were  $p$  value <0.05 (FDR-adjusted) and an absolute  $\log_2$  fold change (FC) of either 1, 1.5 or 2 (results for each were considered during analysis and are highlighted in Spreadsheet S3: <https://zenodo.org/doi/10.5281/zenodo.8377091>).

## Weighted gene coexpression network analysis

To gain a different perspective on the gene expression data and to determine if, and how genes of interest are integrated into native metabolic networks, we used the weighted gene coexpression network analysis (WGCNA) data-reduction technique (Chille et al., 2021). WGCNA was run in RStudio, employing functions from the following packages: DESeq2, genefilter, RcolorBrewer, Biobase, GO.db, impute, ComplexHeatmap, goseq, ClusterProfiler, simplifyEnrichment, tidyverse, flashClust, gridExtra, dplyr and WGCNA (Langfelder & Horvath, 2008, 2012; R code in Codes 2 and 3). Data sets were filtered using the PoverA function in genefilter and checked for quality (Gentleman et al., 2015). The smallest number of replicates was three for SAG21 and two for 5572; thus, genes with fewer than 10 counts in at least 3/48 SAG21 samples (pOverA 0.0625, 10) and 2/475572 samples (pOverA 0.043, 10) were excluded from the analysis. Reads were normalized using the variance stabilizing



**FIGURE 2** Differential gene expression profiles, normalized via the ratio of means (DESeq2 normalization method) for each arsenic or mercury gene of interest. The left heatmap is *Galdieria yellowstonensis* 5572 and the right heatmap is *G. partita* SAG21. Colours within the heatmaps indicate  $\log_2$  fold change (L2FC) values for genes at the indicated time points (TP1–TP4), that is, comparing treatment (i.e., arsenite or mercury) to control cultures. Significant ( $p_{adj} < 0.05$ ) FC values are black and in bold, whereas nonsignificant values are in grey. The grayscale heatmap to the right of each L2FC heatmap shows variation in the magnitude of (normalized) read counts averaged across both control and treatment libraries. This gradient indicates which genes (regardless of differential expression) are generally most highly expressed.



transformation (vst) in DESeq2 after confirming all size factors were less than 4 (one replicate As\_R-NA\_5572\_EXP\_4D was 5.17, but we did not exclude it). After this, the count data were log-transformed. A principal component analysis (PCA) based on sample-to-sample distances was performed on the vst-transformed gene counts for each genome using the *plotPCA* function in DESeq2 (Figures A8 and A9), and an unrooted hierarchical tree (Figure A10) was used to visualize experiment-wide patterns in gene expression and to check for outliers using the R stats *hclust* 'average' function. To assess gene expression adjacency, we constructed a topological overlap matrix similarity network using the WGCNA *pickSoftThreshold* function, displaying soft threshold values from 1 to 20. Soft thresholding powers of 8 for SAG21 and 12 for 5572 were chosen (scale-free topology fit index of 0.8 and 0.65, respectively) and used to construct a topological overlap matrix similarity network using a signed adjacency (Figures A11 and A12). Modules were identified from this network using the *dynamicTreeCut* function, with the settings of *deepSplit* = 2, and minimum module size = 30 (Figure A13). The modules were used for expression plotting, module-trait correlation (Figures A14–A16) and network visualization in Cytoscape v3.10.0 (Shannon et al., 2003; Figure 3).

The final step of this analysis was to generate a heatmap that showed modules with significant up- or downregulation in both experimental conditions (for each genome; Figure A17). We also generated a matrix with genes (nodes), their annotations, and which module they belong to; see Spreadsheet S4: <https://zenodo.org/doi/10.5281/zenodo.8377091>. This includes information on time point significance for each gene (and *p* values), as well as module membership (MM) values per gene in each module (and accompanying *p* values). For downstream analysis, any gene-module correlation that had an adjusted *p* value >0.05 was discarded. We also subsequently filtered each dataset by MM, using a cutoff value of 0.8. Thus, only those genes (nodes) most significantly correlated with their assigned module were retained for analysis. Edge weight thresholds vary across the literature and depend more on the dataset and network than the threshold value itself (Yan et al., 2018). In this case, using Cytoscape, edge weight thresholds were selected for each module independently by determining which cutoff results in unconnected singletons, and then reducing the cutoff by 1% until a streamlined but representative network was produced. Figure 3 (and Figure A18) shows genes from Figure 2 that met these thresholds.

## Phylogenetic analysis of HGTs

For each arsenic and mercury detoxification-related gene, amino acid sequences were retrieved from all

available Cyanidiophyceae genomes and used to query the NCBI nr database using diamond blastp v2.1.2.156 (Buchfink et al., 2021). All BLAST hits with <40% identity and an *e*-value >1e−10 were discarded. Next, the (often tens of thousands of hits to the nr database) were taxonomically downsampled to produce a smaller, more computationally tractable, representative set that could be used for phylogenetic analysis. A custom script v0.1 was used to retrieve at most, two highest-scoring subject sequences per phylum per query sequence based on the taxonomy assigned to the subject sequences in NCBI. The set of downsampled sequences for each set of proteins was aligned using MAFFT-linsi v7.490 (Kato & Standley, 2013) and maximum likelihood (ML) trees were constructed in IQ-TREE v1.6.12 using automated model selection and node support estimated from 1000 ultrafast bootstrap replicates (Hoang et al., 2018; Nguyen et al., 2015).

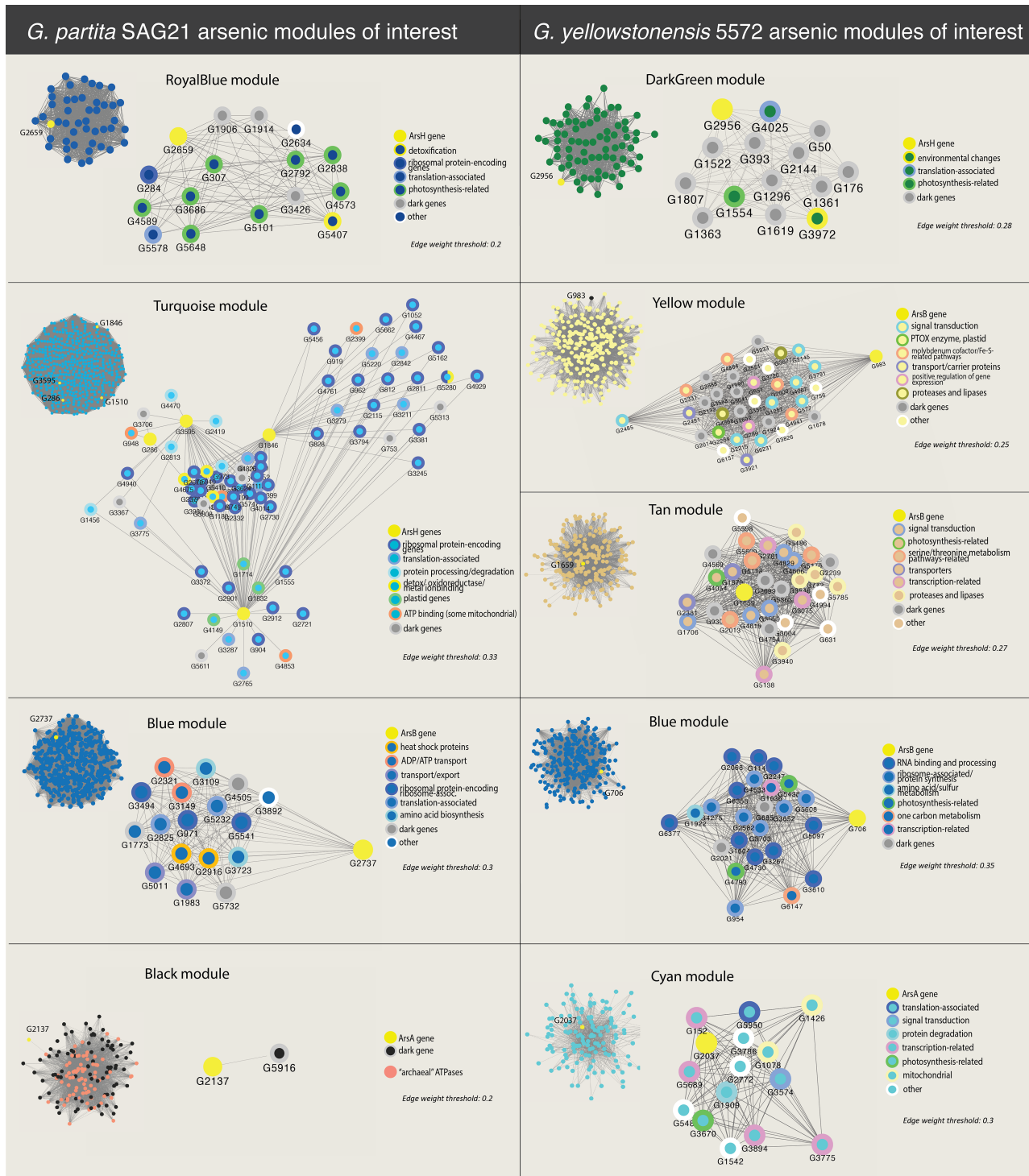
## ArsH protein visualization

To allow structural and functional analysis of ArsH proteins, we aligned all SAG21 and 5572 amino acid sequences, along with those from three bacteria including *Paracoccus denitrificans*, which has one of the only resolved ArsH protein structures in PDB (7PLE; Sedláček et al., 2022). We built the multiple sequence alignment (MSA) using CLUSTAL O v.1.2.4 (Sievers et al., 2011).

## RESULTS

### Differential expression analysis

DESeq2 analysis resulted in sets of differentially expressed genes (DEGs) for each treatment and accompanying time point. All DEGs for various FC values are shown in Table 1, with the full list of DESeq2 output available in Spreadsheet S3: <https://zenodo.org/doi/10.5281/zenodo.8377091>. Table 1 indicates that the magnitude of the arsenite response, in terms of differential transcription, is about 10-fold greater than that of mercury in both *Galdieria* species. *MerA* showed the highest degree of differential upregulation (i.e., it had the highest FC value) across both genomes at TP1 (FC = +3.32 in SAG21, FC = +5.82 in 5572) followed by an immediate return to baseline once (presumably) the mercury was detoxified (Figure 2). The *ars* genes generally show constitutive expression, except *arsH* (and putatively *arsM*), which we discuss below. All arsenic and mercury-related genes of interest and their differential and overall expression profiles and FC values (black or white if significant [ $p_{adj} < 0.05$ ]) are shown in Figure 2.



**FIGURE 3** Networks for heavy metal genes of interest based on WGCNA analysis. On the left are genes from *Galdieria partita* SAG21 and on the right are from *G. yellowstonensis* 5572. Note that not all genes from Figure 2 are explored here. Genes were excluded if they did not pass module membership ( $>0.8$ ) filtering. The small insets on the top right of each box indicate the filtered whole module and the larger networks are the subnetworks of all genes directly connected to each gene of interest at the specified edge weight threshold. Nodes represent genes and edges represent linkage determined by WGCNA. Nodes are coloured based on shared function. A full list of modules and their functional trends can be found in Figure A18. WGCNA, weighted gene coexpression network analysis.

## Heavy metal transport and methylation

In addition to the *merA* and *ars* genes that were the targets of this study, we identified other HGTs involved in

transport and methylation that are possible contributors to heavy metal detoxification and were included in the phylogenetic analysis. In 5572 and SAG21, we identified one and two MIP family aquaglyceroporins,



**TABLE 1** Differential expression results summary of the data generated in DESeq2. (A) shows the number of upregulated DEGs in both species across both conditions and (B) shows the downregulated DEGs.

<b>(A) Genes upregulated (treatment vs. CTRL).</b>													
Organism	Condition	TP1			TP2			TP3			TP4		
		FC (+)	2	1.5	1	2	1.5	1	2	1.5	1	2	1.5
Genes upregulated (treatment vs. CTRL)													
5572	NaAsO <sub>2</sub>	200	317	558	58	94	198	24	45	101	27	47	76
SAG21	NaAsO <sub>2</sub>	290	503	832	67	103	193	164	311	698	123	188	341
5572	HgCl <sub>2</sub>	11	19	49	9	9	10	14	20	25	4	4	6
SAG21	HgCl <sub>2</sub>	29	58	132	64	112	167	12	14	24	10	13	16
<b>(B) Genes downregulated (treatment vs. CTRL)</b>													
Organism	Condition	TP1			TP2			TP3			TP4		
		FC (-)	2	1.5	1	2	1.5	1	2	1.5	1	2	1.5
Genes downregulated (treatment vs. CTRL)													
5572	NaAsO <sub>2</sub>	17	37	147	18	37	161	66	98	158	12	24	41
SAG21	NaAsO <sub>2</sub>	27	73	366	55	105	204	250	395	663	91	140	244
5572	HgCl <sub>2</sub>	14	25	91	4	6	9	45	62	95	5	6	11
SAG21	HgCl <sub>2</sub>	9	12	26	18	28	44	17	20	44	13	16	22

*Note:* For each time point, DEGs with log<sub>2</sub>-fold change (FC) of ±1, 1.5 and 2 are shown. The mercury response is roughly one order of magnitude less than the arsenite response based on the quantities of DEGs shown in this table.

Abbreviation: DEG, differentially expressed gene.

respectively, that are likely importing arsenite based on their differential upregulation at TP1 and subsequent downregulation at later time points (Figure 2), and their previous identification as HGTs in related isolates (Rossoni et al., 2019; Schönknecht et al., 2013). Two methyltransferases in SAG21 (G1984 and G1958) show patterns under arsenite stress that may indicate they provide this function, albeit in a manner that is more difficult to interpret than the *arsH* gene expression patterns. Specifically, both the SAG21 methyltransferases are upregulated at TP2 in the control cultures and decline thereafter, whereas G1984 peaks at TP3 and stays high at both TP2 and TP3. Therefore, these genes may act as methylating agents of Ars(III) to MAs(III) (Figures 1 and 2; see also Schönknecht et al., 2013). We have also listed some putative *arsM* candidates (annotated as ‘methyltransferase type 11’). These genes were identified by Cho et al. (2023) as putative *arsM* homologues via orthogroup analysis. However, due to the phylogenetic breadth of this gene family and the well-conserved nature of methyltransferase activity, it is difficult to determine, using sequence data alone, which of these genes may be acting as *arsM*. Whereas they show differential expression at TP2 or TP3 in both algal species (Figure 2), the FC values are largely nonsignificant, and they have extremely low MM based on the WGCNA analysis (see below). In contrast, the SAM-dependent methyltransferases are more strongly correlated with *arsH* expression. Gene knockdown experiments are needed to assess the function of the putative ArsM-encoding

genes in these *Galdieria* species. All of these genes and their copy numbers in the Cyanidiophyceae tree are shown in Figure 1A.

## Coexpression networks link arsenic HGTs to diverse metabolic functions

We used WGCNA to identify groups of coexpressed genes and focused on networks that include arsenic and mercury-related HGTs. Each algal species (5572 and SAG21) was analysed separately, and the response from both the arsenite and mercury experiments was taken together to inform per-genome coexpression. In this analysis, modules refer to clusters of ≥30 closely interconnected genes in a coexpression network. Clusters of 30 were chosen as the threshold because this value resulted in discrete gene clustering based on TOM-based similarity for both genomes (Figure A12). This value has been established as being suitable by Langfelder and Horvath in other data sets that included taxonomically diverse species (Langfelder & Horvath, 2008; Langfelder & Horvath, 2012; Langfelder & Horvath, 2016). For 5572 and SAG21, 27 and 20 modules were identified (respectively; Figures A13–A17). Of these, only Turquoise, RoyalBlue, Black, Blue, Cyan, Tan and Brown in SAG21 and DarkGreen, Yellow, Tan, Blue, Cyan and LightYellow in 5572 include the target HGT genes. As shown in Figure 3 (and Figure A18), each of these modules is associated with characteristic





functions, some more frequent in the module than others.

In SAG21, the *arsA* gene falls into the black module with many other ATPases, namely the ‘archaeal ATPases’ (Schönknecht et al., 2013). However, the SAG21 *arsA* gene is on the periphery of this module and has lower connectivity than the other ATPase genes. This suggests that SAG21 *arsA* may retain the ATPase function, but has a specialized role, putatively to allow arsenite efflux (Rosen et al., 1990). Interestingly, in 5572, *arsA* is not associated with the other ATPases, but rather with genes involved in transcription, translation, signal transduction, protein degradation and organelle-associated functions. The SAG21 *arsB* gene is in the Blue module which primarily comprises heat shock, transport and translation-associated proteins. One *arsB* paralogue in 5572 (G706) is in the Blue module which has a similar functional profile to the module containing the SAG21 copy (also Blue). The other two 5572 *arsB* catalogues (G983 and G1659) are in the Yellow and Tan modules, respectively, which comprise genes associated with signal transduction, photosynthesis, transport, protein and lipid degradation, and unknown functions. Four *arsH* genes from SAG21 are in the Turquoise module which is composed almost exclusively of genes encoding ribosomal proteins and genes with translation-associated functions, although G286 is not tightly linked to many other genes and G1846 is associated with a more diverse array of functions including photosynthesis (see below). The fifth *arsH* paralogue (G2659) is in the RoyalBlue module which is dominated by photosynthesis-related genes. In 5572, the only *arsH* paralogue to pass filtering (G2956; homologue of G2659 in SAG21) is in the DarkGreen module, which is largely comprised of ‘dark’ genes with unknown function. In SAG21, two SAM-dependent methyltransferase HGTs (G1984 and G1958) are in the Cyan and Tan modules, respectively. Both modules are composed primarily of genes associated with redox processes, RNA and protein metabolism, and energy generation. *G. yellowstonensis* 5572 also has two homologues: G4419 and G4470, although the latter is annotated as dimethylglycine N-methyltransferase. Both genes are in the LightYellow module (although G4419 did not pass MM filtering) which is composed of genes encoding oxidoreductase and those involved in post-translational modification, lipid synthesis, and cell cycle regulation. Only one MIP family aquaglyceroporin (G714 in SAG21 [Brown module]) had adequate MM (>0.8) across both genomes. This gene is coexpressed with a handful of dark genes, oxidoreductases, cell signalling and transport genes, as well as others related to DNA and RNA processes. Finally, *arsC* was analysed as a type of control for our experiment, because the arsenite added to these algae should not have triggered a marked response in the *arsC* gene which functions to detoxify

arsenate (see [Unabridged description of WGCNA results](#) for details). The *arsC* gene in SAG21 (G2556) is in the Magenta module, which is composed of genes with a variety of functions. The two *arsC* catalogues in 5572 are in the Brown module which is dominated by ribosomal protein genes.

## Phylogenetic analysis of HGTs

ML trees for each arsenic and mercury-related protein are shown in Figure A19 (the *ArsH* tree is also in Figure 4). The hallmark of determining if a gene or protein is the result of HGT is if the gene tree shows phylogenetic incongruence with the species tree for the organism within which it resides. These phylogenies support single acquisitions of genes encoding *ArsB*, *ArsC*, *ArsH*, *ArsM*, aquaglyceroporins and SAM-dependent methyltransferases, and multiple acquisitions of genes encoding *MerA* and *ArsA*. In the *ArsA* and *MerA* phylogenies, there are multiple, distinct monophyletic Cyanidiophyceae clades with non-eukaryote outgroups. In the *ArsA* phylogeny, three Cyanidiales proteins group within the Verrucomicrobia clade, indicating the first transfer, whereas all other *ArsA* proteins are monophyletic (including some other protists) with various Asgard archaea forming the outgroup. The same pattern is seen in the *MerA* phylogeny with all Cyanidiales proteins sharing monophyly with Nitrospirae bacteria as the outgroup and then all Galdieriales proteins sharing monophyly with Candidatus Hydrogenedentes bacteria as the outgroup. All other proteins show monophyly with non-eukaryotes as the outgroup, rather than other algae that would appear sister to Cyanidiophyceae in the species tree.

## ArsH comparison

We built an MSA (Figure 4B) of *ArsH* proteins to compare catalogues in each species. The catalytic regions and folding structure (Sedláček et al., 2022) were used to determine if the *ArsH* copies in each species are functional. This analysis identified two clades of *arsH* genes in *Galdieria* which likely originated from a duplication in the common ancestor of this lineage (Figure 4A). The first clade (Figure 4A, clade 1) is in single copy among isolates, with relatively long branches. The second clade (Figure 4A, clade 2) shows extensive gene duplications, with relatively short branch lengths between the genes within and between isolates, suggesting strong purifying selection. Within clade 2, the SAG21 genes G1846 and G1510 are positioned with the other catalogues from this isolate; however, they have significantly longer branch lengths and have lost conserved residues of the NADP<sup>+</sup> binding site, which may lower the efficacy of this enzyme



(Figure 4B). G2074 from 5572 is also a clade 2 homologue and is also missing (different) residues from the active site, potentially explaining its divergent expression pattern compared to the other paralogues and orthologues (Figure 4B). All genes in this analysis

(except G1846 which shows slight deviation) retain a conserved binding motif for the phosphate group of flavin mononucleotide (FMN).

DISCUSSION

Mercury and arsenic elicit contrasting transcriptomic responses in both algal species

*G. partita* SAG21 and *G. yellowstonensis* 5572 are geographically distinct (Figure A20A). The former was isolated from Yangmingshan National Park, Taiwan, and the latter from Yellowstone National Park, United States. These algae are also evolutionarily distinct, having diverged shortly after the emergence of the *Galdieria* clade approximately 1 billion years ago (Figure A20B; Yoon et al., 2017). Furthermore, each isolate represents the earliest divergence within its respective clade (Rossoni et al., 2019; Van Etten, Stephens, & Bhattacharya, 2023). Despite having indistinguishable morphology and nearly identical genome size and features (e.g., few introns, sparse non-coding DNA, similar GC content; Rossoni et al., 2019), these divergent organisms (recently established as different species; Park et al., 2023) have likely resided on opposite sides of our planet for hundreds of millions of years, possibly since the breakup of the last supercontinent, Pangaea (Correia & Murphy, 2020). All sequenced *Galdieria* possess a single *merA* gene and many *ars* operon genes, indicating that prior to the split of SAG21 and 5572, ancestral Cyanidiophyceae were able to respond to mercury and arsenic stress. The latter was likely peaking due to increased oxidation as an outcome of the more ancient Great Oxygenation Event (ca. 2.4 Bya) which created oxidized forms of arsenic (e.g., arsenate) that resulted in the extinction of organisms lacking detoxification/resistance mechanisms (Fru et al., 2019; Zhu et al., 2014).

Multiple Sequence Alignment showing protein structure

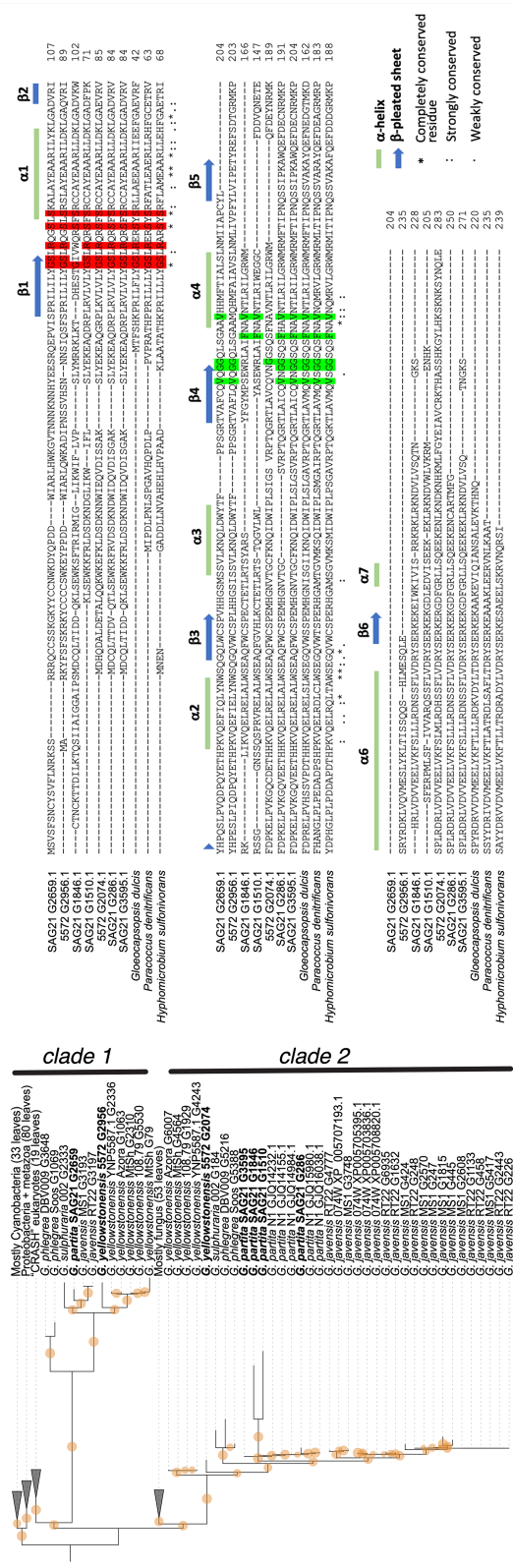


FIGURE 4 The left panel shows a maximum likelihood phylogeny for the ArsH protein. This HGT is the result of a single transfer from bacteria that subsequently diverged into two distinct paralogous groups we refer to as clades 1 and 2. Clade 2 includes gene duplications that have occurred in the SAG21 lineage. The phylogeny was made using iTol v5 (Letunic & Bork, 2021). The right panel shows a multiple sequence alignment of *Galdieria partita* SAG21 and *G. yellowstonensis* 5572 ArsH proteins with various bacterial orthologues. The red-highlighted residues indicate the consensus motif for where binding for the phosphate group on FMN occurs. The green-highlighted residues indicate the consensus sequence for NADP+ binding. The blue arrows indicate  $\beta$ -sheet structure of the folded protein and light green rectangles indicate  $\alpha$ -helix structure of the folded protein. These structural designations are based on those highlighted by Sedláček et al. (2022). HGT, horizontal genetic transfer.

ArsH amino acid ML phylogeny



There are multiple heavy metal detoxification pathways in Cyanidiophyceae. It has been widely hypothesized that the *merA* gene has been transferred multiple times into the Cyanidiophyceae lineage, for example, once each in the Cyanidiales and Galdieriales clades (Figure A19F; Cho et al., 2023; Rossoni et al., 2019; Schönknecht et al., 2013) and that it performs its function without the need for other genes in the prokaryote mercury detoxification pathway (Figure 1). The latter idea is consistent with our results because *merA* follows a pattern of marked differential expression at the onset of mercury(II) exposure that is characteristic of a gene that is responding to an acute stressor (Figure 2) and in the WGCNA analysis, mercury(II) reductase has low MM (<0.8), which demonstrates that it is not tightly coexpressed with other genes (Spreadsheet S4: <https://zenodo.org/doi/10.5281/zenodo.8377091>). These results suggest that mercury detoxification (i.e., *merA*) in *Galdieria* is a direct response to the presence of this toxin, which is likely to be functionally independent of other organismal processes.

Unlike mercury, the arsenic pathway is more complex and does not only involve detoxification, but also, active and passive transport (efflux). *Galdieria* isolates encode different arsenic-related genes in varying copy numbers (Figure 1), with none of the genomes sequenced thus far encoding all *ars* operon genes. The absence of the *arsR* (the operon repressor) gene from all sequenced isolates, the lack of colocalization (i.e., neighbours of each other) of *ars* genes within the genomes of these species, and the phylogenetic analysis (Figure A19) support independent (non-operon) acquisitions of genes encoding ArsB, ArsC, ArsH, ArsM, aquaglyceroporins and methyltransferases, and multiple independent acquisitions of ArsA. This suggests that arsenic detoxification has a complex evolutionary history in this group and does not likely result from a single HGT event involving a prokaryotic donor. Arsenite compounds, such as the sodium arsenite in our experiment, likely enter the cell via aquaporins (e.g., MIP family aquaglyceroporins), bidirectional transmembrane channel proteins that indiscriminately take up a variety of solutes, including metalloids (Mukhopadhyay et al., 2014; Yang & Rosen, 2016). After gaining entry, the arsenite can encounter one of two fates. First, it can remain chemically static and be effluxed from the cell via the ArsA-ArsB complex, wherein ArsA is the pump-driving ATPase that provides energy to ArsB, the transport protein, that extrudes the arsenite (Yang et al., 2012). ArsA and B homologues in both genomes show constitutive expression, although one *arsA* (G2037) and one *arsB* (G983) gene in 5572 share a highly similar pattern of increased expression at later time points, suggesting that the two proteins encoded by these genes are the ArsA-B complex and their expression is linked (Figure 2).

In addition, arsenite can be methylated by ArsM or potentially, another methyltransferase (e.g., SAM-dependent methyltransferase) (Figure 1, right panel), converting it into the more toxic MAs(III) which can then be oxidized by ArsH, an NADPH-dependent FMN oxidoreductase, to less toxic MAs(V) or various other organoarsenicals that can diffuse out of the cell (Nowack et al., 2016; Qin et al., 2009; Yang & Rosen, 2016). These observations suggest that arsenic detoxification, either via efflux or chemical modification in *Galdieria* is responsive to the presence of arsenite compounds; however, each set of genes is also functionally associated with other diverse organismal processes (Figures 3 and A18). Taken together, these results show that all genes encoding arsenite and mercury metabolic pathways in these two algae are of HGT origin; however, the response to mercury is more direct, whereas the arsenite response is complex and difficult to elucidate based on transcriptomic data alone. It is also possible that in the natural environment, arsenic detoxification is a communal process (see below). Analysis of environmental metatranscriptomic and/or metaproteomic data is needed to test this hypothesis.

## The lifecycle of an HGT

Genes encoding processes related to arsenic or mercury tolerance fall under the category of ‘operational’ genes, because they carry out specific functions, such as detoxification or metabolism. In contrast, ‘informational’ genes function as part of tightly linked protein–protein interaction and signalling networks such as transcription and translation (Jain et al., 1999; Schönknecht et al., 2014). The ‘complexity hypothesis’ posits that operational genes have higher transferability because they are modular and do not rely on a network of accessory genes to achieve their function (Jain et al., 1999). Furthermore, the ‘continual horizontal transfer hypothesis’ proposes that, in prokaryotes, the HGT of operational genes is continuous over time and argues against a few ancient instances of many transfers at once (Rivera et al., 1998). Both hypotheses were originally posited for prokaryotes, but aspects hold true for HGT events in *Galdieria* (and likely most eukaryotes, however, more research is needed in this area). Specifically, the functions encoded by these genes are operational and have multiple, independent origins (Figure A19). The recently proposed ‘simplicity hypothesis’ (Jones et al., 2022) is also relevant to our study. This hypothesis invokes connectivity, defined as the extent of protein–protein interactions of the transferable gene product in its native genome, and suggests that the connectivity of a transferable gene (which will vary over time due to environmental constraints) will control its rate of fixation post-HGT. The evolution of high connectivity would render genes less

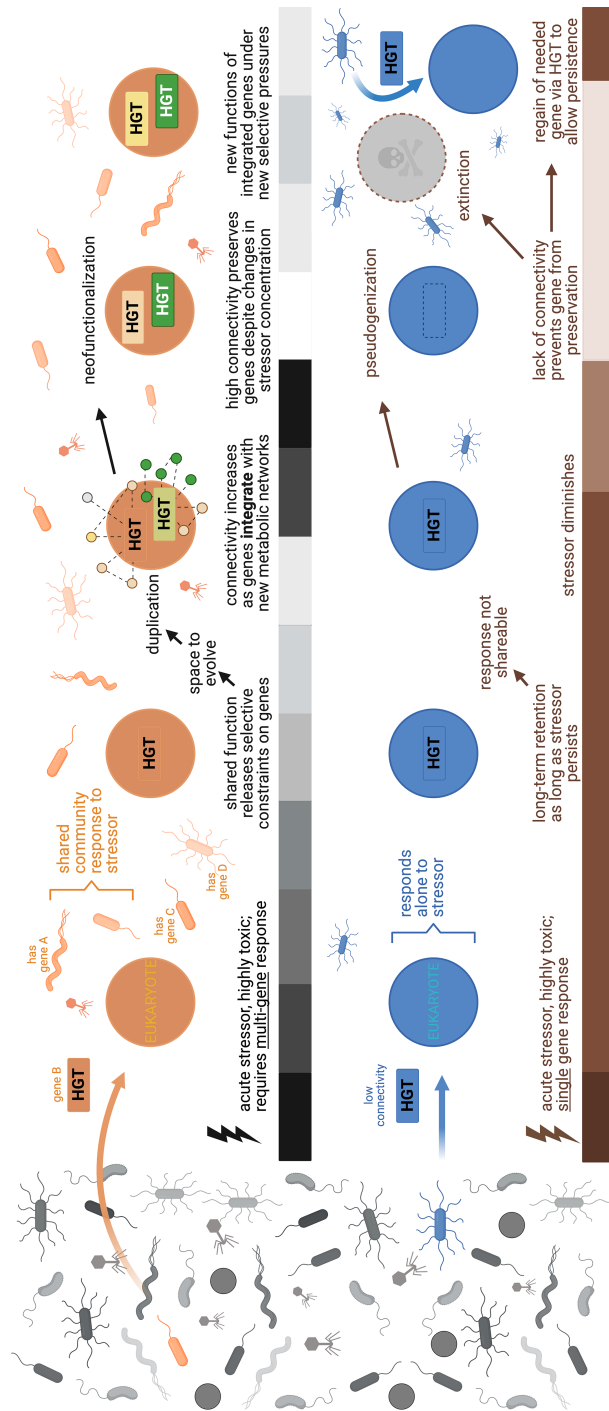
transferable, whereas the opposite would hold for genes with low connectivity (Figure 5).

These hypotheses, however, do not consider the effects that ecology can have on the acquisition and retention of HGTs. Specifically, under the simplicity hypothesis, it would be expected that the *ars* genes, which are all required for detoxification of arsenic, would be harder to transfer than *merA* given that they have developed higher connectivity in the recipient organism, that is, *Galdieria*. Here, we refer to MM from

the WGCNA analysis as a rough proxy for connectivity. However, in contrast to *merA*, *ars* genes have a convoluted evolutionary history in Cyanidiophyceae. The acquisition and retention of these genes are likely dependent not only on their connectivity with other genes in the recipient organism but also on their interaction with genes encoded by other taxa in the environment. That is, the acquisition of, for example, *ars* genes, may depend on how efficiently that function can be done by co-habiting taxa. If part of a pathway can be performed by other organisms, then selection would not drive the acquisition of this function in a naïve organism. As the environment and community (and by extension functional) composition of an environment changes, then selection would drive the acquisition of different genes, and potentially, loss of existing genes that have functions that can be performed by other organisms in the community.

## The integrated HGT model

The IHM extends the simplicity and complexity hypotheses to eukaryotes, which have traits distinct from most prokaryotes, such as larger genomes where HGTs may survive intact and undergo low background expression and fixation if they are selectively advantageous, potentially followed by gene duplication and divergence



**FIGURE 5** The Integrated HGT Model (IHM). We show the fates of two different types of HGTs. The top panel shows HGT from prokaryote to eukaryote following the introduction of an acute stressor (its concentration over time is represented by the black gradient). To combat this stressor requires a multigene response. This can be accomplished by one organism possessing an operon or by splitting up these genes among different organisms that provide some redundancy and preserve gene product functions in the pathway. Over time, the ongoing shared community response to the stressor reduces selective constraints on the stress-responsive HGTs, allowing them to gain new functions, possibly via duplication and divergence. The reduced selective pressures also allow connectivity to increase, and the HGT-derived genes integrate into host metabolic networks that may or may not have functions related to the original purpose. The stressor may diminish or fluctuate over time but having the HGTs connected to large networks of proteins hinders loss. Over time, diversification will increase, and new selective pressures related to new functions will ensue. This scenario is represented by our arsenic data. The bottom panel shows HGT from prokaryote to eukaryote following the introduction of an acute stressor (brown gradient) that requires a single gene response. This gene will be retained while the stressor persists; however, the response is not divisible, and therefore, this gene does not integrate into host metabolic networks. If the stressor diminishes over time, the lack of connectivity of the HGT gene products to other proteins will prevent its preservation and the accumulation of mutations may lead to pseudogenization. If the stressor returns but the HGT is no longer functional, the organism will perish. Alternatively, the eukaryote may once again acquire the gene from a prokaryote in a cycle of gain/loss that ultimately preserves the function in the ecosystem. Image made in [Biorender.com](https://www.biorender.com). HGT, horizontal genetic transfer.



(De Clerck et al., 2018; Schönknecht et al., 2014; Van Etten & Bhattacharya, 2020). We postulate that as in prokaryotes, operational genes with low connectivity in the donor tend to be transferred more often, with changes in connectivity post-transfer impacting their fates. However, these prokaryote-driven frameworks do not account for the processes of retention and integration of HGTs that we propose with the IHM. The IHM is based on our findings in two detoxification pathways which lead us to consider how community-level response to environmental stressors may lessen selective constraints on HGTs, allowing them to explore functional space and thus integrate into new metabolic networks (Figure 5). However, to better understand the linkage of *ars* HGTs to algal gene expression networks, future studies should explore the responses of these networks to other environmental changes such as light, temperature, pH or oxidative stress. These results may enlarge upon the functions encoded by these genes when compared to their origin function in donor prokaryotes. In the case of increased connectivity following an HGT event, as is evident for the *ars* genes in *Gal-dieria*, the IHM suggests that due to duplications, HGT-derived gene copies may not only become connected to other proteins in recipient metabolic networks but may take on new functions. This is the result of lowered selective constraints on these genes, possibly due to their presence and expression in other microbes in the environment. These new functional connections reduce the likelihood that this HGT will be lost and thus, integration leads to preservation, duplication and putatively, novel functions (Figure 5, top panel). Conversely, in the case of *merA*, a single, presumably non-divisible gene is responsible for facilitating response to the chemical stressor and thus, the HGT remains less connected and is easier to gain and lose, if the stressor disappears. The latter can lead to extinction if the stressor reappears or alternatively, a subsequent *merA* HGT rescues the lineage. The IHM predicts that genes with low connectivity may show multiple rounds of gain and loss based on the presence/absence of the stressor (Figure 5, bottom panel). This scenario is supported by the multiple *merA* gene acquisitions in Cyanidiophyceae, and the complex evolutionary history of *ars* genes which is likely the result of variable selective pressure from changes in community functional composition (Figure A19).

## The curious case of ArsH

The *arsH* gene family in Cyanidiophyceae was likely the result of a single HGT event that underwent a duplication early in the evolution of the lineage (Figure 4). ArsH is found primarily in bacteria, where it forms part of some, but not all, prokaryotic *ars* operons (Páez-Espino et al., 2009). In some organisms, ArsH is

hypothesized to be an integral part of the arsenite detoxification toolkit; however, its role in this pathway is indirect and sometimes enigmatic. ArsH can reduce  $O_2$  to  $H_2O_2$ , which can then oxidize trivalent forms of arsenite compounds to pentavalent forms such as  $MAs(V)$ , which are less toxic and gaseous and can thus diffuse out of the cell (Chen et al., 2015; Ye et al., 2007). Alternatively, as shown in *Pseudomonas putida*, ArsH may play a more accessory role by quenching oxidative stress caused by exposure to arsenic-containing salts or other types of redox stressors via the reduction of FMN by NADPH (Páez-Espino et al., 2020). All *Gal-dieria* species encode two paralogues of this gene (termed clades 1 and 2 in Figure 4). Clade 1 has remained a single copy in all genomes, whereas clade 2 has duplicated in some genomes, such as SAG21 (four copies) but not 5572. In SAG21, the four clade 2 homologues share an identical expression pattern with a marked transcriptional peak at TP3, coincident with SAM-dependent methyltransferases and the putative *arsM* gene. All clade 2 paralogues from SAG21 are grouped into the same WGCNA module (Turquoise, Figure 3) and are strongly coexpressed with translation-associated and ribosomal protein-encoding genes. This may indicate increased protein turnover caused by arsenic stress or alternatively, that ArsH is taking on a second function that is linked with well-conserved informational genes that thus, aids in the preservation of the HGT. This is an example of increased connectivity driving altered function (IHM). Furthermore, the two SAG21 clade 2 paralogues (G1846 and G1510) have long branch lengths compared to other members of the clade (Figure 4A) and have a weaker transcriptomic response (i.e., are expressed in much lower numbers) upon arsenite exposure. This may be evidence of weakened selective constraints on these genes that are likely undergoing neofunctionalization. These two genes have the highest connectivity to native genes in the network (Figure 3). The MSA (Figure 4B) supports this idea, showing that both genes have lost conserved residues of the  $NADP^+$  binding site and G1846 has an altered phosphate binding motif for FMN.

The SAG21 clade 1 *arsH* paralogue (G2659) has an expression pattern that has deviated from the other four copies (Figure 2), suggesting it has undergone neofunctionalization. This idea is supported by the WGCNA results, which for SAG21 place this gene in the Royal-Blue module (rather than Turquoise) that is composed of a majority of photosynthesis-related genes, linking it to cellular processes distinct from detoxification but still relevant to the redox stress response (Figure 3). Moreover, three of the RoyalBlue photosynthesis genes are transketolases in the pentose phosphate pathway, which is responsible for producing NADPH (Solovjeva et al., 2020). ArsH is a NADPH-dependent oxidoreductase, therefore this finding may represent a novel,



beneficial function that is independent of arsenite detoxification and consistent with the IHM. In 5572, *arsH* was placed in the smaller DarkGreen module that is largely comprised of dark genes of unknown function, indicating that this copy may also have taken on a novel function that cannot yet be discerned (Figure 3).

## Open questions

Our study leaves several key questions open that will require the use of newly developed genetic tools for *Galdieria* (Hirooka et al., 2022). Importantly, the gene expression data do not allow us to determine if *Galdieria* species are using extrusion and/or detoxification pathways for arsenite detoxification because the relevant genes are all expressed. Organisms grown in monoculture provide a narrower perspective on physiology that may not reflect natural conditions (Figure 1C, ‘*Galdieria in monoculture*’). Wild Cyanidiophyceae coexist with many other microbes, often in biofilms, in extreme habitats. It is possible that in nature, *Galdieria* spp. primarily tolerate arsenite through their functional and conserved efflux system (ArsAB), whereas a different microbe (one with clear ArsM activity, e.g., *Cyanidioschyzon* sp., which lacks most of the other *ars* genes; Yang & Rosen, 2016) methylates environmental arsenite to MAs(III), which reenters *Galdieria* cells where ArsH can aid in detoxification (Figure 1C, ‘*Galdieria in nature*’). As in endosymbiosis, where genetic drift results in gene loss, genetic transfer and genome reduction of the endosymbiont, we see extreme genome reduction in all Cyanidiophyceae (Miyagishima & Tanaka, 2021; Qiu et al., 2013). This pattern invokes the Black Queen Hypothesis, whereby gene loss ‘in free-living organisms may leave them dependent on cooccurring microbes for lost metabolic functions’ and ‘can provide a selective advantage by conserving an organism’s limiting resources, provided the gene’s function is dispensable’ (Lee et al., 2022; Morris et al., 2012). Implicating the Black Queen Hypothesis lends support to the eukaryote-based IHM because the community-level response leads to the sharing of common goods and conserves gene product functions on an ecosystem scale, which may foster longer-term stability. Therefore, within eukaryotes, it is insufficient to ascribe adaptive significance to HGTs based only on their role in prokaryotes. Investigation of their integration into host metabolic networks may more fully capture the role of HGTs in driving biodiversity and adaptation within the eukaryotic tree of life.

## AUTHOR CONTRIBUTIONS

**Debashish Bhattacharya:** Conceptualization; writing – original draft; supervision; funding acquisition; resources; writing – review and editing. **Julia Van Etten:** Conceptualization; data curation; formal

analysis; visualization; writing – original draft; methodology; writing – review and editing; software. **Timothy G. Stephens:** Conceptualization; methodology; supervision; validation; writing – review and editing. **Erin Chille:** Investigation; software; formal analysis. **Anna Lipzen:** Data curation; methodology. **Daniel Peterson:** Data curation; methodology. **Kerrie Barry:** Resources; funding acquisition; writing – review and editing. **Igor V. Grigoriev:** Funding acquisition; resources; writing – review and editing.

## ACKNOWLEDGEMENTS

JVE was supported by the National Aeronautics and Space Administration Future Investigators in NASA Earth and Space Science and Technology (FINESST grant 80NSSC19K1542). TGS and DB were supported by a grant from NASA (80NSSC19K0462) awarded to DB. DB was also supported by a NIFA-USDA Hatch grant (NJ01180). We acknowledge Prof Peter Lammers for providing us with meticulously cultivated axenic cultures of *G. yellowstonensis* 5572 and *G. partita* SAG21. We would also like to acknowledge our collaboration with the JGI. The work (proposal: 10.46936/10.25585/60000481) conducted by the U.S. Department of Energy Joint Genome Institute (<https://ror.org/04xm1d337>), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We thank two anonymous reviewers for their constructive comments.

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

RNA-sequencing data are publicly available at the National Library of Medicine under the BioProject numbers PRJNA999762 to PRJNA999857. Supplementary files, including spreadsheets, code and Newick files, can be found in Zenodo at <https://zenodo.org/doi/10.5281/zenodo.8377091>. The custom script is available on GitHub: [https://github.com/TimothyStephens/Utils/tree/main/Blast/taxonomically\\_downsample\\_hits](https://github.com/TimothyStephens/Utils/tree/main/Blast/taxonomically_downsample_hits).

## ORCID

**Debashish Bhattacharya**  <https://orcid.org/0000-0003-0611-1273>

## REFERENCES

- Allen, M.B. (1959) Studies with *Cyanidium caldarium*, an anomalously pigmented chlorophyte. *Archiv für Mikrobiologie*, 32, 270–277.
- Alsmark, C., Foster, P.G., Sicheritz-Ponten, T., Nakjang, S., Martin Embley, T. & Hirt, R.P. (2013) Patterns of prokaryotic lateral gene transfers affecting parasitic microbial eukaryotes. *Genome Biology*, 14, 1–6.
- Ben Fekih, I., Zhang, C., Li, Y.P., Zhao, Y., Alwathnani, H.A., Saquib, Q. et al. (2018) Distribution of arsenic resistance genes in prokaryotes. *Frontiers in Microbiology*, 9, 2473.



- Boyd, E.S. & Barkay, T. (2012) The mercury resistance operon: from an origin in a geothermal environment to an efficient detoxification machine. *Frontiers in Microbiology*, 3, 349.
- Buchfink, B., Reuter, K. & Drost, H.G. (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, 18(4), 366–368.
- Burch, C.L., Romanchuk, A., Kelly, M., Wu, Y. & Jones, C.D. (2023) Empirical evidence that complexity limits horizontal gene transfer. *Genome Biology and Evolution*, 15(6), evad089.
- Bushnell, B. (2014) *BBMap: a fast, accurate, splice-aware aligner*. Berkeley, CA: Lawrence Berkeley National Lab. (LBNL).
- Castenholz, R.W. & McDermott, T.R. (2010) *The Cyanidiales: ecology, biodiversity, and biogeography*. the Netherlands: Springer.
- Chen, J., Bhattacharjee, H. & Rosen, B.P. (2015) ArsH is an organoarsenical oxidase that confers resistance to trivalent forms of the herbicide monosodium methylarsenate and the poultry growth promoter roxarsone. *Molecular Microbiology*, 96(5), 1042–1052.
- Chen, S.C., Sun, G.X., Rosen, B.P., Zhang, S.Y., Deng, Y., Zhu, B.K. et al. (2017) Recurrent horizontal transfer of arsenite methyltransferase genes facilitated adaptation of life to arsenic. *Scientific Reports*, 7(1), 7741.
- Chille, E., Strand, E., Neder, M., Schmidt, V., Sherman, M., Mass, T. et al. (2021) Developmental series of gene expression clarifies maternal mRNA provisioning and maternal-to-zygotic transition in a reef-building coral. *BMC Genomics*, 22(1), 1–7.
- Cho, C.H., Park, S.I., Huang, T.Y., Lee, Y., Ciniglia, C., Yadavalli, H.C. et al. (2023) Genome-wide signatures of adaptation to extreme environments in red algae. *Nature Communications*, 14(1), 10.
- Christakis, C.A., Barkay, T. & Boyd, E.S. (2021) Expanded diversity and phylogeny of mer genes broadens mercury resistance paradigms and reveals an origin for MerA among thermophilic archaea. *Frontiers in Microbiology*, 12, 682605.
- Correia, P. & Murphy, J.B. (2020) Iberian-Appalachian connection is the missing link between Gondwana and Laurasia that confirms a Wegenerian Pangaea configuration. *Scientific Reports*, 10(1), 2498.
- De Clerck, O., Kao, S.M., Bogaert, K.A., Blomme, J., Foflonker, F., Kwantes, M. et al. (2018) Insights into the evolution of multicellularity from the sea lettuce genome. *Current Biology*, 28(18), 2921–2933.
- Doemel, W.N. & Brock, T.D. (1971) The physiological ecology of *Cyanidium caldarium*. *Microbiology*, 67(1), 17–32.
- Francisco, M.J., Hope, C.L., Owolabi, J.B., Tisa, L.S. & Rosen, B.P. (1990) Identification of the metalloregulatory element of the plasmid-encoded arsenical resistance operon. *Nucleic Acids Research*, 18(3), 619–624.
- Fru, E.C., Somogyi, A., El Albani, A., Medjoubi, K., Aubineau, J., Robbins, L.J. et al. (2019) The rise of oxygen-driven arsenic cycling at ca. 2.48 Ga. *Geology*, 47(3), 243–246.
- Gentleman RC, Carey V, Huber W, Hahne F. Genefilter: methods for filtering genes from high-throughput experiments. R package version. 2015;1(1).
- Hirooka, S., Itabashi, T., Ichinose, T.M., Onuma, R., Fujiwara, T., Yamashita, S. et al. (2022) Life cycle and functional genomics of the unicellular red alga *Galdieria* for elucidating algal and plant evolution and industrial use. *Proceedings of the National Academy of Sciences of the United States of America*, 119(41), e2210665119.
- Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. & Vinh, L.S. (2018) UFBoot2: improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution*, 35(2), 518–522.
- Huang, J. (2013) Horizontal gene transfer in eukaryotes: the weak-link model. *BioEssays*, 35(10), 868–875.
- Jain, R., Rivera, M.C. & Lake, J.A. (1999) Horizontal gene transfer among genomes: the complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 96(7), 3801–3806.
- Jones, C.T., Susko, E. & Bielawski, J.P. (2022) Evolution of the connectivity and indispensability of a transferable gene: the simplicity hypothesis. *BMC Ecology and Evolution*, 22(1), 1–3.
- Kalia, K. & Joshi, D.N. (2009) Detoxification of arsenic. In: *Handbook of toxicology of chemical warfare agents*. Cambridge, MA: Academic Press, pp. 1083–1100.
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780.
- Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. (2019) Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*, 37(8), 907–915.
- Koonin, E.V. (2016) Horizontal gene transfer: essentiality and evolvability in prokaryotes, and roles in evolutionary transitions. *F1000Research*, 5, F1000 Faculty Rev-1805.
- Kovaka, S., Zimin, A.V., Pertea, G.M., Razaghi, R., Salzberg, S.L. & Pertea, M. (2019) Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, 20(1), 1–3.
- Langfelder, P. & Horvath, S. (2008) WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9(1), 1–3.
- Langfelder, P. & Horvath, S. (2012) Fast R functions for robust correlations and hierarchical clustering. *Journal of Statistical Software*, 46(11), i11.
- Langfelder P, Horvath S. (2016) Tutorials for the WGCNA package. UCLA. Available from: <https://web.archive.org/web/20230813135706/https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/Rpackages/WGCNA/Tutorials/>.
- Lee, I.P., Eldakar, O.T., Gogarten, J.P. & Andam, C.P. (2022) Bacterial cooperation through horizontal gene transfer. *Trends in Ecology & Evolution*, 37(3), 223–232.
- Lehr, C.R., Frank, S.D., Norris, T.B., D'Imperio, S., Kalinin, A.V., Toplin, J.A. et al. (2007) Cyanidia (Cyanidiales) population diversity and dynamics in an acid-sulfate-chloride spring in Yellowstone National Park 1. *Journal of Phycology*, 43(1), 3–14.
- Letunic, I. & Bork, P. (2021) Interactive tree of life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, 49(W1), W293–W296.
- Love, M.I., Huber, W. & Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 1–21.
- Marcet-Houben, M. & Gabaldón, T. (2010) Acquisition of prokaryotic genes by fungal genomes. *Trends in Genetics*, 26(1), 5–8.
- Miyagishima, S.Y. & Tanaka, K. (2021) The unicellular red alga *Cyanidochytrion merolae*—the simplest model of a photosynthetic eukaryote. *Plant and Cell Physiology*, 62(6), 926–941.
- Morris, J.J., Lenski, R.E. & Zinser, E.R. (2012) The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *MBio*, 3(2), 10–128.
- Mukhopadhyay, R., Bhattacharjee, H. & Rosen, B.P. (2014) Aquaglyceroporins: generalized metalloid channels. *Biochimica et Biophysica Acta (BBA)—General Subjects*, 1840(5), 1583–1591.
- Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. & Minh, B.Q. (2015, 1) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32, 268–274.
- Nowack, E.C., Price, D.C., Bhattacharya, D., Singer, A., Melkonian, M. & Grossman, A.R. (2016) Gene transfers from diverse bacteria compensate for reductive genome evolution in the chromatophore of *Paulinella chromatophora*. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43), 12214–12219.
- Páez-Espino, A.D., Nikel, P.I., Chavarría, M. & de Lorenzo, V. (2020) ArsH protects *Pseudomonas putida* from oxidative damage caused by exposure to arsenic. *Environmental Microbiology*, 22(6), 2230–2242.



- Páez-Espino, D., Tamames, J., de Lorenzo, V. & Cánovas, D. (2009) Microbial responses to environmental arsenic. *Biometals*, 22, 117–130.
- Palmgren, M., Engström, K., Hallström, B.M., Wahlberg, K., Søndergaard, D.A., Säll, T. et al. (2017) AS3MT-mediated tolerance to arsenic evolved by multiple independent horizontal gene transfers from bacteria to eukaryotes. *PLoS One*, 12(4), e0175422.
- Park, S.I., Cho, C.H., Ciniglia, C., Huang, T.Y., Liu, S.L., Bustamante, D.E. et al. (2023) Revised classification of the Cyanidophyceae based on plastid genome data with descriptions of the Cavernicolales ord. nov. and Galdieriales ord. nov. (Rhodophyta). *Journal of Phycology*, 59, 444–466.
- Pereira L, Christin PA, Dunning LT. The mechanisms underpinning lateral gene transfer between grasses. *Plants, People, Planet*. 2022, 5(5), 672–682.
- Pertea, G. & Pertea, M. (2020) GFF utilities: GffRead and GffCompare. *F1000Research*, 9, ISCB Comm J-304.
- Qin, J., Lehr, C.R., Yuan, C., Le, X.C., McDermott, T.R. & Rosen, B.P. (2009) Biotransformation of arsenic by a Yellowstone thermoacidophilic eukaryotic alga. *Proceedings of the National Academy of Sciences of the United States of America*, 106(13), 5213–5217.
- Qiu, H., Price, D.C., Weber, A.P., Reeb, V., Yang, E.C., Lee, J.M. et al. (2013) Adaptation through horizontal gene transfer in the cryptoendolithic red alga *Galdieria phlegrea*. *Current Biology*, 23(19), R865–R866.
- Reeb, V., Bhattacharya, D., Seckbach, J. & Chapman, D. (n.d.) Red algae in the genomic age. Cellular origin, life in extreme habitats and astrobiology.
- Ribeiro, G.M. & Lahr, D.J.G. (2022) A comparative study indicates vertical inheritance and horizontal gene transfer of arsenic resistance-related genes in eukaryotes. *Molecular Phylogenetics and Evolution*, 173, 107479.
- Rivera, M.C., Jain, R., Moore, J.E. & Lake, J.A. (1998) Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences of the United States of America*, 95(11), 6239–6244.
- Rosen, B.P., Hsu, C.M., Karkaria, C.E., Owolabi, J.B. & Tisa, L.S. (1990) Molecular analysis of an ATP-dependent anion pump. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 326(1236), 455–463.
- Rossoni, A.W., Price, D.C., Seger, M., Lyska, D., Lammers, P., Bhattacharya, D. et al. (2019) The genomes of polyextremophilic cyanidiales contain 1% horizontally transferred genes with diverse adaptive functions. *eLife*, 8, e45017.
- Schönknecht, G., Chen, W.H., Ternes, C.M., Barbier, G.G., Shrestha, R.P., Stanke, M. et al. (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, 339(6124), 1207–1210.
- Schönknecht, G., Weber, A.P. & Lercher, M.J. (2014) Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *BioEssays*, 36(1), 9–20.
- Seckbach, J. (1972) On the fine structure of the acidophilic hot-spring alga *Cyanidium caldarium*: a taxonomic approach. *Microbios*, 5(18), 133–142.
- Sedláček, V., Kryl, M. & Kučera, I. (2022) The ArsH protein product of the *Paracoccus denitrificans* ars operon has an activity of organoarsenic reductase and is regulated by a redox-responsive repressor. *Antioxidants*, 11(5), 902.
- Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D. et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11), 2498–2504.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W. et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), 539.
- Solovjeva, O.N., Kovina, M.V., Zavialova, M.G., Zgoda, V.G., Shcherbinin, D.S. & Kochetov, G.A. (2020) The mechanism of a one-substrate transketolase reaction. *Bioscience Reports*, 40(8), BSR20180246.
- Soucy, S.M., Huang, J. & Gogarten, J.P. (2015) Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, 16(8), 472–482.
- Van Etten, J. & Bhattacharya, D. (2020) Horizontal gene transfer in eukaryotes: not if, but how much? *Trends in Genetics*, 36(12), 915–925.
- Van Etten, J., Cho, C.H., Yoon, H.S. & Bhattacharya, D. (2023) Extremophilic red algae as models for understanding adaptation to hostile environments and the evolution of eukaryotic life on the early earth. *Seminars in cell & developmental biology*, 134, 4–13.
- Van Etten, J., Stephens, T.G. & Bhattacharya, D. (2023) A k-mer-based approach for phylogenetic classification of taxa in environmental genomic data. *Systematic Biology*, 72(5), 1101–1118.
- Yan, X., Jeub, L.G., Flammini, A., Radicchi, F. & Fortunato, S. (2018) Weight thresholding on complex networks. *Physical Review E*, 98(4), 042304.
- Yang, H.C., Fu, H.L., Lin, Y.F. & Rosen, B.P. (2012) Pathways of arsenic uptake and efflux. *Current topics in membranes*, 69, 325–358.
- Yang, H.C. & Rosen, B.P. (2016) New mechanisms of bacterial arsenic resistance. *Biomedical Journal*, 39(1), 5–13.
- Ye, J., Yang, H.C., Rosen, B.P. & Bhattacharjee, H. (2007) Crystal structure of the flavoprotein ArsH from *Sinorhizobium meliloti*. *FEBS Letters*, 581(21), 3996–4000.
- Yoon, H.S., Nelson, W., Lindstrom, S.C., Boo, S.M., Poeschel, C., Qiu, H. et al. (2017) Rhodophyta. In: *Handbook of the protists*, 2nd edition. Switzerland: Springer International Publishing, pp. 89–133.
- Zhu, Y.G., Yoshinaga, M., Zhao, F.J. & Rosen, B.P. (2014) Earth abides arsenic biotransformations. *Annual Review of Earth and Planetary Sciences*, 42, 443–467.

**How to cite this article:** Van Etten, J., Stephens, T.G., Chille, E., Lipzen, A., Peterson, D., Barry, K. et al. (2024) Diverse fates of ancient horizontal gene transfers in extremophilic red algae. *Environmental Microbiology*, 26(5), e16629. Available from: <https://doi.org/10.1111/1462-2920.16629>



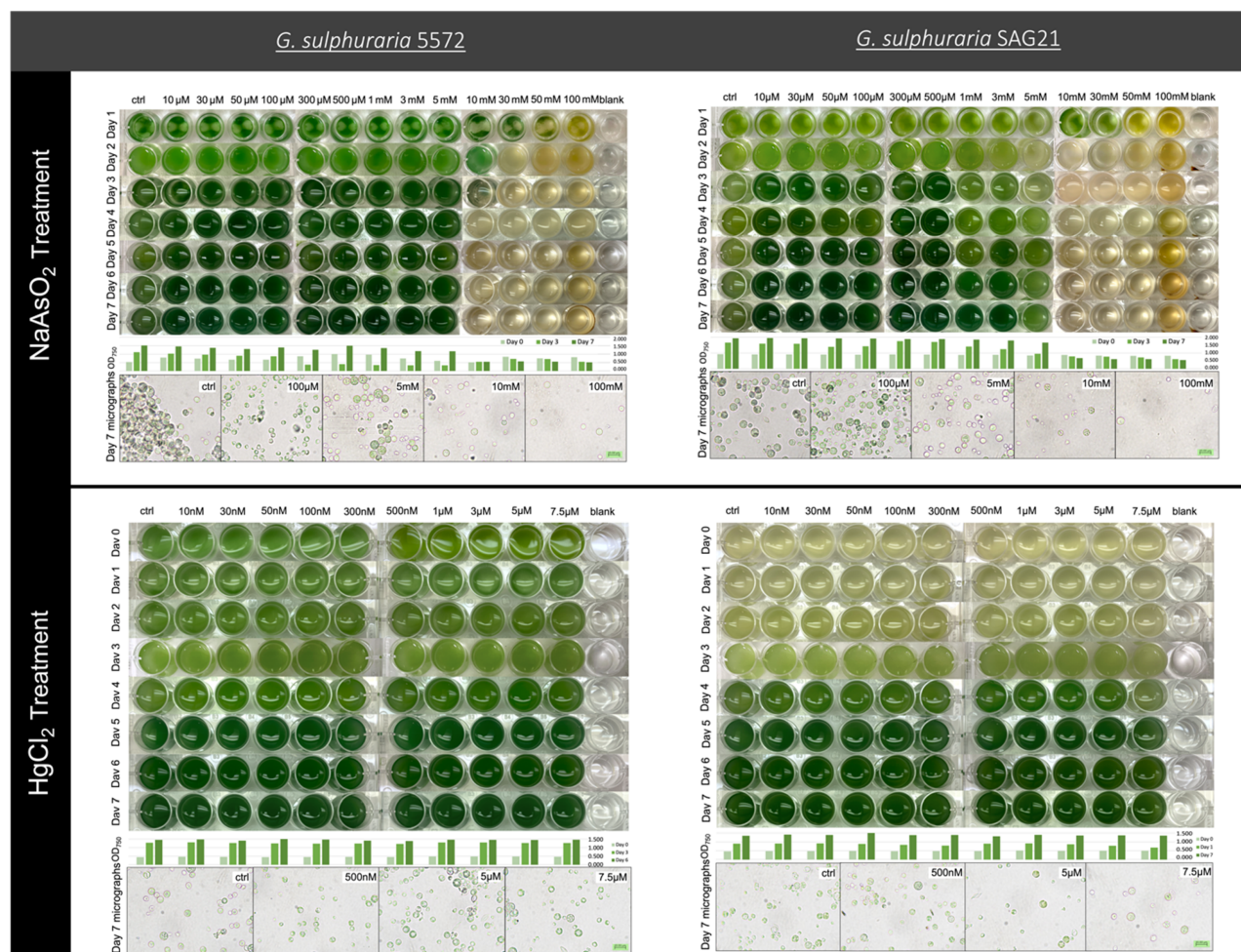
## APPENDIX

## Preliminary growth experiments methods and results

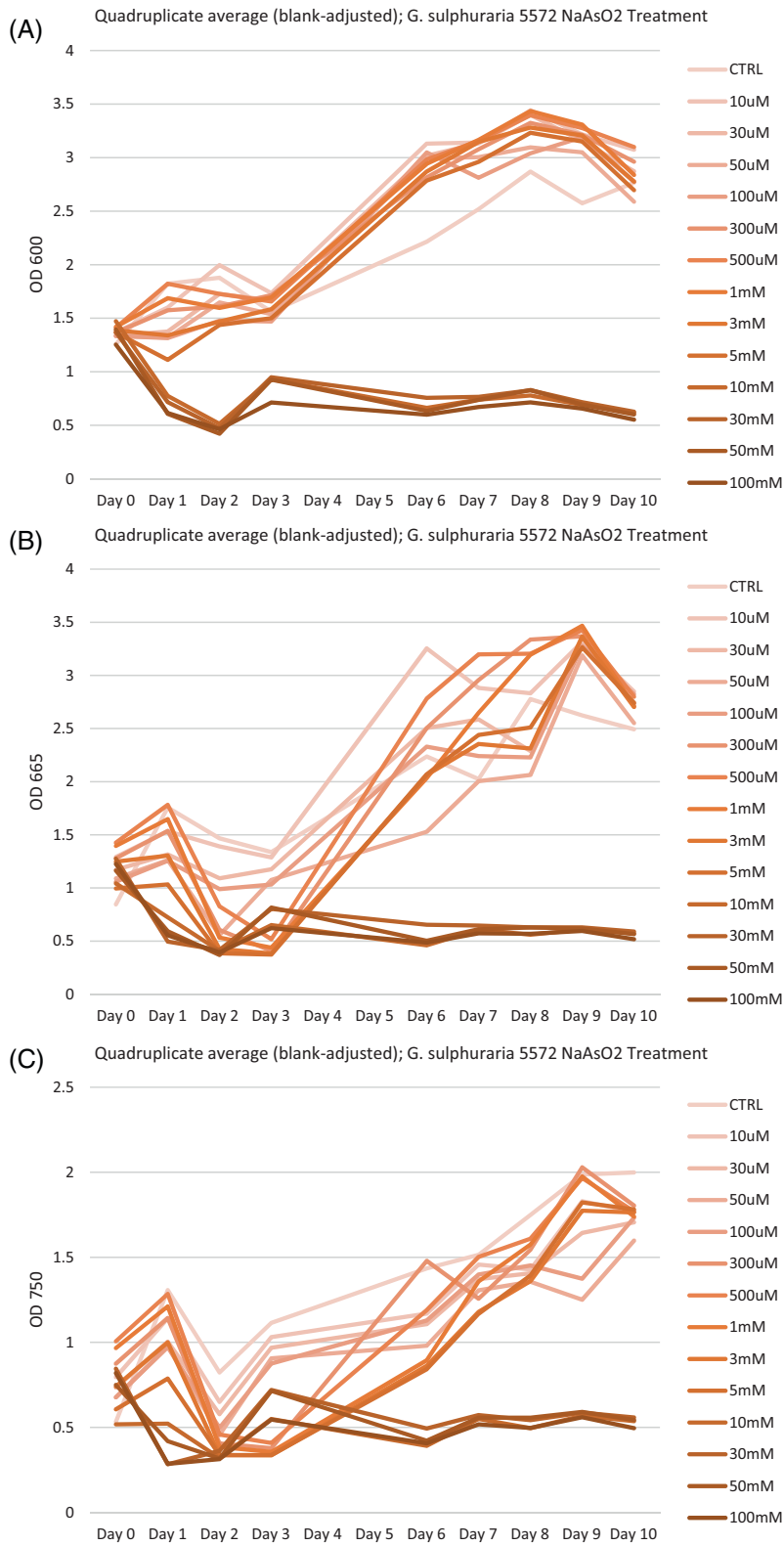
## Methods

To determine concentrations of sodium arsenite ( $\text{NaAsO}_2$ ) and mercuric chloride ( $\text{HgCl}_2$ ) that would likely elicit a transcriptional response from which the organism would recover a well plate experiment was devised. About 2 mL of each algal strain (*G. sulphuraria* 5572 and SAG21; hereinafter 5572 and SAG21, respectively) in  $2\times$  modified Allen medium with 25 mM glucose (pH 2) was pre-diluted and then added to each well to ensure constant concentration across the plate (Allen, 1959).  $\text{NaAsO}_2$  and  $\text{HgCl}_2$  were then added differentially to each plate column (four wells per column; quadruplicate) to achieve the desired concentrations which were as follows: 10  $\mu\text{M}$ , 30  $\mu\text{M}$ , 50  $\mu\text{M}$ , 100  $\mu\text{M}$ , 300  $\mu\text{M}$ , 500  $\mu\text{M}$ , 1 mM, 3 mM, 5 mM, 10 mM, 30 mM, 50 mM, 100 mM for  $\text{NaAsO}_2$ ; and 10 nM, 30 nM, 50 nM, 100 nM, 300 nM, 500 nM, 1  $\mu\text{M}$ , 3  $\mu\text{M}$ , 5  $\mu\text{M}$  and 7.5  $\mu\text{M}$  for  $\text{HgCl}_2$ . Each set of experiments also included a control in which algae were grown in media without treatment and a blank that was media with no algae or treatment. Each day, photos were taken with the plates under controlled lighting, and optical density measurements were taken using a BioTek Synergy 4 microplate reader. Specifically, measurements of  $\text{OD}_{600}$ ,  $\text{OD}_{680}$  and  $\text{OD}_{750}$  were recorded simultaneously for each well. Furthermore, subsamples from each condition were observed under the microscope every 2 days, and micrographs were taken on Day 7. The experiments were all run at the same time and lasted 7 days. Visual summary data can be seen in Figure A1 below which shows photos of a single replicate per day, select  $\text{OD}_{750}$  values (from Days 0, 3 and 7), and select micrographs. Full charts of all ODs can be found in Spreadsheet S1 (<https://zenodo.org/doi/10.5281/zenodo.8377091>) and the growth curves from

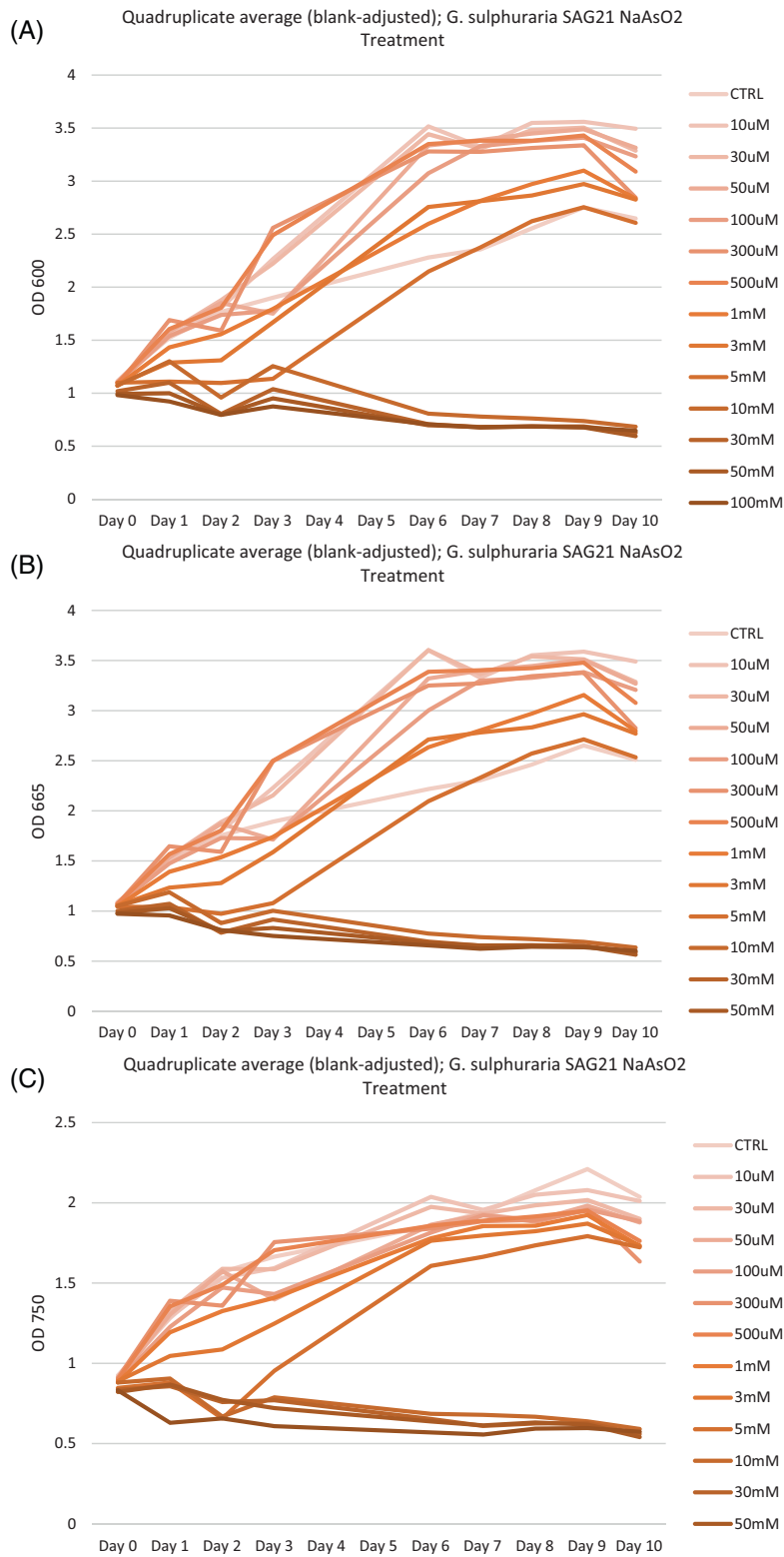
30 mM, 50 mM and 100 mM for  $\text{NaAsO}_2$ ; and 10 nM, 30 nM, 50 nM, 100 nM, 300 nM, 500 nM, 1  $\mu\text{M}$ , 3  $\mu\text{M}$ , 5  $\mu\text{M}$  and 7.5  $\mu\text{M}$  for  $\text{HgCl}_2$ . Each set of experiments also included a control in which algae were grown in media without treatment and a blank that was media with no algae or treatment. Each day, photos were taken with the plates under controlled lighting, and optical density measurements were taken using a BioTek Synergy 4 microplate reader. Specifically, measurements of  $\text{OD}_{600}$ ,  $\text{OD}_{680}$  and  $\text{OD}_{750}$  were recorded simultaneously for each well. Furthermore, subsamples from each condition were observed under the microscope every 2 days, and micrographs were taken on Day 7. The experiments were all run at the same time and lasted 7 days. Visual summary data can be seen in Figure A1 below which shows photos of a single replicate per day, select  $\text{OD}_{750}$  values (from Days 0, 3 and 7), and select micrographs. Full charts of all ODs can be found in Spreadsheet S1 (<https://zenodo.org/doi/10.5281/zenodo.8377091>) and the growth curves from



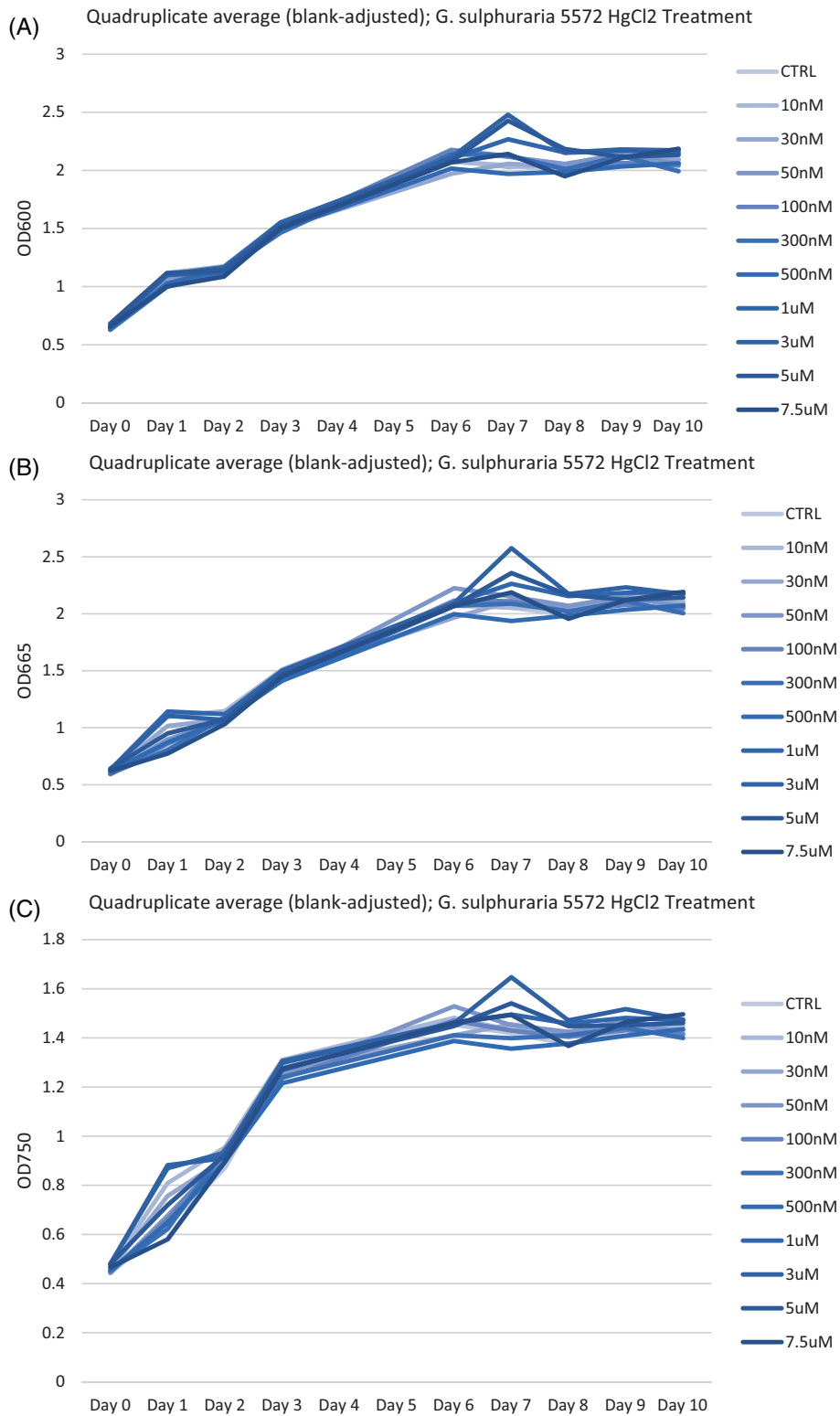
**FIGURE A1** Visual summary data for preliminary growth experiments in *Galdieria sulphuraria* 5572 (left) and *G. sulphuraria* SAG21 (right) treated with  $\text{NaAsO}_2$  (top) and  $\text{HgCl}_2$  (bottom). The same well from each condition was photographed each day and used to make this composite picture. Below the well photos are select optical density (OD) data, specifically  $\text{OD}_{750}$  measurements shown for Days 0, 3 and 7. Below the OD graphs are select micrographs taken at the end of the experiment.



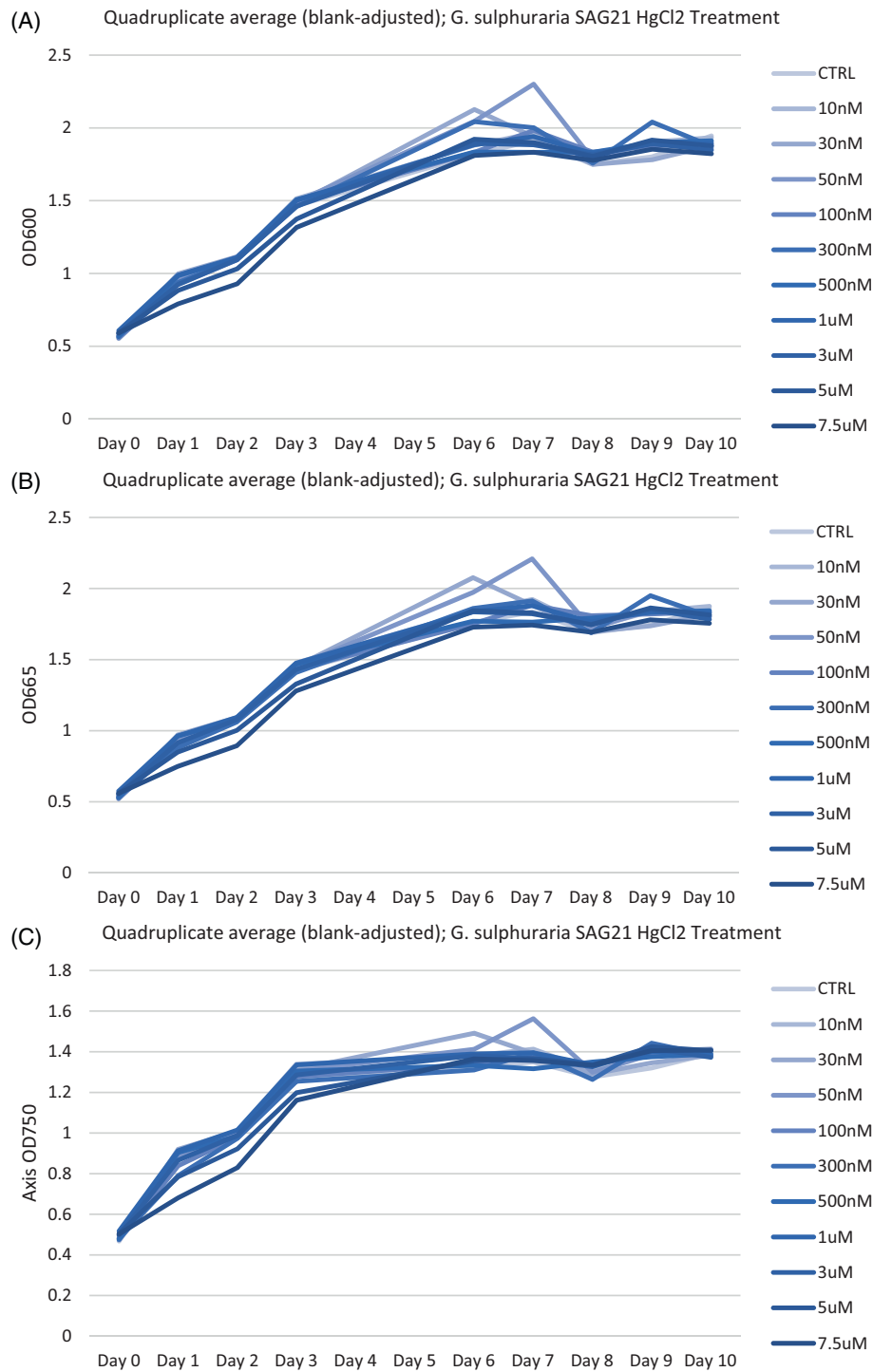
**FIGURE A2** *Galdieria sulphuraria* 5572 sodium arsenite treatment quadruplicate average OD for (A) OD<sub>600</sub>, (B) OD<sub>665</sub> and (C) OD<sub>750</sub>. OD, optical density.



**FIGURE A3** *Galdieria sulphuraria* SAG21 sodium arsenite treatment quadruplicate average OD for (A) OD<sub>600</sub>, (B) OD<sub>665</sub> and (C) OD<sub>750</sub>. OD, optical density.



**FIGURE A4** *Galdieria sulphuraria* 5572 mercuric chloride treatment quadruplicate average OD for (A) OD<sub>600</sub>, (B) OD<sub>665</sub> and (C) OD<sub>750</sub>. OD, optical density.

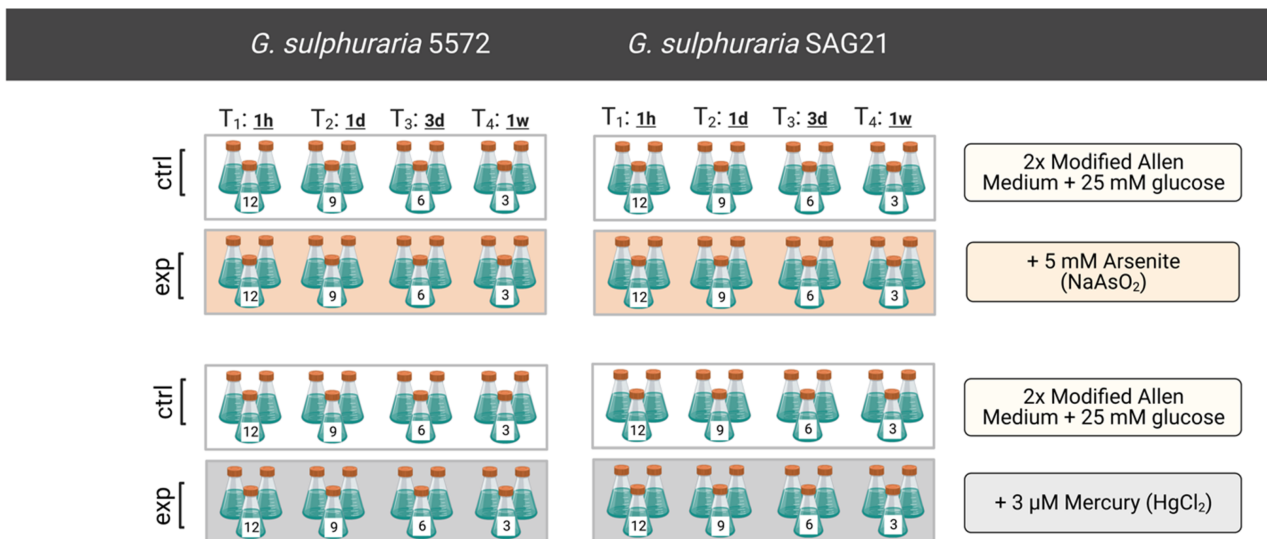


**FIGURE A5** *Galdieria sulphuraria* SAG21 mercuric chloride treatment quadruplicate average OD for (A) OD<sub>600</sub>, (B) OD<sub>665</sub> and (C) OD<sub>750</sub>. OD, optical density.

each experiment are below as Figures A2–A5. Based on colour changes of the cultures, OD fluctuations and visual inspection under the microscope, 5 mM NaAsO<sub>2</sub> and 3 μM HgCl<sub>2</sub> were chosen as treatment concentrations for the RNA-seq experiments.

## Results

For mercuric chloride, both SAG21 and 5572 were resistant to all concentrations of mercury; however, visual inspection of the cells and slight declines in



**FIGURE A6** RNA-seq experimental setup. For safety reasons, all arsenite experiments were run at the same time and all mercury experiments were run at the same time but at different times. This means that for each of these experiments and each strain involved, there was a control group and treatment group (i.e., there is not a shared control group between the two sets of experiments). Both sets of experiments followed an identical run and sampling protocol. Each sampling effort was undertaken by a team of three people, involved 12 flasks at a time, and took under 15 min from start to finish to ensure the minimal effect of the sampling procedure on transcription.

growth rate (measured via OD) indicated potentially slowed growth at 5 and 7.5  $\mu\text{M}$ . For sodium arsenite treatment, both strains were resistant to arsenite toxicity through 5 mM concentrations. 5572 perished at 10 mM after 3 days and SAG21 perished at the same concentration after 1 day. Despite the slightly higher resistance of 5572, OD measurements showed a slowed growth rate for Day 3 at all concentrations  $\geq 300 \mu\text{M}$ . See below for the comprehensive results of these experiments.

### JGI QC pipeline

Using BBDuk (Bushnell, 2014), raw reads were evaluated for artefact sequence by kmer matching (kmer = 25), allowing one mismatch, and the detected artefact was trimmed from the 3' end of the reads. RNA spike-in reads, PhiX reads and reads containing any Ns were removed. Quality trimming was performed using the phred trimming method set at Q6. Finally, following trimming, reads under the length threshold were removed (minimum length 25 bases or 1/3 of the original read length—whichever is longer).

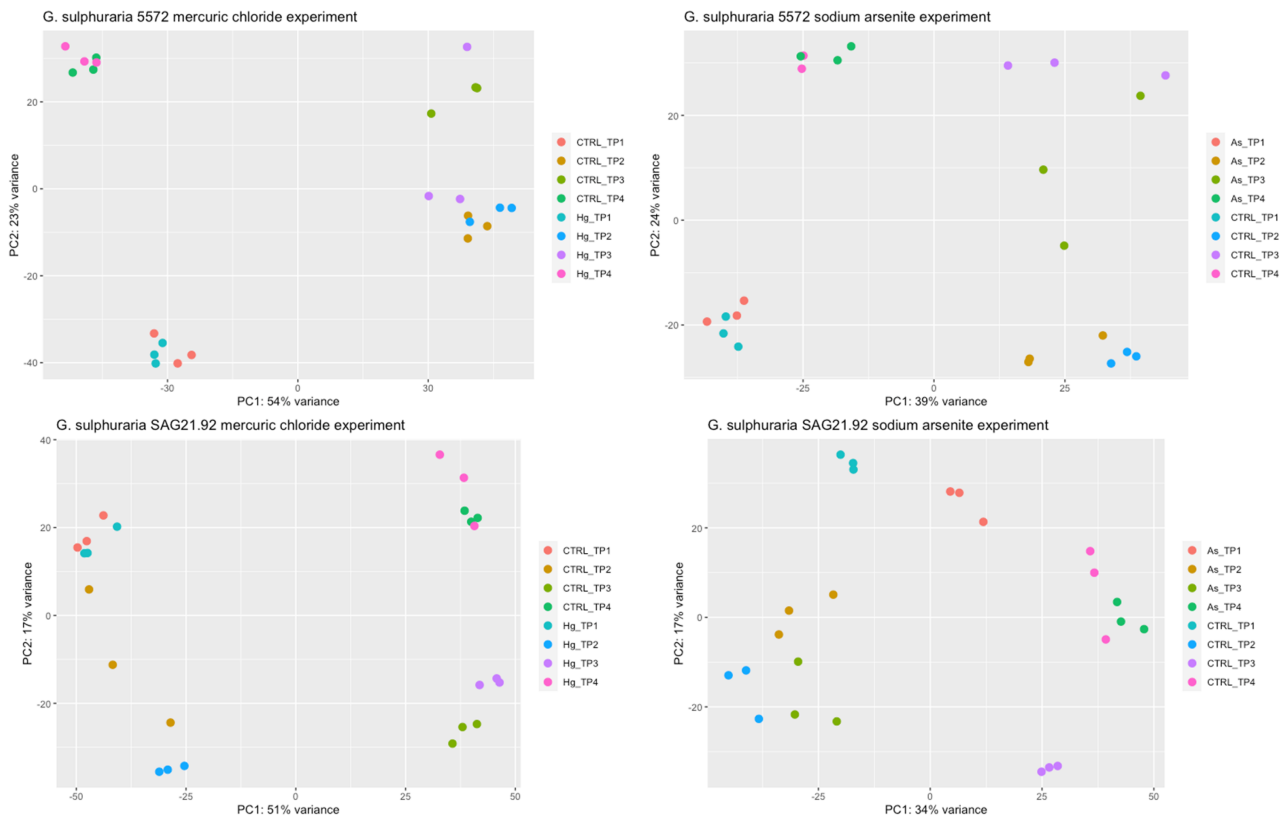
### Unabridged description of WGCNA results

#### Coexpression networks show most heavy metal HGTs are linked to central metabolism

In SAG21, the *arsA* gene (encoding the arsenical pump-driving ATPase) falls into the black coexpression module with many other ATPases, namely the 'archaeal ATPases' identified by Schönknecht et al. (2013) and further characterized by Rossoni

et al. (2019). However, the SAG21 *arsA* gene is on the outside of this module and has lower connectivity to the other genes in this group than the other ATPases have to each other, suggesting it still has its ATPase function but is specialized, in this case to power arsenite (or alternatively, antimonite) efflux (Rosen et al., 1990). Interestingly, in 5572, *arsA* is not associated with the other ATPases but is coexpressed with genes of various functions related to transcription, translation, signal transduction, protein degradation and organelle processes. SAG21 only contains one arsenic transporter gene (*arsB*) in its genome which is in the blue module. Here, it is coexpressed with genes encoding mainly heat shock, transport and translation-associated proteins. One *arsB* paralogue in 5572 (G706, coincidentally also in the blue module) shares a similar functional coexpression profile, whereas the other two paralogues (G983 and G1659) are in the yellow and tan modules, respectively, and are coexpressed with genes that encode a variety of functions such as signal transduction, photosynthesis, transport, protein and lipid degradation, and unknown functions.

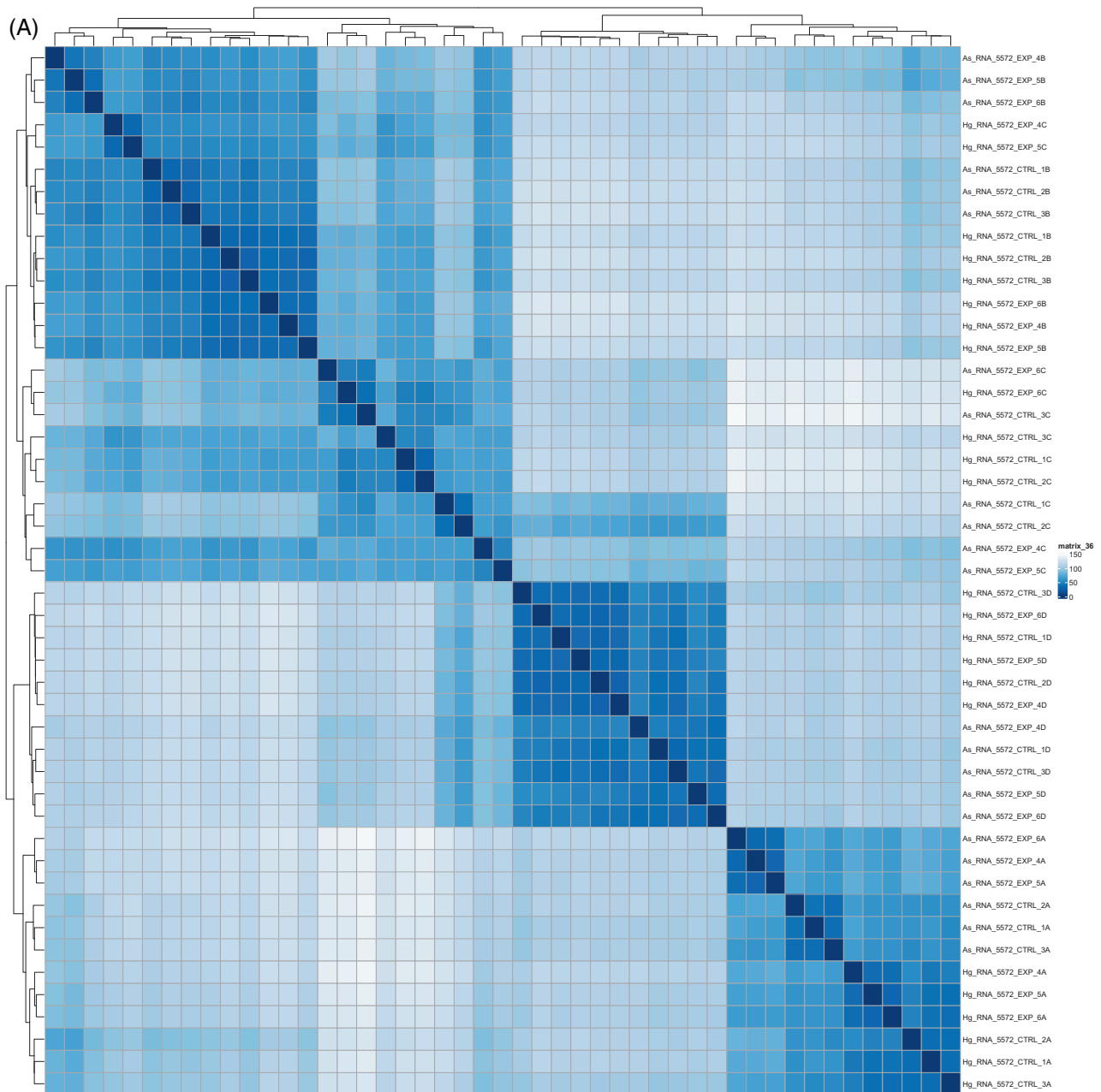
Three *arsH* genes from SAG21 that passed filtering (G286 had low module membership and was excluded) were almost exclusively coexpressed with genes encoding ribosomal proteins or translation-associated functions. The fifth *arsH* paralogue (G2659) in SAG21 is in the RoyalBlue module and is coexpressed mostly with photosynthesis genes. In the 5572 strain, the only *arsH* paralogue (of the two encoded in the genome) to pass filtering was G2956 (homologous to G2659 in



**FIGURE A7** PCA plots showing TreatmentTimeGroup clustering for differential expression data. From DESeq2; see Code 1 file. PCA, principal component analysis.

SAG21) which was placed in the DarkGreen module and is coexpressed with a smaller network of genes that are almost exclusively dark genes of unknown function. In SAG21, the Cyan and Tan modules contain G1984 and G1958 (respectively), the two SAM-dependent methyltransferase HGTs. Both are coexpressed mainly with genes associated with redox processes, RNA and protein metabolism, and energy generation. This is consistent with 5572 which also has two methyltransferases homologous to those mentioned above: G4419, another SAM-dependent methyltransferase, and G4470, annotated as dimethylglycine N-methyltransferase. Both genes fall into the LightYellow module; however, G4419 did not pass module membership filtering. G4470 is associated with genes that encode oxidoreductases and are involved in post-translational modification, lipid synthesis and cell cycle regulation. Only one MIP family aquaglyceroporin (G714 in SAG21 [brown module]) had adequate module membership ( $>0.8$ ) across both genomes. This gene is coexpressed with a handful of dark genes, oxidoreductases, cell signalling and transport genes, as well as others related to DNA and RNA processes. Finally, *arsC* was analysed as its expression can be

interpreted as a type of control for our experiment because adding arsenite to these algae should not have triggered a marked response in the *arsC* gene as it only functions to detoxify arsenate. *ArsC* is only present in 5572 (in two copies, G299 and G388). G388 is still expressed constitutively in high numbers and G299 is expressed in lower numbers but shows some differential expression, despite no arsenate being present to induce a specific transcriptional response. Whereas *arsC* in both genomes does get assigned to a WGCNA module and pass all filtering steps (magenta in SAG21 and brown in 5572), neither of those modules was significantly upregulated based on the WGCNA module-trait association heatmaps generated (Figure A17). Furthermore, there is not a clear functional pattern in either of these modules, that is, magenta has many different functions, and brown is largely translation-associated but has such high connectivity (even when stringently filtered) that it is hard to parse out meaningful functional information. This indirectly supports the results found for those modules discussed above that had statistically significant ( $p > 0.05$ ) module-trait eigencorrelations and the functional suppositions ascribed to them.



**FIGURE A8** These figures show a heatmap of sample-to-sample distances for (A) *Galdieria sulphuraria* 5572 experiments and (B) *G. sulphuraria* SAG21 experiments. To understand the sample labels: As and Hg indicate arsenite and mercuric chloride treatments, respectively. The number indicates the treatment group and the letters A, B, C and D correspond to time points 1, 2, 3 and 4, respectively.



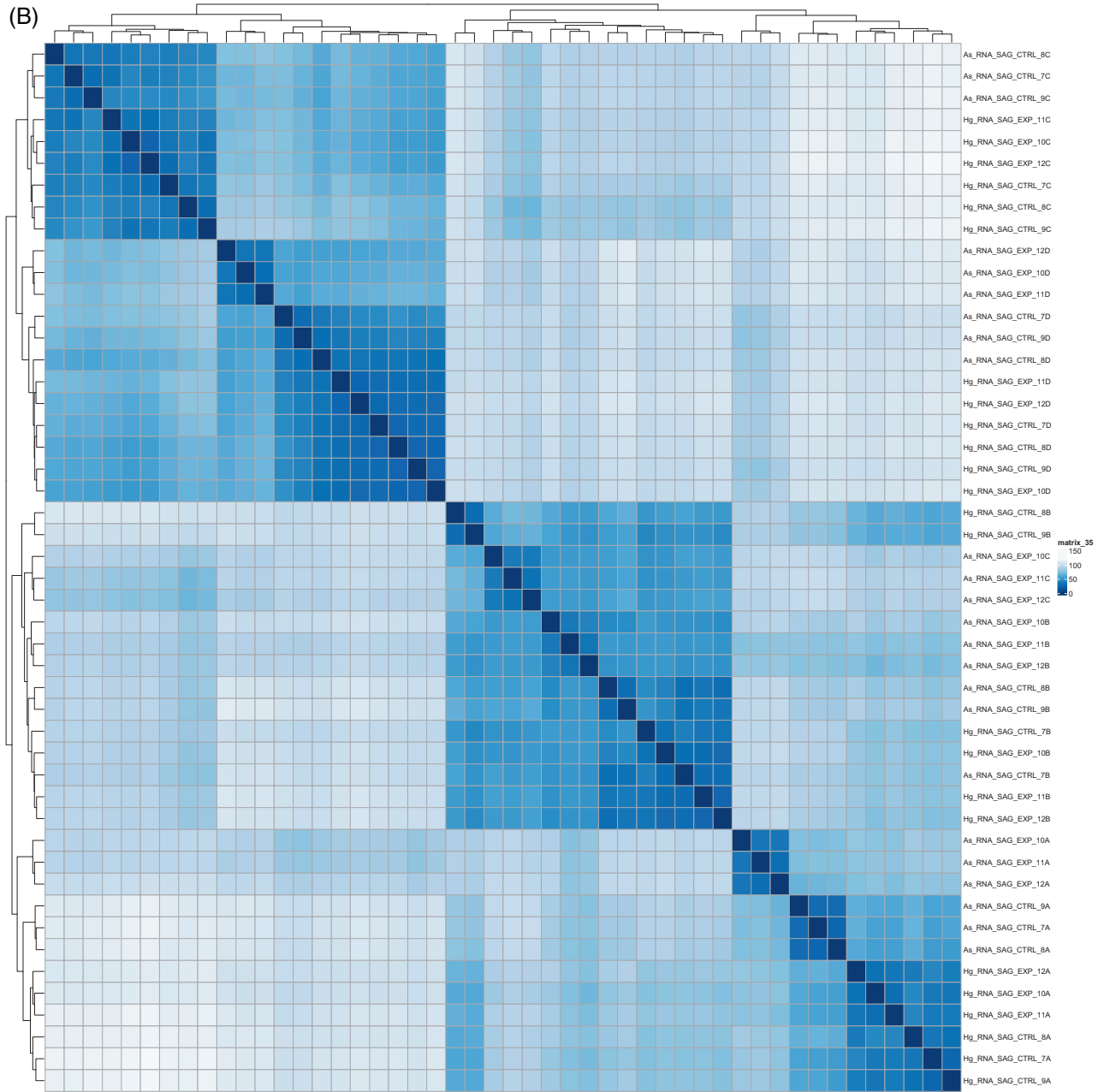
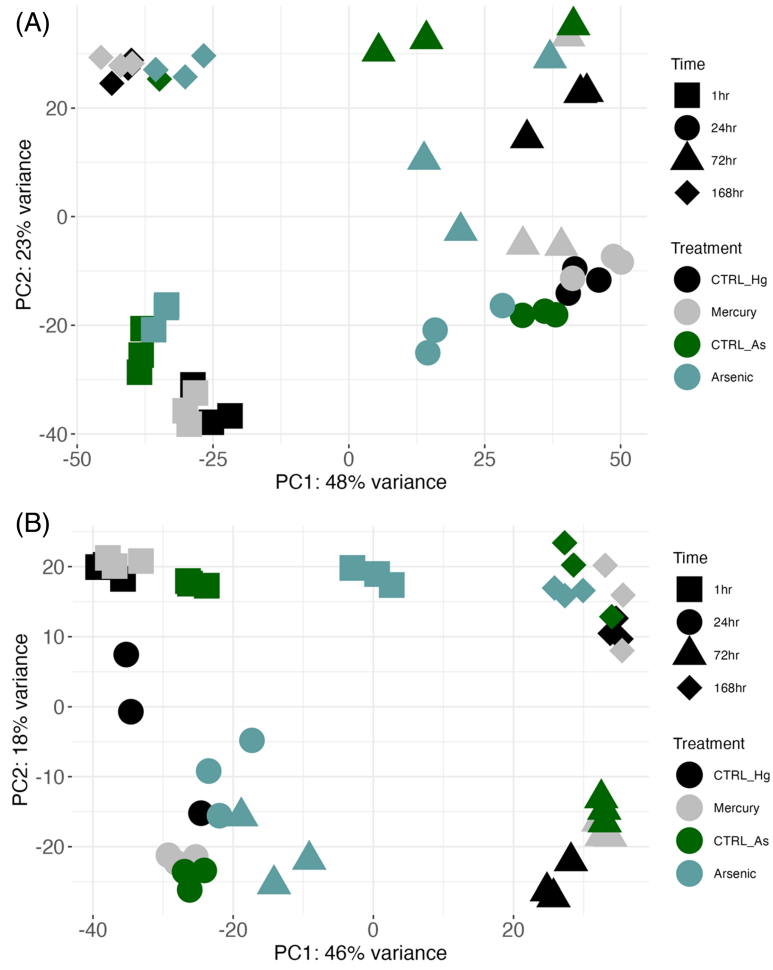
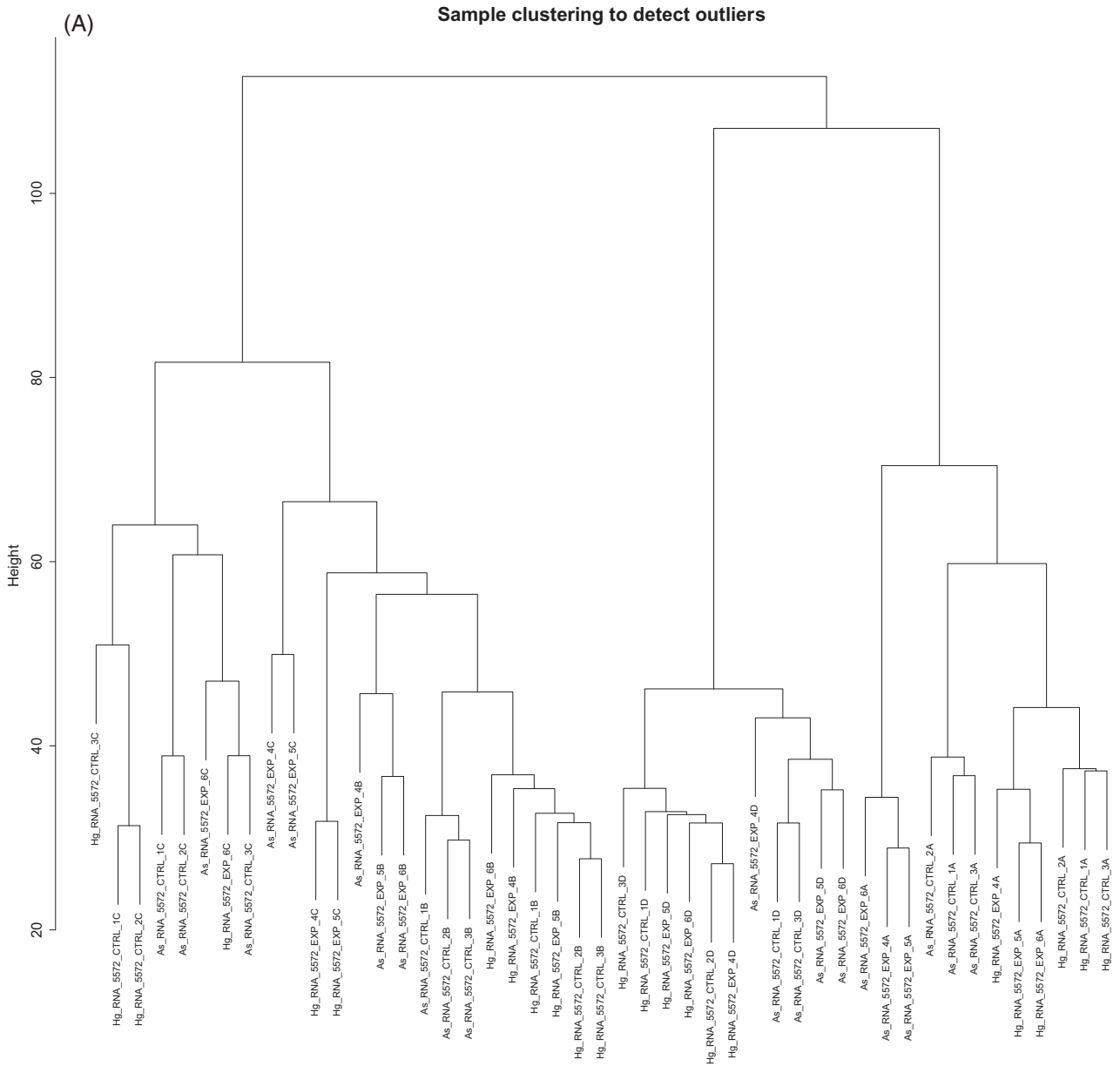


FIGURE A8 (Continued)



**FIGURE A9** PCA plots for samples in (A) 5572 and (B) SAG21 experiments. PCA, principal component analysis.



**FIGURE A10** Sample clustering (hierarchical trees) to look for outliers in (A) 5572 and (B) SAG21. No outliers were detected as most clustering reflected similarities in treatment and/or time points.

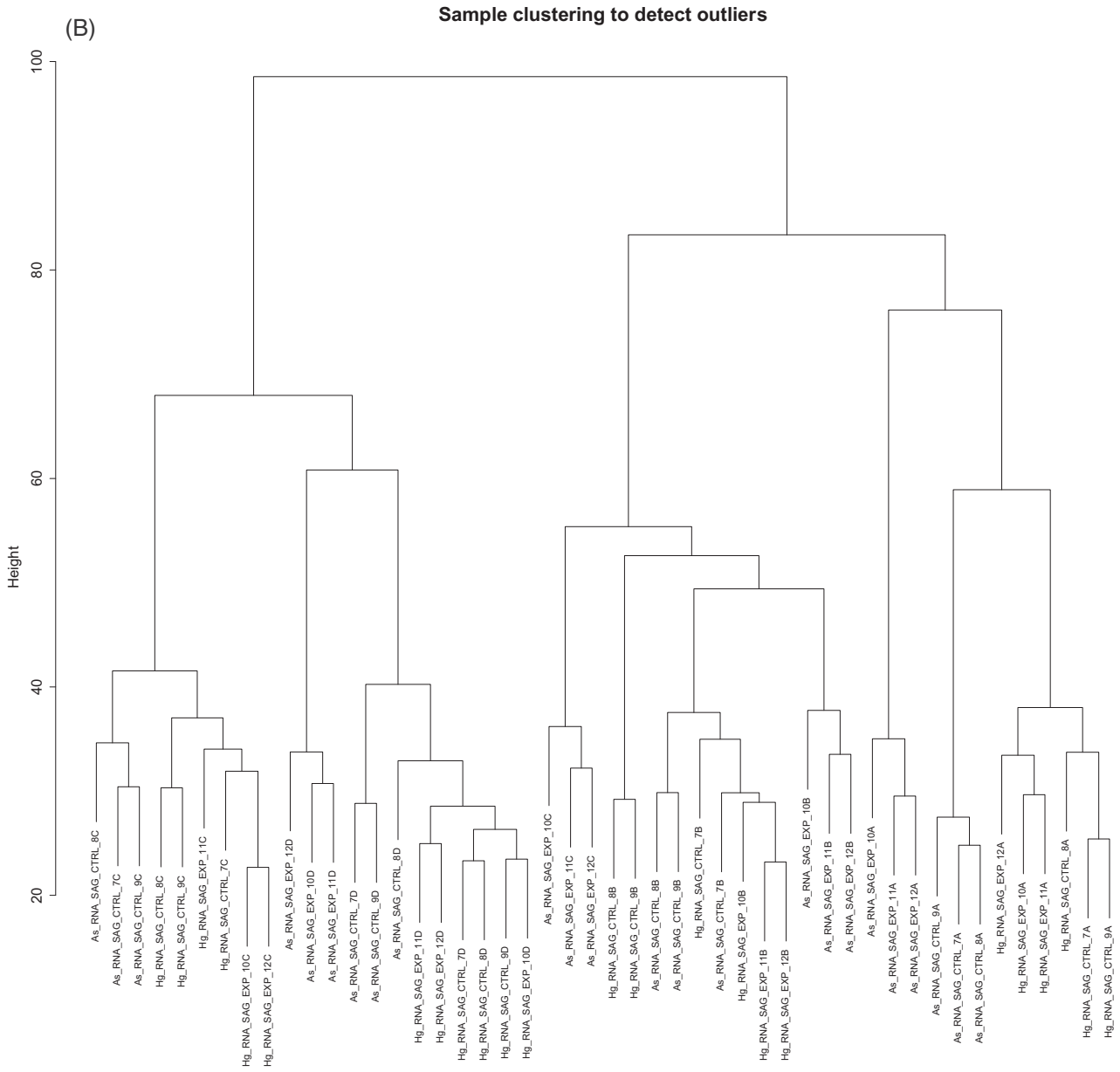
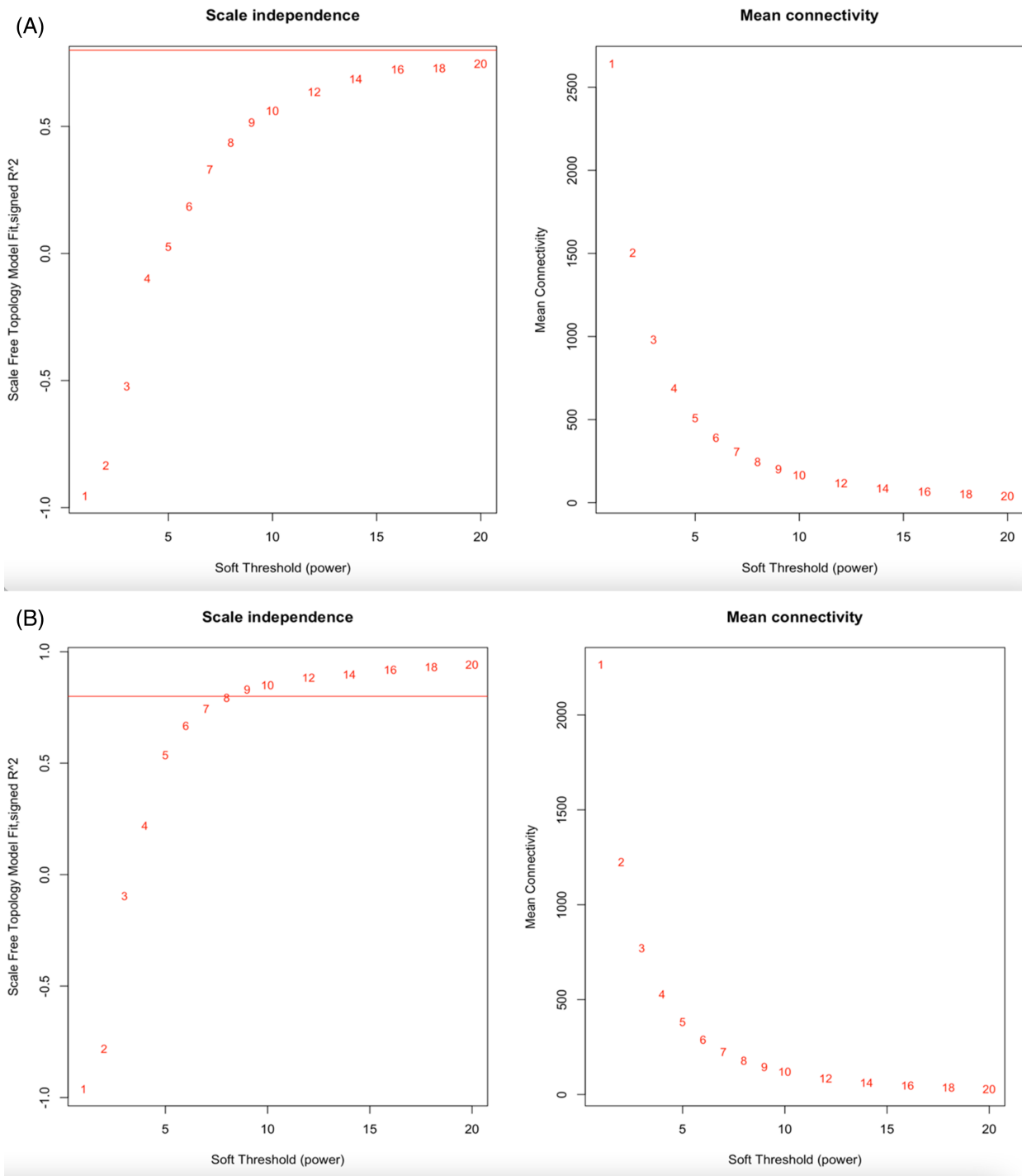
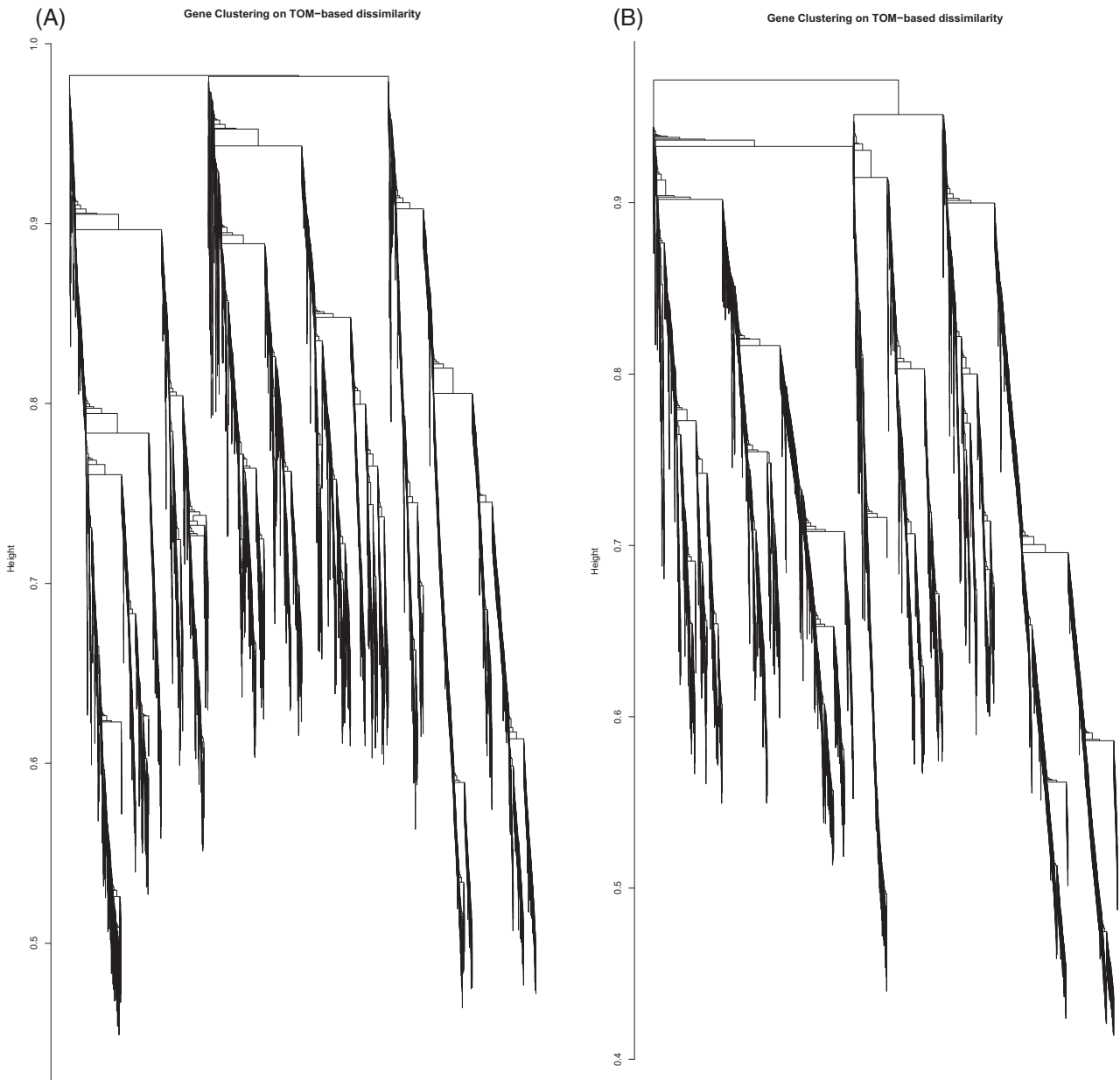


FIGURE A10 (Continued)



**FIGURE A11** Choosing a soft-thresholding power: analysis of network topology  $\beta$ . The soft thresholding power ( $\beta$ ) is the number to which the coexpression similarity is raised to calculate adjacency. The function *pickSoftThreshold* performs a network topology analysis. The user chooses a set of candidate powers; however, the default parameters are suitable values. Choice of soft power threshold for (A) 5572 and (B) SAG21. For 5572, a value of 12 was chosen, and for SAG21, a value of 8 was chosen. This is based on where the graph begins to plateau.



**FIGURE A12** Gene clustering based on TOM-based dissimilarity for (A) 5572 and (B) SAG21. TOM is a 'topological overlap measure'.

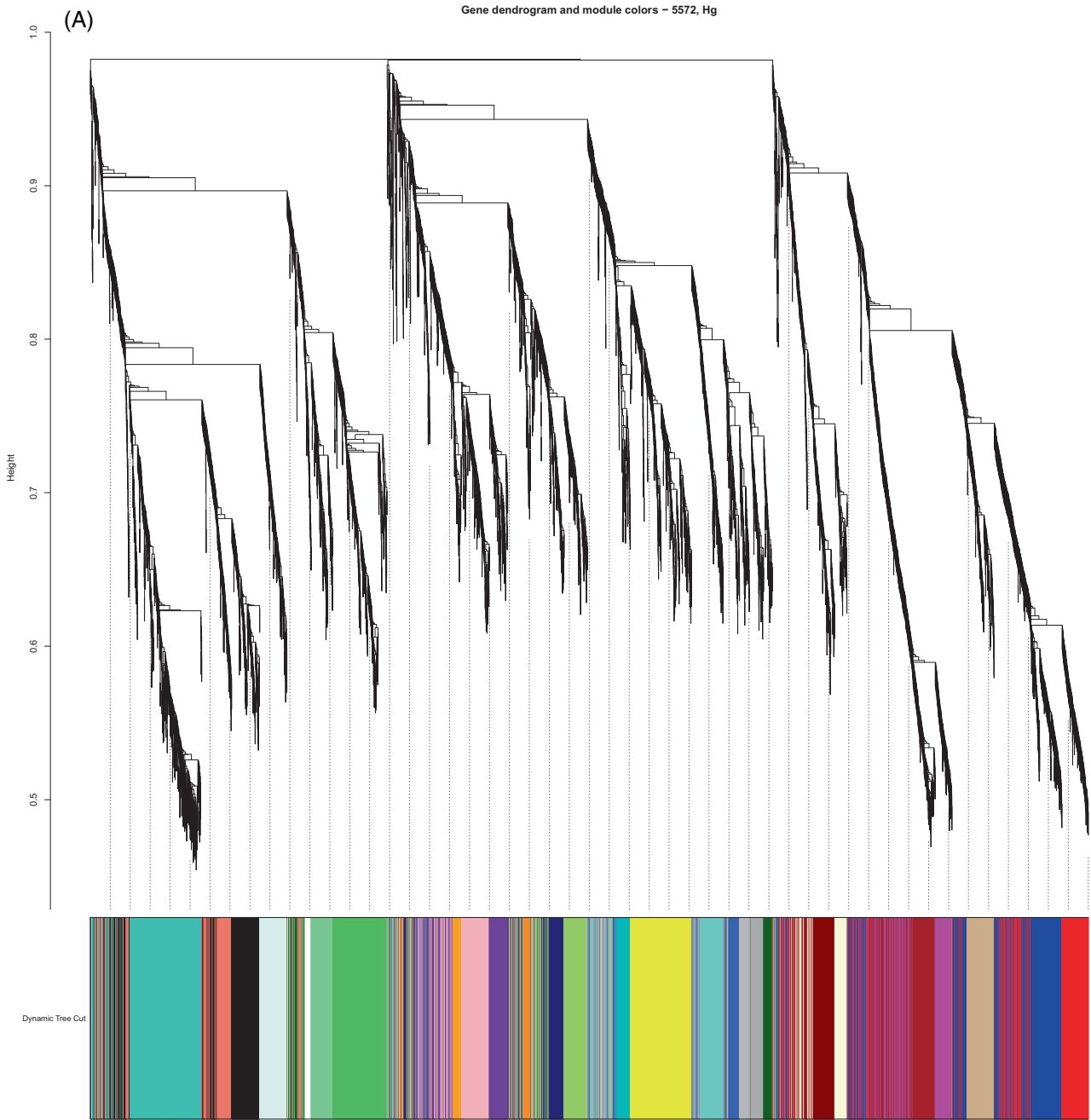


FIGURE A13 Gene dendrograms with corresponding module colours for (A) 5572 and (B) SAG21.

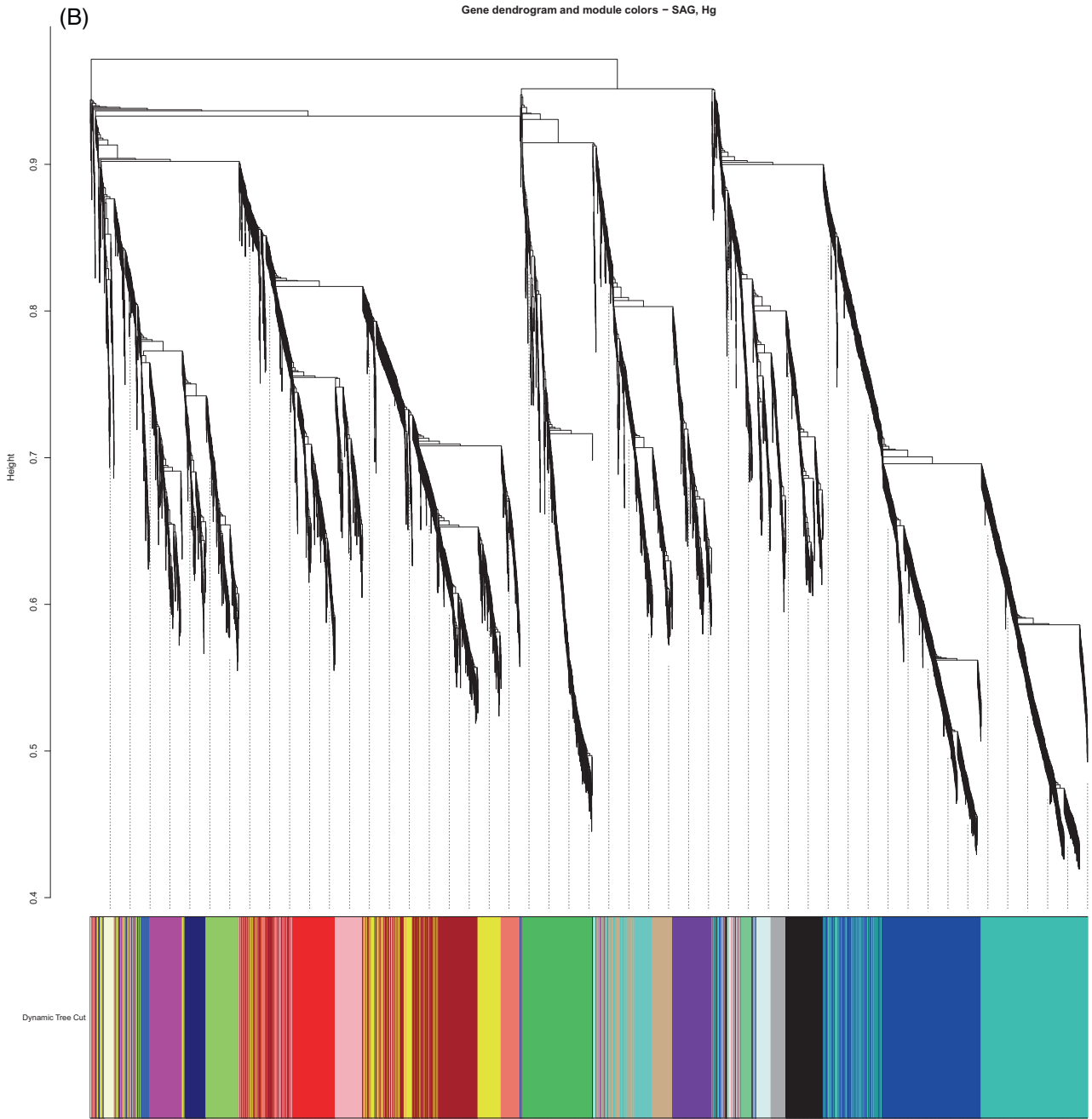


FIGURE A13 (Continued)



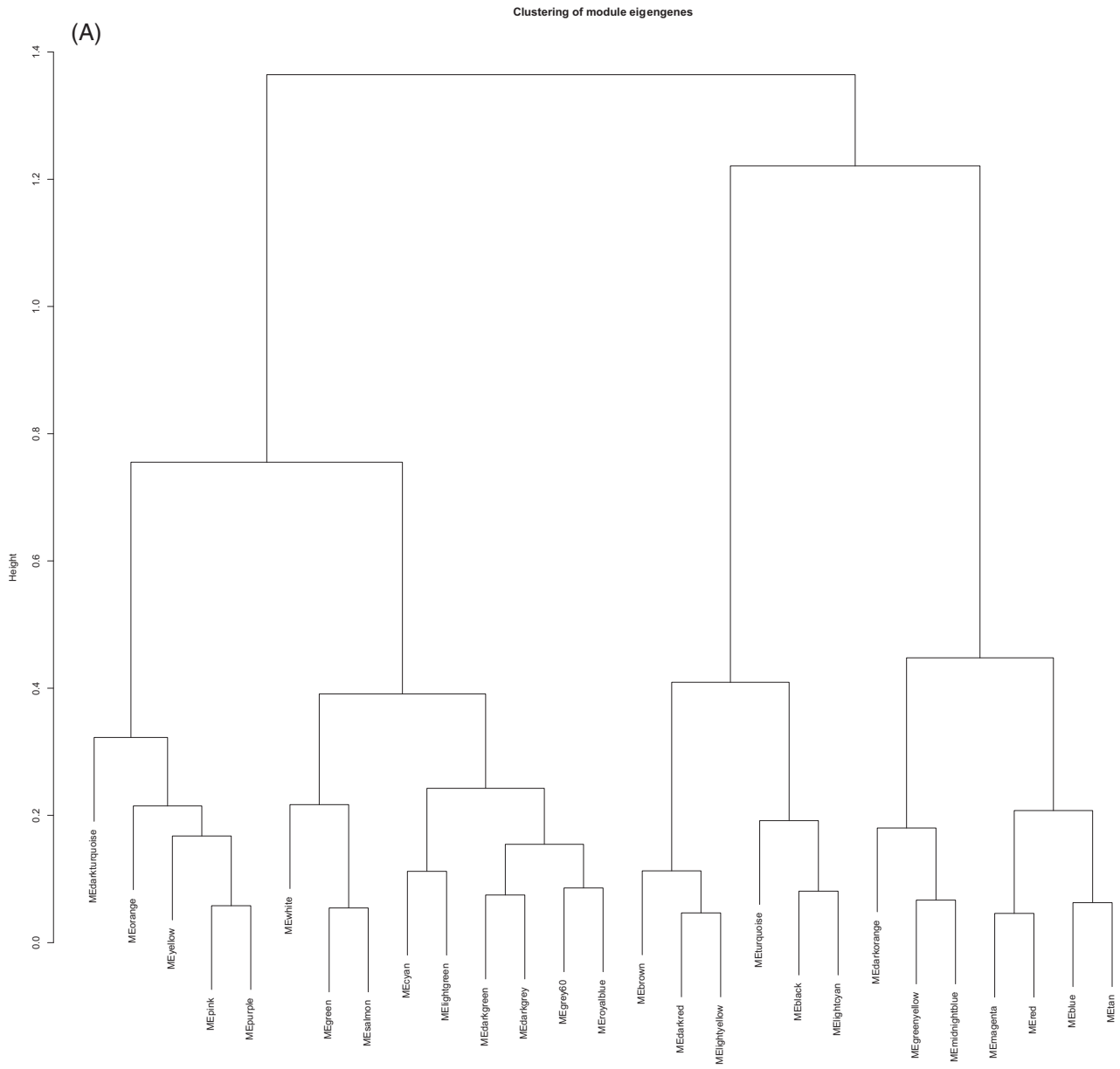


FIGURE A14 Clustering of module eigengens for (A) 5572 and (B) SAG21.

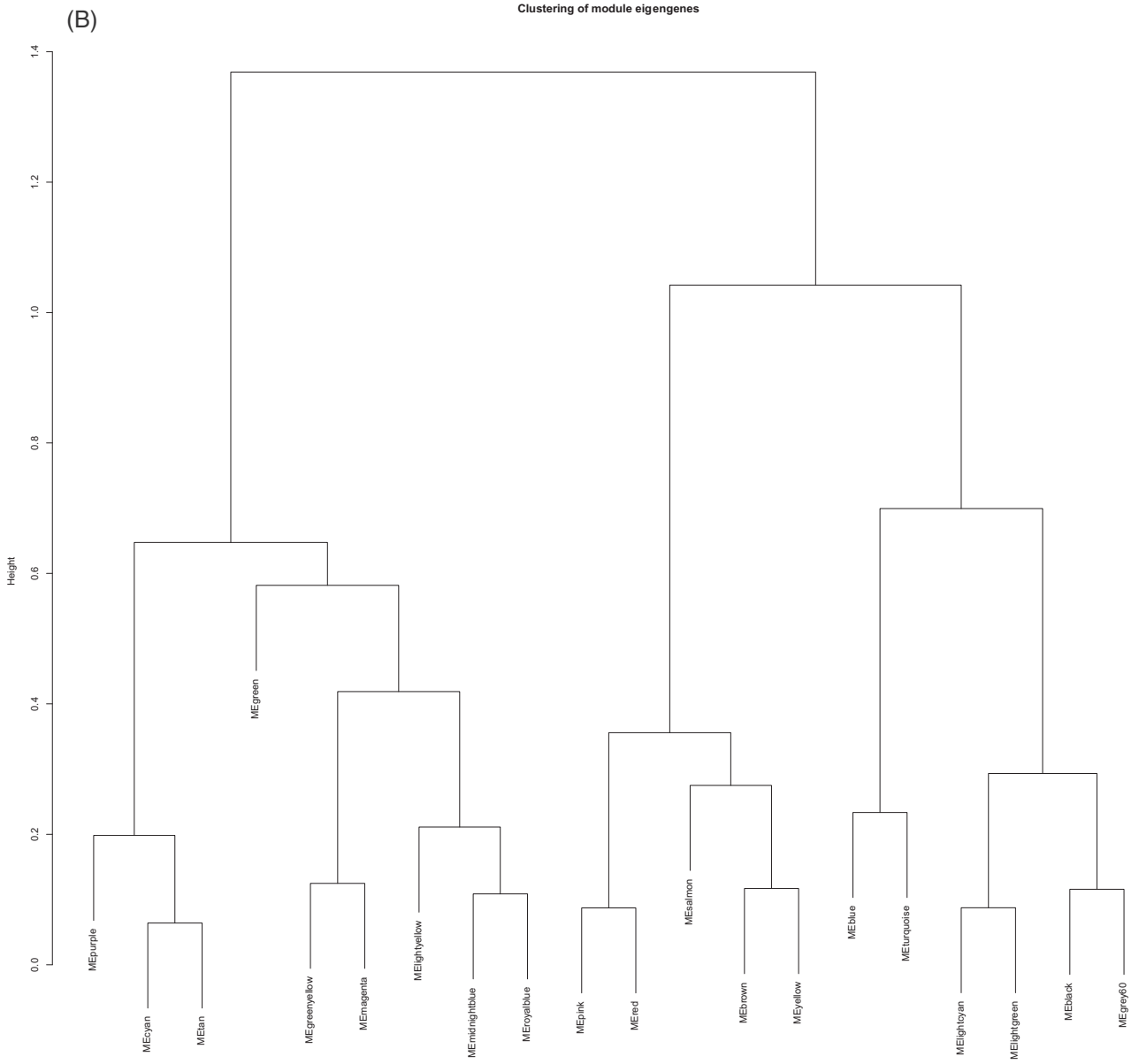


FIGURE A14 (Continued)

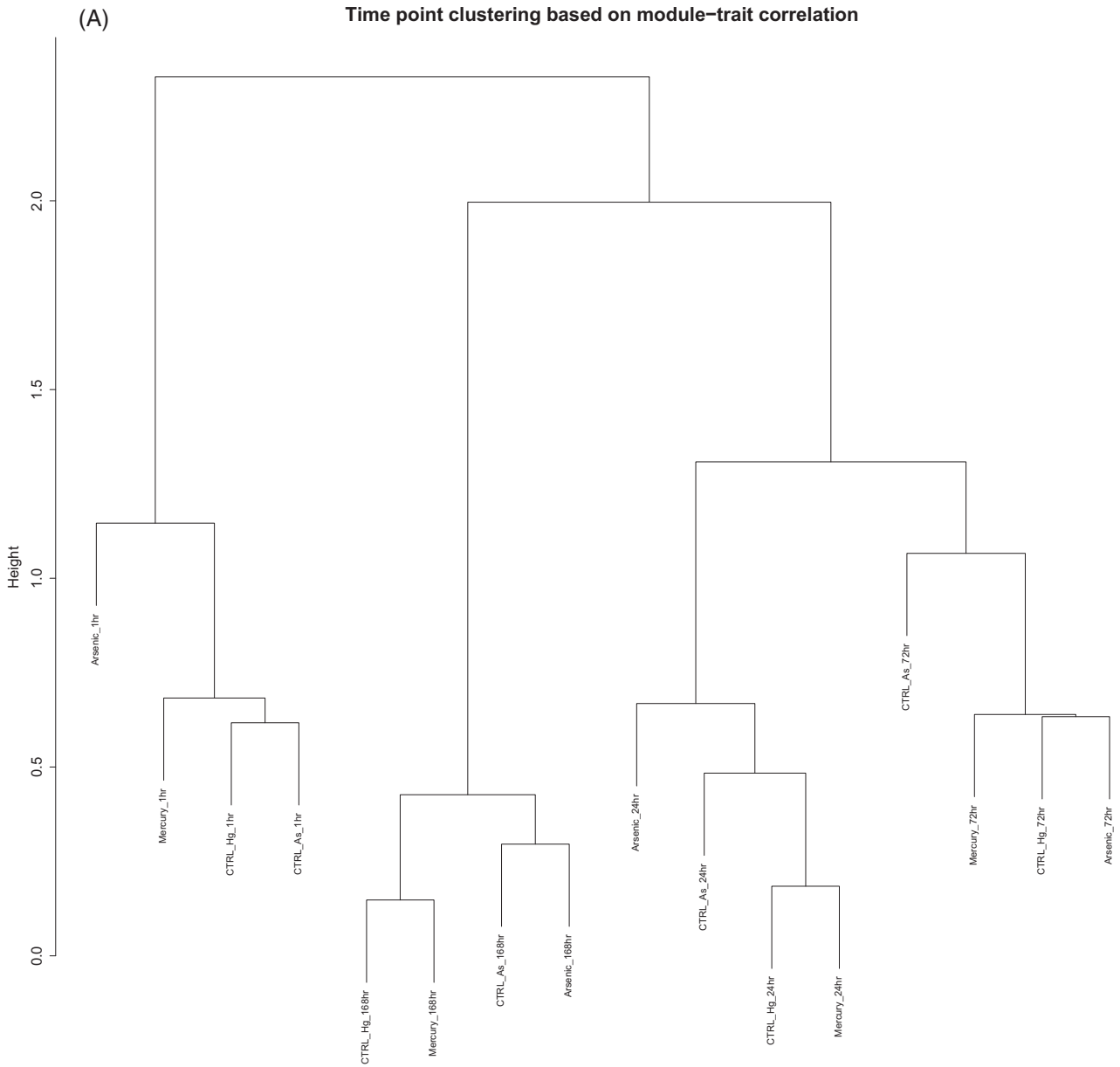


FIGURE A15 Timepoint clustering based on module-trait correlation for (A) 5572 and (B) SAG21.

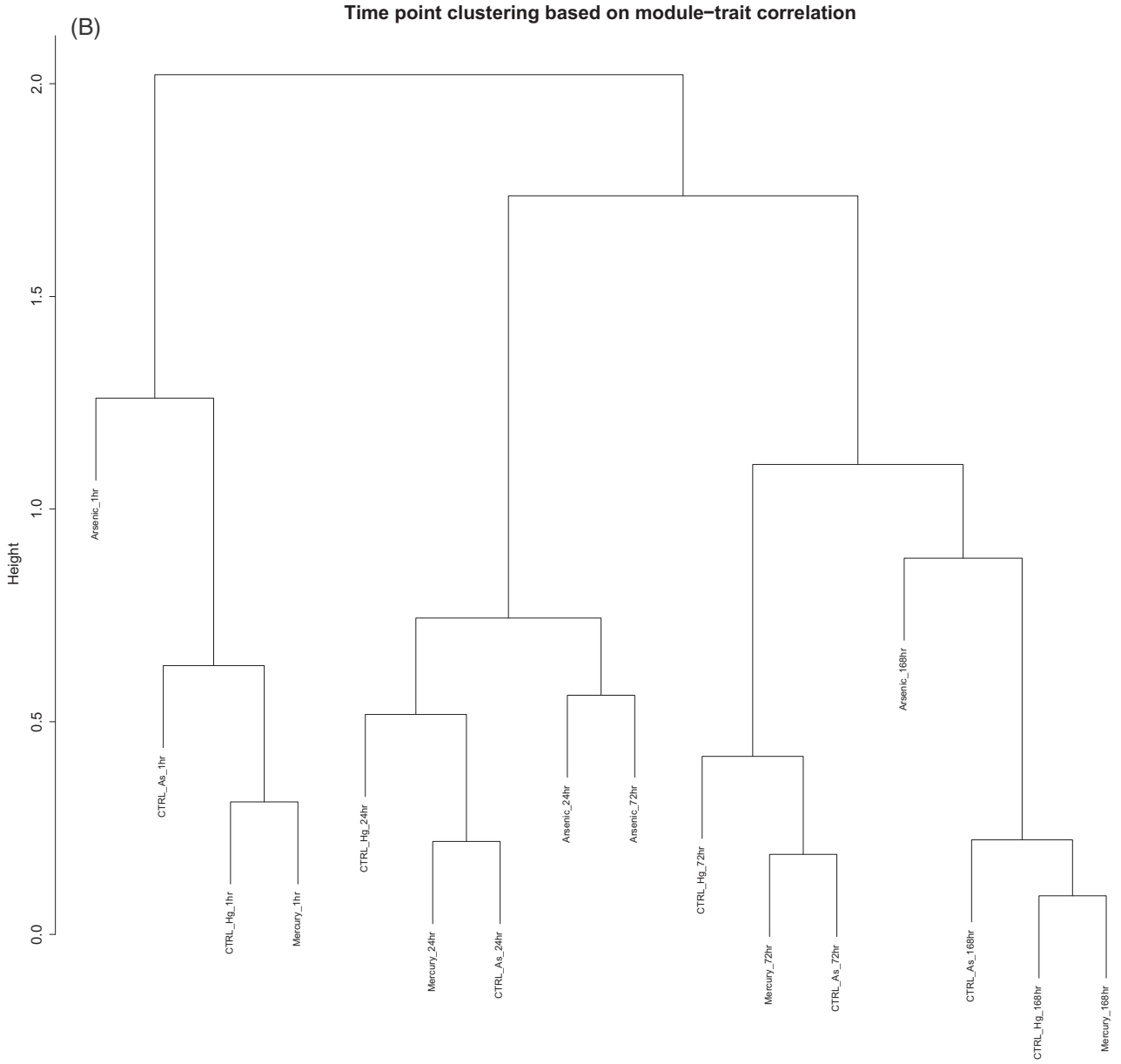


FIGURE A15 (Continued)

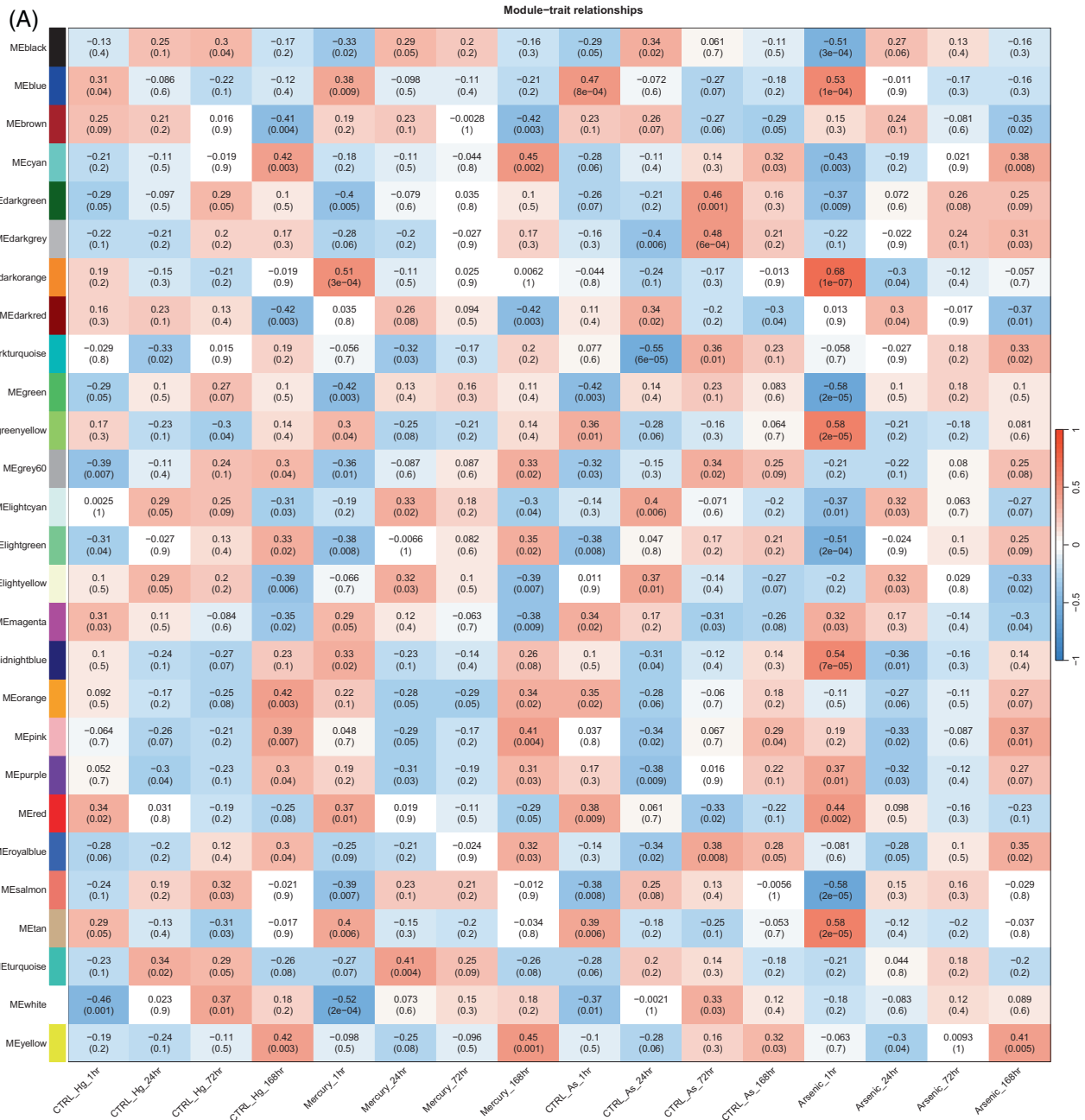


FIGURE A16 Heatmap representing module-trait associations for (A) 5572 and (B) SAG21.

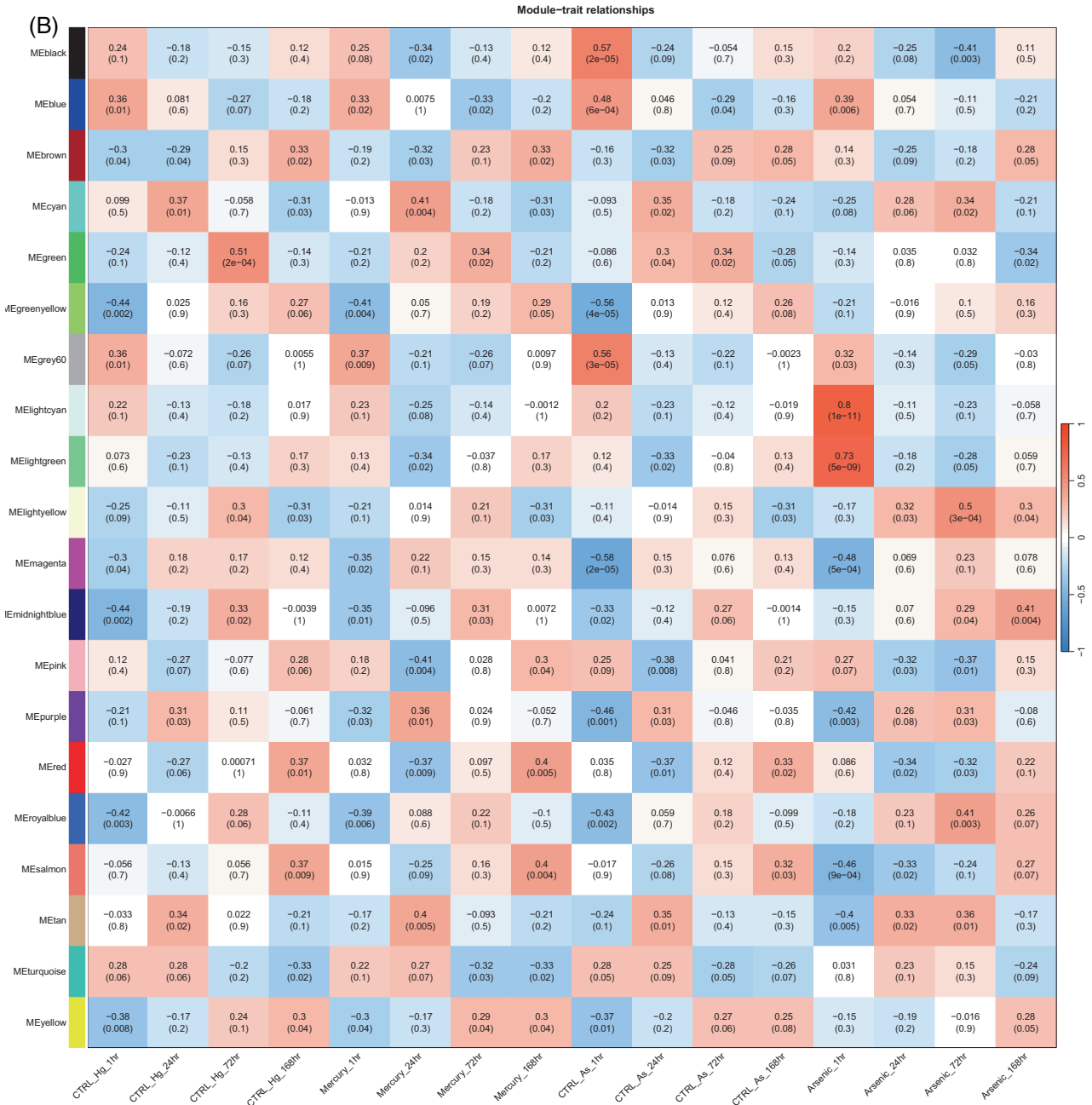
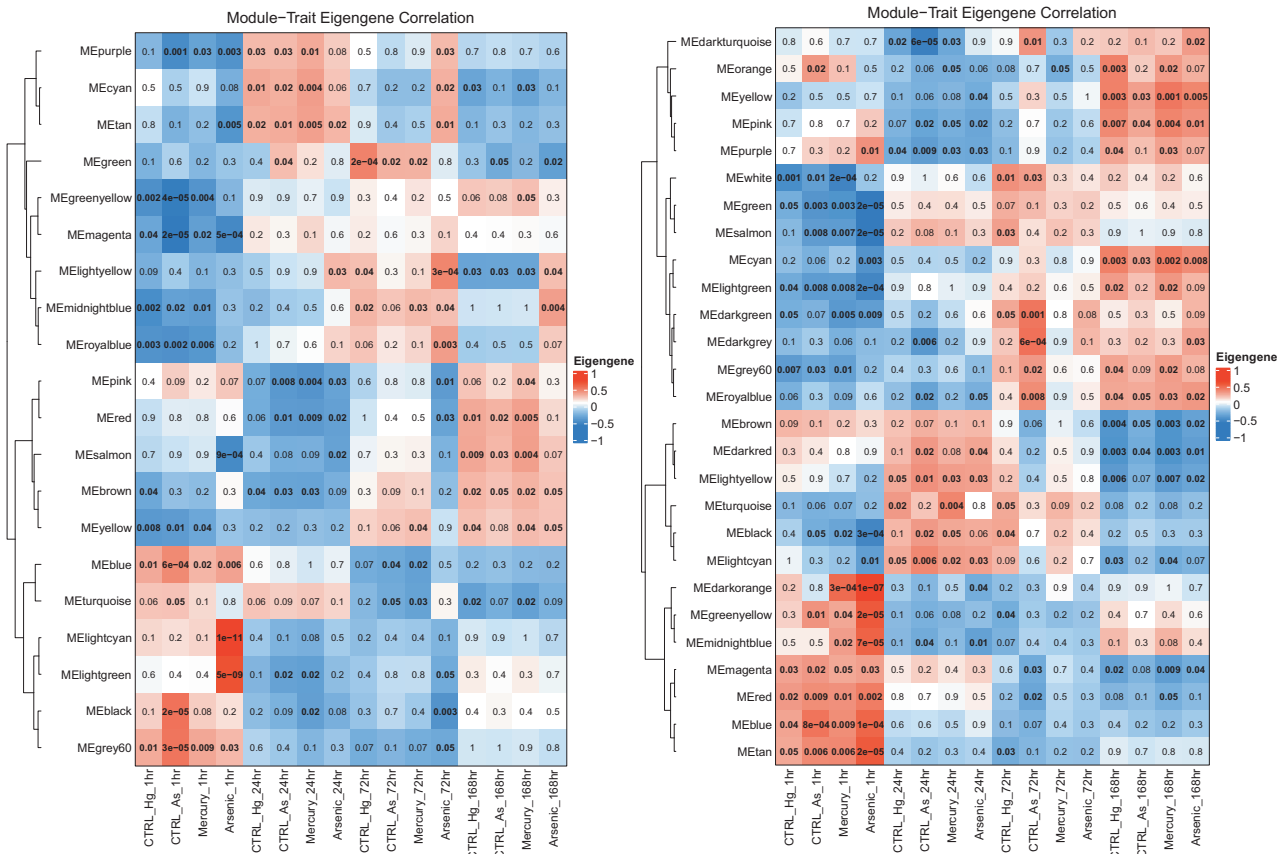


FIGURE A16 (Continued)



**FIGURE A17** Final heatmaps showing module-trait eigengene correlation. (A) SAG21 and (B) 5572. Bold values in the heatmap boxes indicate statistically significant ( $p < 0.05$ ) correlations for the treatment time group (x-axis). If an arsenic or mercury HGT of interest fell into a module (colour, y-axis) that had a significant correlation for any time point, we investigated further; see Figures 3 and A20. HGT, horizontal genetic transfer.

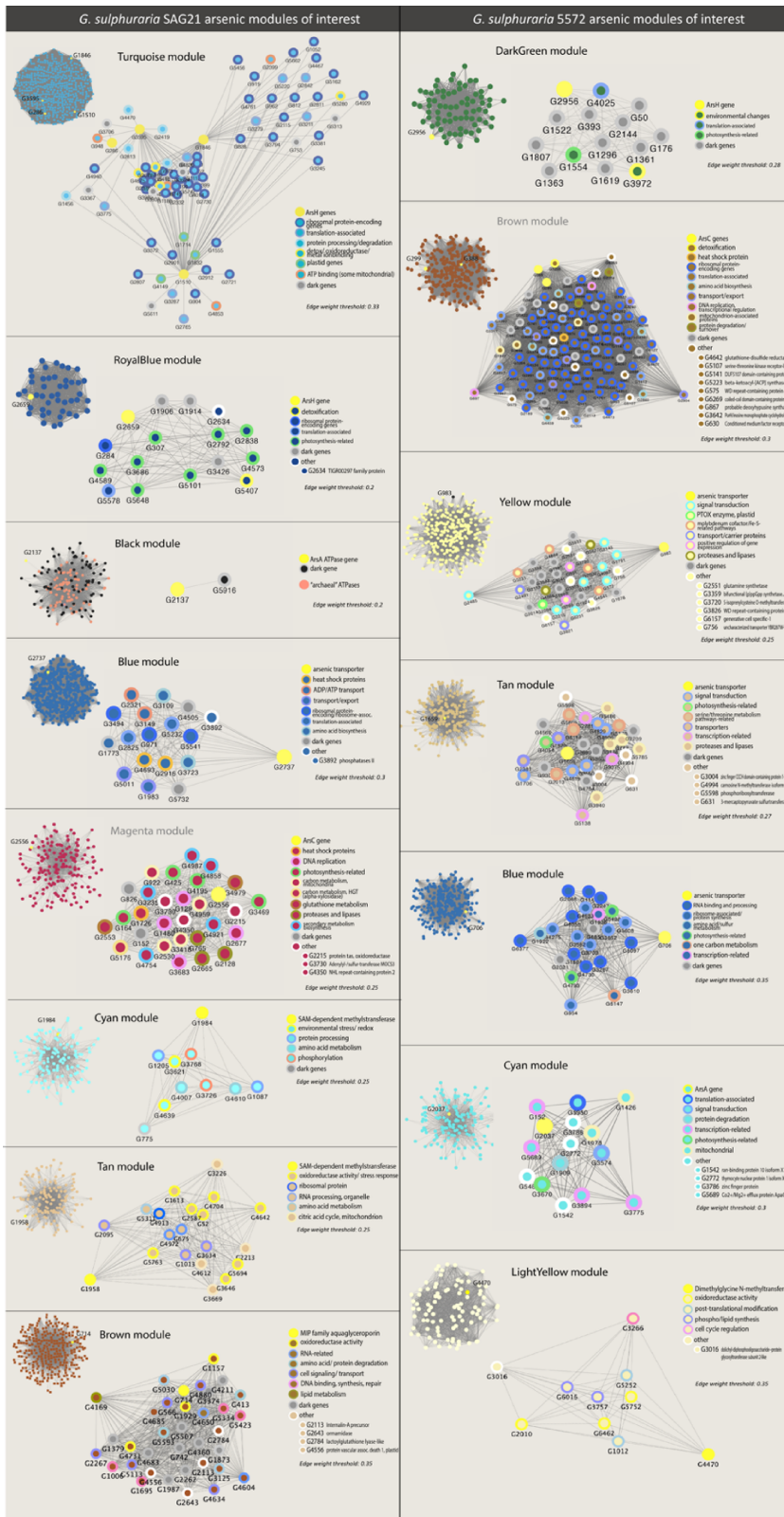
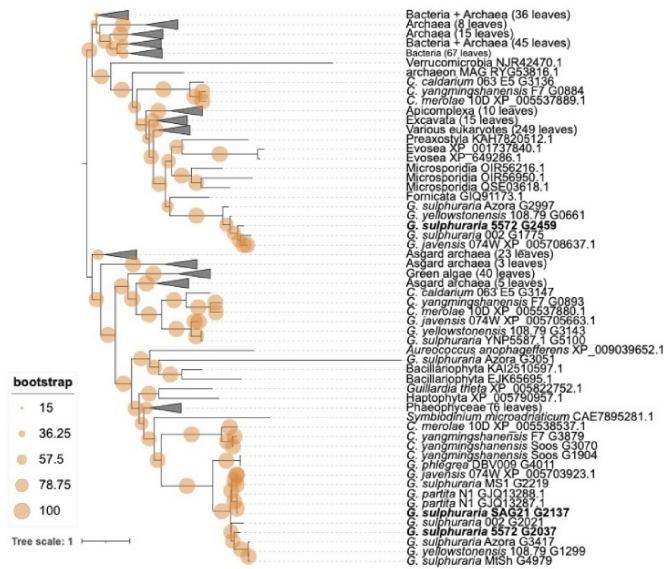


FIGURE A18 Full version of Figure 3, including all genes of interest that were assigned to modules and passed module membership (MM) filtering. Note any genes in Figure 2 that are not here were excluded due to low MM.



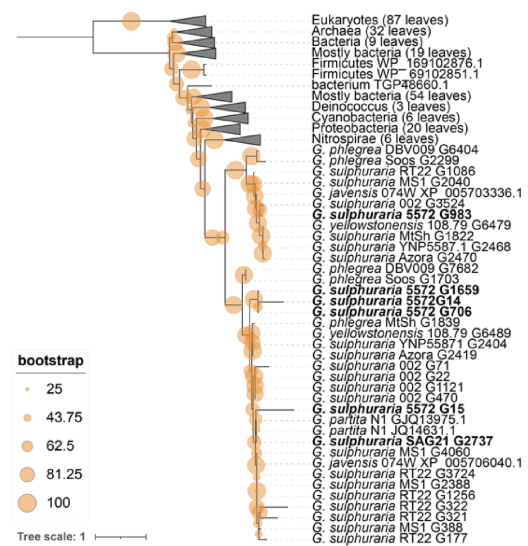
## ArsA - arsenical pump-driving ATPase

(A)



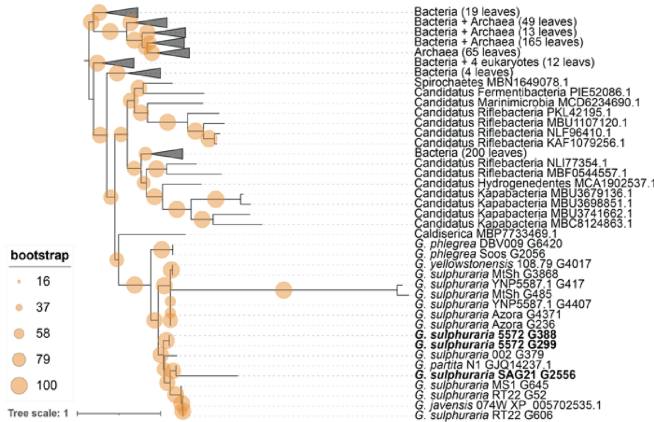
## ArsB - arsenite efflux transporter

(B)



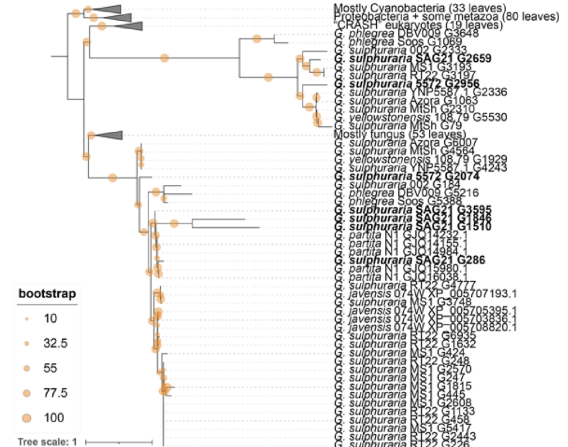
## ArsC - arsenate reductase

(C)



## ArsH - arsenical resistance protein

(D)

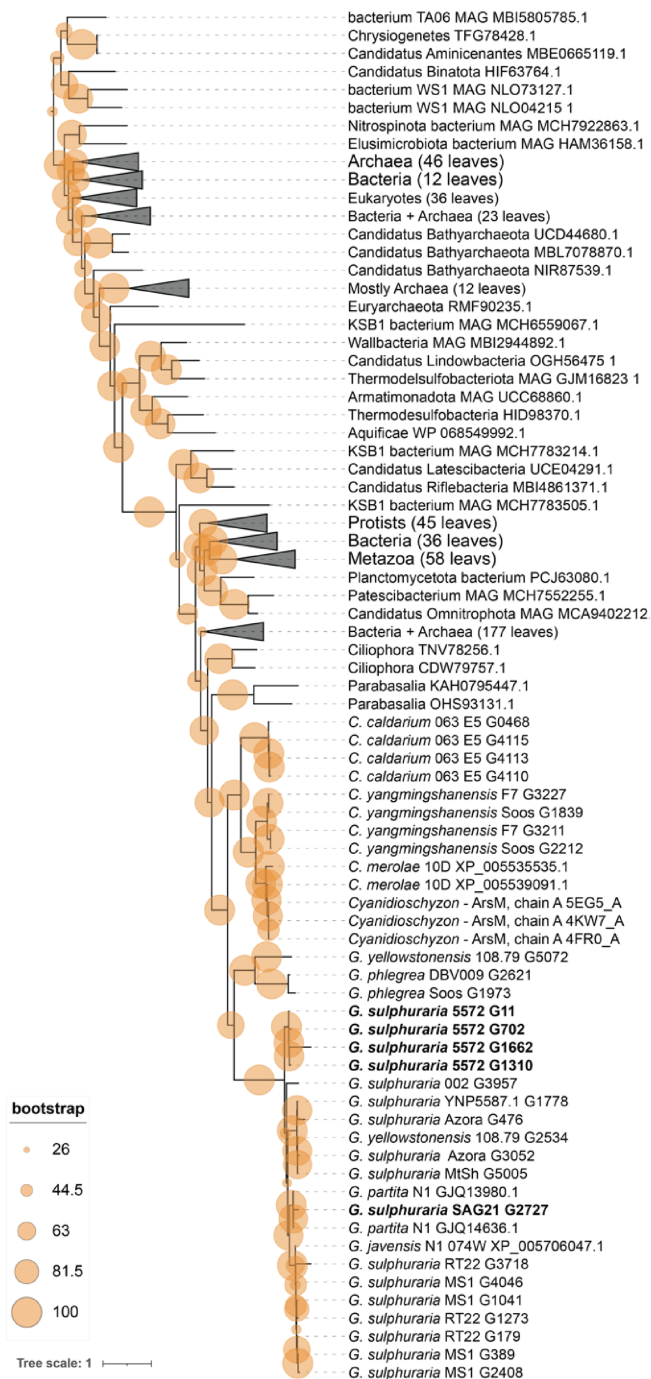


**FIGURE A 19** Single amino acid phylogenies for all HGTs of interest. See methods for how they were built. All Newick tree files can be found as additional supplemental files if you want to rebuild the trees and view the taxa in the collapsed clades. It is interesting to see how the HGT sequences fall within these unrelated species clades. HGT, horizontal genetic transfer.



**Arsm - arsenite methyltransferase**

(E)



**MerA - mercury(II) reductase**

(F)

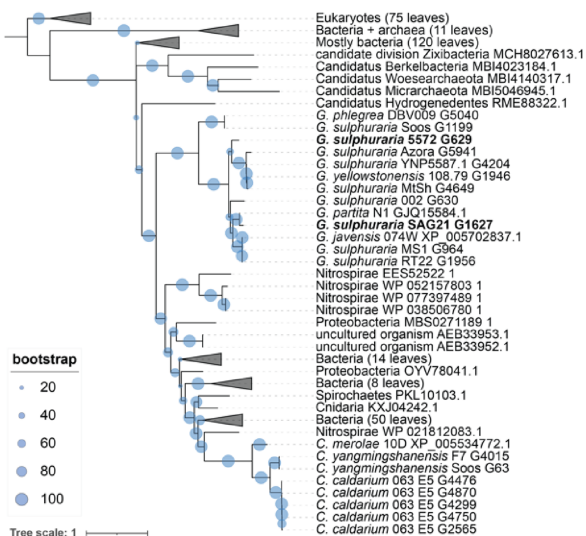
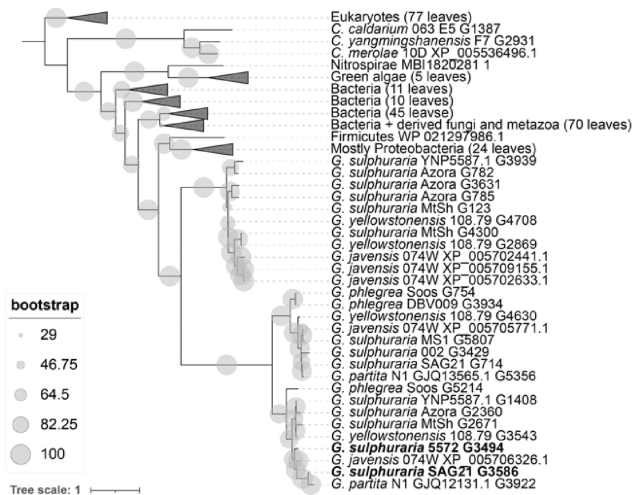


FIGURE A19 (Continued)

## MIP family aquaglyceroporin

(G)



## SAM-dependent methyltransferase

(H)

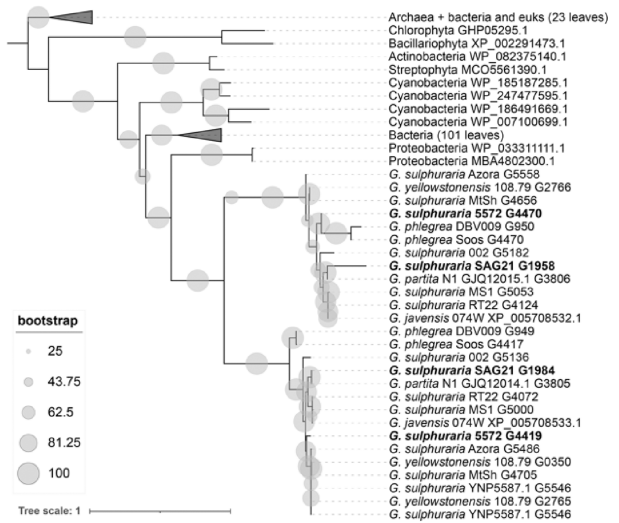
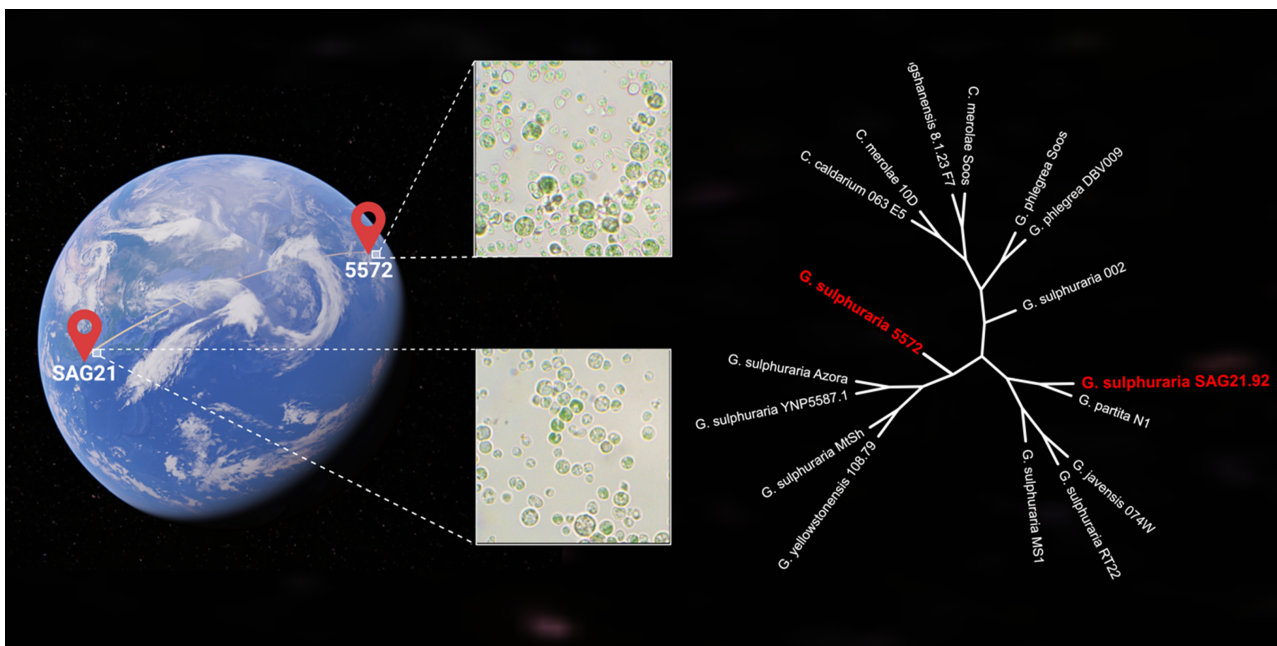


FIGURE A19 (Continued)



**FIGURE A20** Geographic and phylogenetic isolation of SAG21 and 5572. As shown on the left, SAG21 is from Yangmingshan National Park in Taiwan and 5572 is from Yellowstone National Park in the United States. On the right is an un-rooted, alignment-free whole-genome *k*-mer-based phylogeny of all sequenced Cyanodiophyceae, based on the Newick tree file from Van Etten, Cho, et al. (2023) and Van Etten, Stephens, and Bhattacharya (2023).