

Regulatory Element Analysis and Comparative Genomics Study of Heavy Metal-Resistant Genes in the Complete Genome of *Cupriavidus gilardii* CR3

Bioinformatics and Biology Insights
Volume 18: 1–14
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322241299905



Duguma Dibbisa^{1,2}, Tadesse Daba^{1,3} and Seid Mohammed¹

¹Department of Applied Biology, School of Applied Natural Science, Adama Science and Technology University, Adama, Ethiopia. ²School of Biological Sciences and Biotechnology, CNCS, Haramaya University, Dire Dawa, Ethiopia. ³Ethiopia Institute of Agricultural Research Center, Addis Ababa, Ethiopia.

ABSTRACT: Environmental pollution has become a worldwide concern that requires rigorous efforts from all sectors of society to monitor, control, and remediate it. In environmental pollution control, *Cupriavidus gilardii* CR3 has become a model organism to study resistance to heavy metals as a means of bacterial bioremediation. This research aimed to single out regulatory element analysis and conduct a comparative genome study of the heavy metal resistance genes in the complete genome of *C. gilardii* CR3 using bioinformatics and omics tools. Comparative genome analysis, promoter prediction, common motif identification, transcriptional start site identification, gene annotation, and transcription factor identification search are the major steps to understanding gene expression and regulation. MEME Suit, TOMTOM, Prokka, Rapid Annotation utilizing Subsystem Technology (RAST), Orthologous Average Nucleotide Identity Software Tool (OAT), and EziBio databases or programs were the major tools used in this study. Fourteen transcriptional factors were identified and predicted from the most credible and lowest candidate motifs with an e-value of 3.0e-009, which was statistically the utmost remarkable candidate motif. A detailed evaluation was further performed, and 14 transcriptional factors were identified as in activation, repression, and dual functions. The data revealed that most transcriptional factors identified were used for activation rather than repression. The *C. gilardii* CR3 genome contains many genes responsible for resisting heavy metals such as mercury, cadmium, zinc, copper, and arsenate. As a result, regulatory elements will lay a solid basis for understanding genes responsible for heavy metal bioremediation. It was concluded that further studies with wet lab support could be conducted for confirmation. Moreover, other advanced bioinformatics and omics technologies are needed to strengthen the results.

KEYWORDS: Bioremediation, heavy metal, pollution, promoters and transcriptional factors

RECEIVED: August 13, 2024. **ACCEPTED:** October 28, 2024.

TYPE: Research Article

FUNDING: The author(s) received no financial support for the research, authorship, and/or publication of this article.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Duguma Dibbisa, Department of Applied Biology, School of Applied Natural Science, Adama Science and Technology University, P.O. Box 1888 Adama, Ethiopia. Email: chemduguma2013@gmail.com

Introduction

Environmental pollution is the current challenge to sustainable development in the world. The challenge arises from the natural and anthropogenic activities of humankind. Pollution has become a global concern requiring rigorous efforts from all sectors of society to monitor, control, and remediate. As industrialization and urbanization have accelerated over the past century, pollution has become a major problem for both developing and developed countries. The use of pesticides and herbicides in modern agricultural practices also contributes to environmental pollution.¹ Unmanaged industrial waste can leak into soil and groundwater, contaminating the environment with heavy metals. Pollution caused by mining is also one of the main causes of environmental degradation.² Heavy metal effluents from wastewater and agricultural inputs also contribute to major environmental pollution due to their extreme poisoning. Heavy metals are also dissolved in water bodies contaminating ground water, lakes, rivers, and oceans. This causes damage to the food chain and web systems. Environmental pollution threatens humans, animals, and marine life by affecting ecosystem services that support life.¹

Microorganisms have developed specific adaptation capabilities to thrive in harsh environmental conditions.³ Bacteria have a

variety of coping mechanisms despite challenging environmental conditions such as food scarcity, metabolic and biological alterations, and extreme temperatures.³ The mechanisms include extracellular barrier, metal ions transporting actively, extracellular, bioaccumulation, intracellular sequestration, and reduction of heavy metals.^{3,4} *C. gilardii* CR3 is a strain of bacteria belonging to the genus *Cupriavidus*, which is known for its ability to metabolize a variety of substances including heavy metals. It is often studied for its potential for bioremediation particularly in contaminated environments where heavy metals are present. This bacterium has been shown to utilize various carbon sources anaerobically and aerobically through pathways. Therefore, it is used in cleansing up contaminated sites and its role in nutrient cycling in various ecosystems. *Cupriavidus* spp., *Pseudomonas putida*, *Pseudomonas aeruginosa*, *Pseudomonas nitroreducens*, and *Pseudomonas alcaligenes*,⁵ *Klebsiella* species,⁶ and recombinant *Escherichia coli* are few bacteria that have mechanisms to resist cadmium heavy metal.⁷ *C. gilardii* CR3 (see Figures 1–4) encodes all heavy metal resistance genes such as arsenate, cadmium, chromate, copper, mercury, and nickel among *Cupriavidus* spp. On the contrary, *C. gilardii* CR3 encodes for genes on conjugative elements, which are chromosomally integrated. For efficient uses of bacteria in HMR, temperature is one of the utmost essential



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (<https://creativecommons.org/licenses/by-nc/4.0/>) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (<https://us.sagepub.com/en-us/nam/open-access-at-sage>).

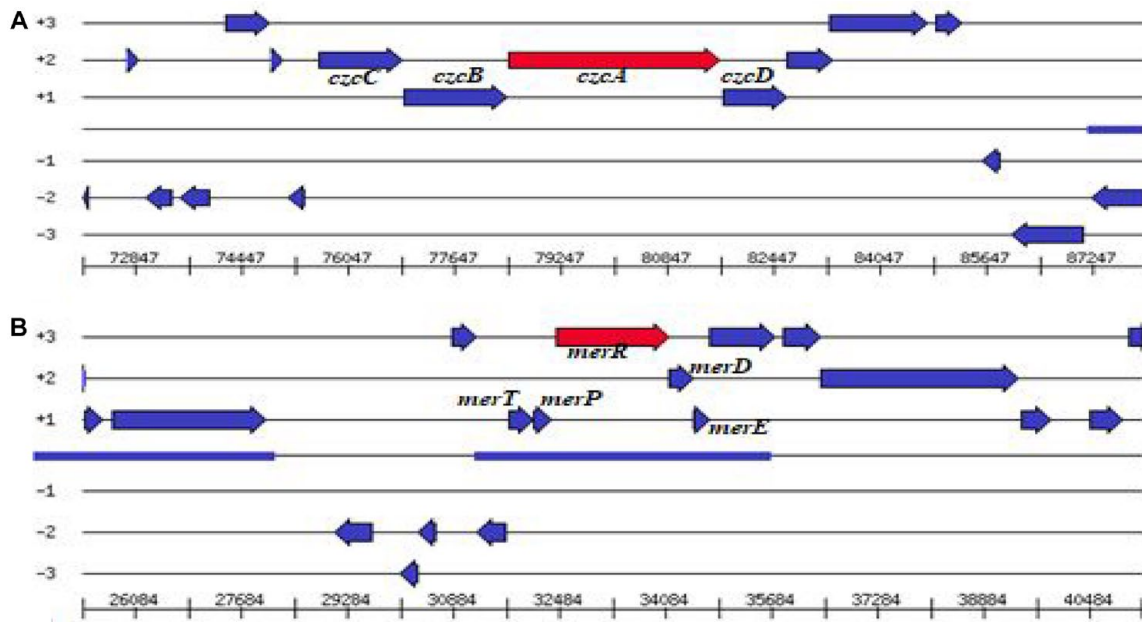


Figure 2. (A) *Czc ABC* family, (B) mercury operon families, and Cobalt-Resistant and Copper-Zinc- and Cadmium-Resistant gene families predicted by RAST. RAST indicates Rapid Annotation utilizing Subsystem Technology.

Resource (<https://meme-suite.org/meme/tools/meme>).¹⁹ The common candidate motifs were identified as binding sites of transcriptional factors (TFs) to regulate HMR gene expressions. The MEME-Suite was utilized for motif identification and exploration, motif alignment investigation, motif scanning, and motif comparisons.²⁰ Statistically, significant candidate motifs predicted by MEME-Suit in the sequence were picked and used for further comparison. The MEME-suit predicted and discovered gene sequences that fit the best motifs (fixed-length repetitive patterns with the lowest e-values) were submitted to TOMTOM online database to obtain TFs. This technique was utilized to discover common motifs that served as binding spots for TFs believed to influence HMR gene expression stages. However, before beginning the search for the selected sequences, the motif distribution menu's basic criteria for searching were defined. These included motif distribution, motif positions, zero or one incidence per sequence, or multiple incidences per sequence.

Furthermore, the number of motifs and persistent motif width (6–50 bps) were set to default. Following the completion of the MEME searches, a search results page was linked to the MEME output in HTML. This stage is an important preliminary point of view for the expected value (e-value). The smaller the e-value, the better result was expected as reported in earlier studies.²⁰ At the bottom of the MEME HTML output, one or more potential motifs were provided for additional investigation. TOMTOM Web server database tool was used to identify common TFs. The sequences matching with the best candidate motifs in their respective TF were identified for the

imported gene sequences. As part of TOMTOM output, LOGOSS represents the alignment of the candidate motif and TF with the *P*-value and *q*-value (a measure of false discovery rate) and links back to the parent transcription database for more detailed information on the matches in TOMTOM output.^{20,21}

Comparative genome analysis of C. gilardii CR3 genome by Prokka and Rapid Annotation utilizing Subsystem Technology

The sequences responsible for the HMR were obtained from the NCBI database, which is freely available at <https://www.ncbi.nlm.nih.gov/>. Heavy metals such as arsenate, cadmium, mercury, chromium, zinc, and cobalt resistance of DNA sequences were used in the study. Comparative genome analysis of *C. gilardii* CR3 (CP010516.1), *Cupriavidus metallidurans* CH34 (NC_007973.1), *Cupriavidus nantongensis* X1 (CP014844.1), *Cupriavidus* ISTL7 (CP066227.1), *Cupriavidus* MP-37 (NZ_CP085344.1), and *Cupriavidus gilardii* WM02 (CP104386.1) for the HMR was performed by using Rapid Annotation utilizing Subsystem Technology (RAST) online engine accessed through (<https://rast.nmpdr.org/rast.cgi>). The RAST is a prokaryotic genome annotation service used to re-annotate for gene prediction.¹⁸ Similarly, a *C. gilardii* CR3 circular genome analysis was performed using CGView (Circular Genome Viewer) a Java application and library for generating high-quality and zoomable maps of circular genomes.¹⁷ Finally, gene expression, gene sequences, location, and function are

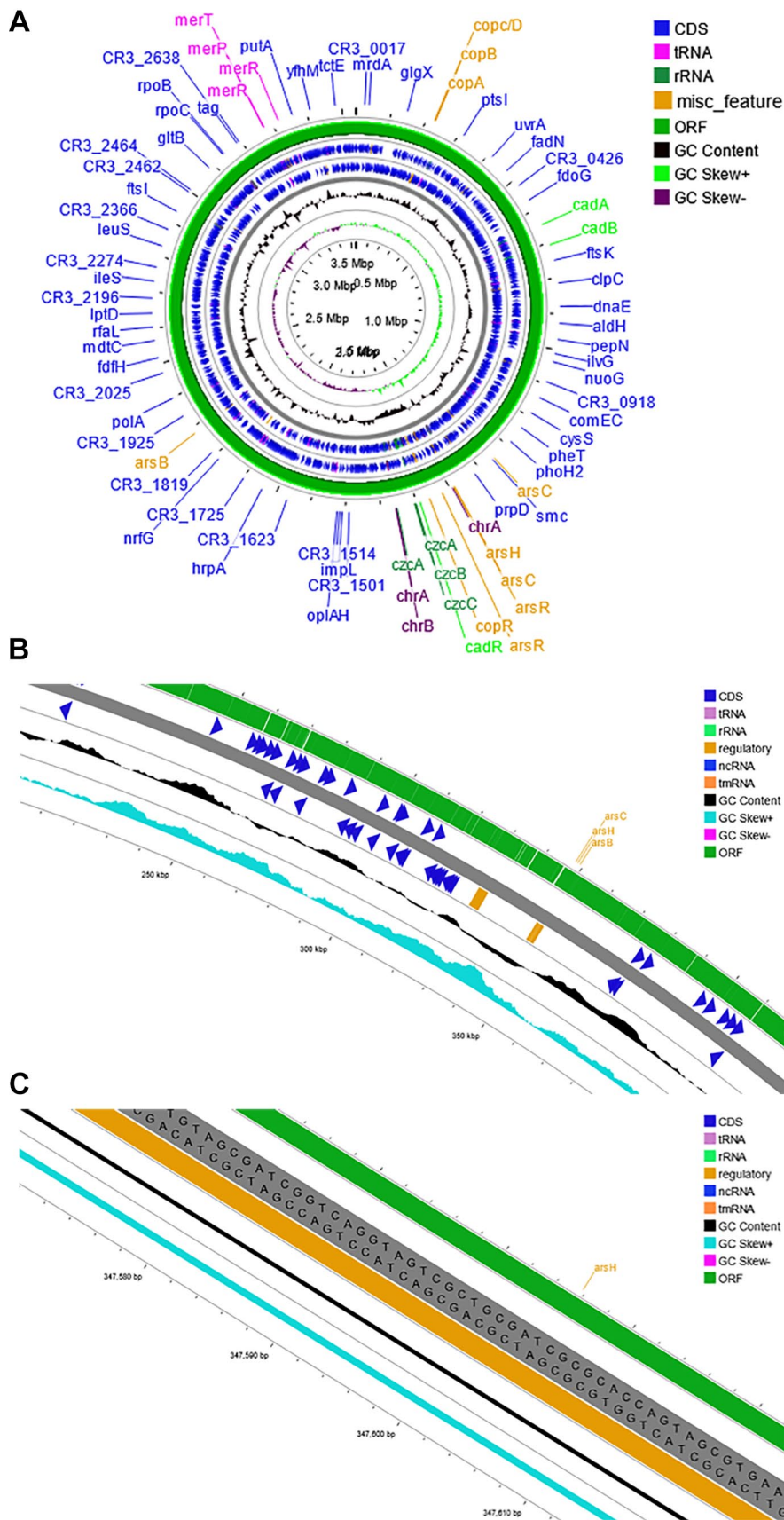


Figure 3. (A) Graphic circular genome mapping for *C. gilardii* CR3 and a sequence obtained from the NCBI database with accession number CP010516.1. (B) Zoomed circular genome map for *C. gilardii* CR3 obtained from NCBI. (C) Zoomed circular genome map with the location of sequence length and/or contig, stop, start, ORF, GC skew⁺ and GC skew⁻. ORF indicates open reading frame; RAST, Rapid Annotation utilizing Subsystem Technology.

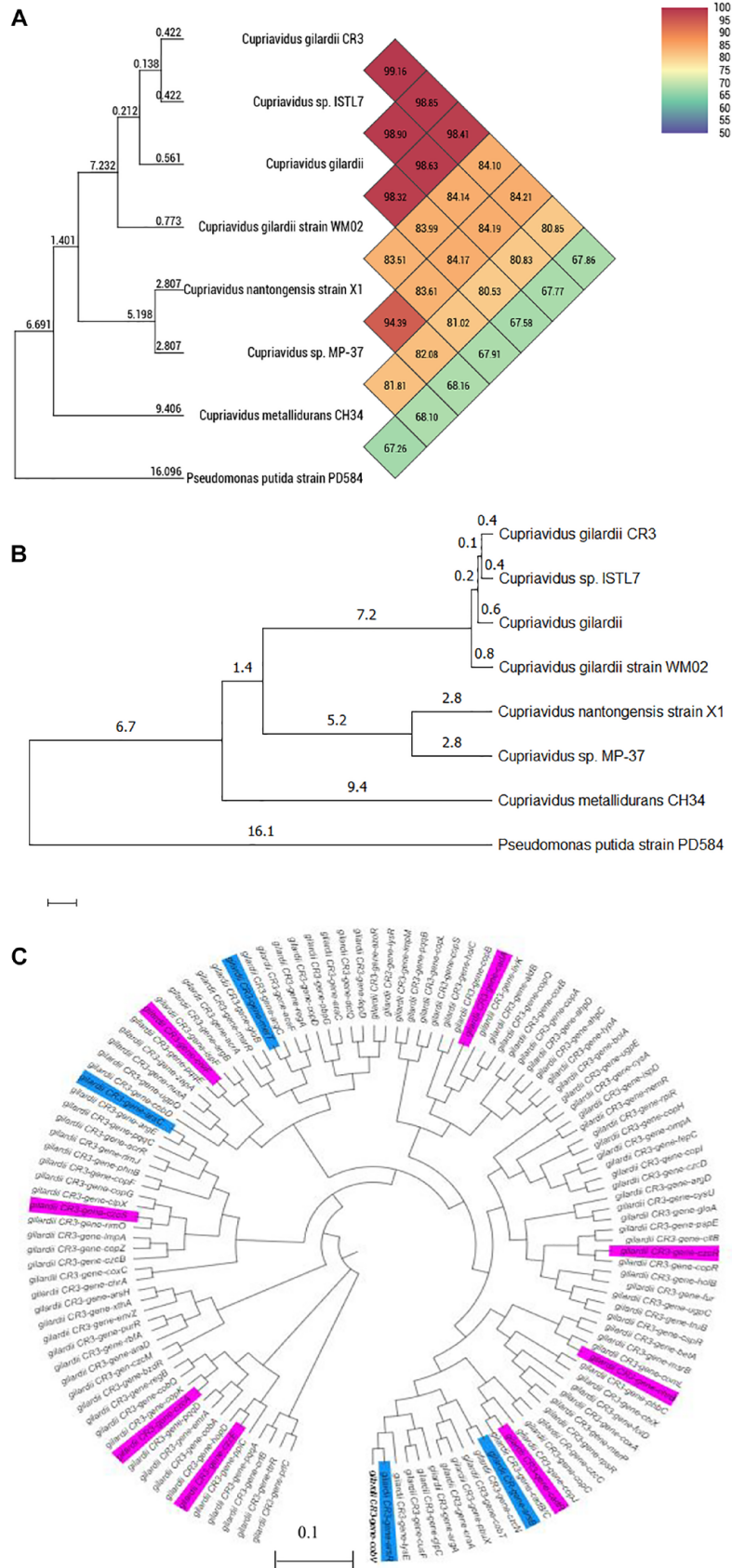


Figure 4. (A) Sequence similarity of OrthoANI and the other closely related strains, (B) phylogenetic tree and evolutionary relationships of taxa for *Cupriavidus gilardii* CR3, and (C) a circular phylogenetic analysis of the complete genomes of *C. gilardii* CR3. The phylogenetic tree shows heavy metal resistance gene locations.

Table 1. TSS, its promoter predictive score values, and distance from 5'-UTR region of the corresponding gene predicted and function.

S. NO.	GENE ID	GENE SYMBOL	NO. OF TSS IDENTIFIED	PREDICTIVE SCORE CUT-OFF VALUE OF 0.80	L. OF TSS FROM THE START CODON	GENE FUNCTION
1	66323649	<i>ChrA</i>	5	0.97,0.90,0.81,0.98,0.81	-45	Chromate efflux
2	66323648	<i>ChrB</i>	3	0.81,0.98,0.97	-550	Chromate transporter
3	60820444	<i>CzcA</i>	1	0.88	2994	Zinc efflux pump
4	60820445	<i>CzcB</i>	1	0.88	1415	Zinc efflux pump
5	50136911	<i>CzcD</i>	19	1.00,0.94,0.99,0.85,0.83,0.99,0.98,0.89,0.83,0.87,0.98,1.00,0.99,1.00,0.99,1.00,0.99,0.90,0.96	-323	Zinc transporter
6	72447367	<i>CadA</i>	19	0.92,0.95,0.97,0.83,0.97,0.87,0.94,0.99,0.94,0.99,0.93,0.93,1.00,0.87,0.83,0.96,0.97,0.98,0.90	303	Cadmium translocating
7	69569638	<i>CadC</i>	16	0.95,0.84,0.97,0.98,0.88,0.92,0.95,0.87,0.83,0.97,0.87,0.94,0.99,0.94,0.99,0.99	30	Sensing metalloregulator transcriptional repressor
8	66840924	<i>cadD</i>	38	0.88,0.91,0.99,0.95,0.91,0.91,0.95,0.98,0.99,0.97,0.87,0.86,0.99,1.00,0.96,0.96,0.82,1.00,0.91,0.99,1.00,1.00,0.97,1.00,0.89,0.92,0.93,0.81,0.99,0.85,0.90	379	Cadmium resistance and transporter
9	67467524	<i>CutC</i>	20	0.93,0.98,0.82,0.93,0.90,0.86,0.94,0.93,0.98,0.84,0.85,0.96,0.99,0.87,0.81,1.00,0.97,0.92,0.98,1.00,0.94,0.99,0.97,0.95,0.99,0.97	-38	Copper resistance
10	67451619	<i>CopC/D</i>	4	0.83,0.99,0.89,0.90	688	Copper resistance
11	75085266	<i>ZnuA</i>	12	0.99,0.98,0.85,0.86,0.93,0.95,0.92,0.81,0.86,0.89,1.00,0.98	99	Zinc transporter substrate binding
12	75088734	<i>ArsA</i>	10	0.95,0.90,0.84,0.90,0.95,0.84,0.95,0.81,0.82,0.83	423	Arsenate pump
13	68608930	<i>ArsB</i>	21	0.90,0.83,0.93,0.96,0.81,0.88,0.97,0.82,0.95,0.83,0.91,0.85,0.92,0.81,0.92,0.99,1.00,0.86,0.98,0.95,0.84	440	Arsenate efflux transporter
14	68608929	<i>ArsC</i>	17	0.85,1.00,0.92,0.90,0.98,0.83,1.00,0.94,0.99,0.81,0.90,0.92,0.89,0.87,0.96,0.95,0.95	607	Arsenate reduction
15	68610477	<i>ArsR</i>	24	1.00,0.89,0.92,0.99,0.85,0.99,0.98,0.93,0.92,0.99,0.99,1.00,0.99,0.97,0.92,0.96,0.82,0.93,0.94,0.82,0.94,0.83,0.93,0.97	537	Arsenate regulation

TSS, transcriptional start site; UTR, untranslated region.

The NNPP (Neural Network Promoter Prediction) tool prediction findings are considered reliable at 0.8 cut-off values for the prokaryotic organism.

used to identify genes. Similarly, SEED Subsystem category distribution approach technology was used to characterize functional genes from the *C. gilardii* CR3 genome and other related strains.¹⁹

Comparative genome analysis of *C. gilardii* CR3 by using OAT software

In this study, we used Orthologous Average Nucleotide Identity Software Tool (OAT) online to figure out the

Orthologous average nucleotide identity (OrthoANI) of 16S rRNA from closely related species acquired from EziBio Cloud (www.ezi.biocloud.net/sw/oat).^{22,23} Genomic sequences of *C. gilardii* CR3 CP010516.1 and related strains such as *Cupriavidus gilardii* WM02, *C. metallidurans* CH34, *C. nantongensis* X1, *Cupriavidus* ISTL7, and *Cupriavidus* MP-37 were investigated for comparative genomic analysis of HMR genes for heavy metal bioremediation. *P. putida* PD584 (CP115665.1) was exploited as an out-group for the comparisons.²²

Table 2. List of predicted motifs and the number and proportion of promoter-containing motifs.

S. NO.	SEARCHED CANDIDATES MOTIFS	NO. OF PROMOTER % FOR EACH MOTIF	E-VALUE ^A	MOTIFS WIDTH	NO. OF BINDING SITES
1	CMFs_1	8 (72.73%)	3.0e-010	40	8
2	CMFs_2	6 (54.55%)	3.8e-010	50	6
3	CMFs_3	7 (63.64%)	1.3e-009	50	7
4	CMFs_4	7 (63.64%)	1.6e-010	50	7
5	CMFs_5	6 (54.55%)	3.0e-009	50	6

^APossibility of discovery of a similarly well-conserved motif in random sequences.

Results

TSS and promoter determination

Identifying the transcription start point is critical for understanding gene expression. In a conceptual sense, where an RNA polymerase enzyme interacts before the initiation spot. With primer extensions, TSSs can be pinpointed to the exact nucleotide. This approach involves binding an oligonucleotide primer to mRNA to start initiations. To accurately predict promoter regions and interpret patterns of gene expression, as well as build and understand genetic regulatory networks, a correct prediction of promoter sites has become a necessary element.²⁴ Nowadays, bioinformatics tools can also be used to determine the exact sites of TSS. The NNPP version 2.20 databases were utilized to work out the TSS, which is often used in HMR. The promoter region located upstream of 1 kb of the TSS was characterized with the assumption that the promoter's functional gene elements could be discovered within the region. Table 1 summarizes and provides the TSS predicted values for each of the HMR gene's coding sequences. Consequently, the gene variety has several TSS values ranging from 1 to 38. Interestingly, the minimum and maximum of the TSS values identified correspond to genes *CzcA/CzcB* and *CadD*, respectively, as shown in Table 1. Similarity genes symbolized by *CzcD* and *CadA* had the same TSS values on the different strands with -323 and 303 far away from codons, respectively.

A start codon's location has a great deal to do with transcriptional initiation and gene regulation. There can be an enhancement or hindrance to gene regulation based on the location of the start codon. Accordingly, about 36.67% and 73.33% of TSS was found on the negative strands and the positive strands, respectively, as summarized in Table 1. The minimum and maximum TSS values were 38 and 550 for genes found on negative strands. These values were linked to the *CutC* and *ChrB* genes. In contrast, the minimum and maximum TSS values were 30 and 2994 for those located on the positive strand that corresponds to *CadC* and *CzcA*, respectively, as found in Table 1. The *CzcA* and *CzcB* genes showed 2994 and 1415 the highest location of the best TSS predicted by NNPP at the fourth and second rounds, respectively. The

CadC gene had the lowest location of the best TSS values of 30 values.

Prediction of common motif and TFs

There is evidence that a wide range of factors affect gene expression. One of the most significant things was the motif. Thus, the five common candidate motifs were predicted and searched by the MEME suite algorithm version 5.5.2 available <https://meme-suite.org/meme/>. For this study, data show that the CMFs_3 by e-values of 1.3e_009 having seven numbers of promoters with motive width of 50 was the best candidate motif as mentioned in Table 2 compared to 84 in the databases of MEME. Thus, CMFs_3 had the lowest e-values having 63.64% binding sites. Table 2 also showed that the highest and lowest numbers of promoter binding sites had been eight and six, respectively with 3.0e_010 and 3.8e_010/3.0e_009 e-values. As presented in Table 2, the CMFs_4 candidate motif shares with that of the best candidate motifs of the number of binding sites.

The prospective common motif with the lowest e values (1.3e-009) indicates a significantly different and functionally important motif, which was successfully imported into TOMTOM a motif comparison tool version 5.5.2 for further comparison, a database that is freely available for transcription element predictions that may be matching to known regulatory elements.^{25,26} TOMTOM includes LOGOSS, which stands for the matching of observed motifs with probable TFs. The database's TOMTOM output includes interactions with a parent TFs database for more information. This includes the matched motif's activation, repression, and dual regulatory functions. Furthermore, conformational information connected to the TF databases of monomers, dimers, tetramers, unpaired, and other factors is anticipated (see Tables 3 and 4). The data showed that CMFs_3 had the lowest e-value (1.3e-009) and was statistically significant with nine matching TFs from the databases obtained with matched e-value criteria of 14 or below, as tested and seen in the TOMTOM database. Figure 1 describes a forward and reverse strand of statistically significant strands.

Table 3. Lists of matching candidates from EXPREG transcription factor (TF).

S. NO.	CANDIDATE OF TF	STRAINS THAT SHOW MOTIF SEQUENCE BINDING	GC (%)	REGULATORY ELEMENTS				STATISTICAL SIGNIFICANCE
				ACTIVATION (%)	REPRESSION (%)	DUAL (%)	NOT SPECIFIED (%)	
1	<i>HrpX</i>	<i>Xanthomonas oryzae</i>	53.20	100	00	0	0	6.42e-02
2	<i>AmrZ</i>	<i>Pseudomonas aeruginosa</i> PAO1	60.21	16	31	0	51	4.11e-01
3	<i>RpoN</i>	<i>Vibrio cholerae</i>	47.75	10	0	0	0	9.60e-01
4	<i>HexR</i>	<i>Shewanella oneidensis</i>	21.72	30	69	0	0	1.69e + 00
5	<i>VqsM</i>	<i>P aeruginosa</i> PAO1	59.33	7	0	0	92	2.02e + 00
6	<i>PvdS</i>	<i>P aeruginosa</i> PAO1	48.33	100	0	0	0	2.23e + 00
7	<i>Lrp</i>	<i>Escherichia coli</i>	40.00	1	1	0	97	5.34e + 00
8	<i>IhfA</i>	<i>Pseudomonas putida</i>	10.00	100	0	0	0	5.82e + 00
9	<i>Fur</i>	<i>Helicobacter pylori</i>	14.52	13	78	0	8	7.33e + 00
10	<i>ArgR</i>	<i>P aeruginosa</i> PAO1	50.20	61	27	0	11	7.57e + 00
11	<i>Fur</i>	<i>Salmonella enterica</i>	28.33	26	73	0	0	8.08e + 00
12	<i>CcpA</i>	<i>Lactococcus lactis</i>	33.46	69	30	0	0	8.20e + 00
13	<i>PhhR</i>	<i>P putida</i>	44.67	90	10	0	0	8.27e + 00
14	<i>CcpA</i>	<i>Streptococcus pneumoniae</i>	32.21	68	31	0	0	9.28e + 00

Abbreviations: *AmrZ*, transcriptional activation and biofilm formation; *ArgR*, arginine biosynthesis; *CcpA*, catabolite control protein A; *Fur*, ferric uptake regulation protein; *HexR*, regulation glucose metabolism; *HrpX*, transcriptional regulator; *IhfA*, integration host factor subunit alpha; *Lrp*, leucine-responsive regulatory protein; *PhhR*, phenylalanine hydroxylase regulator; *PvdS*, encode putative sigma factor; *RpoN*, gene expression; *VqsM*, virulence and QS modulator.

Transcription factors regulate gene transcription, that is, their copying into RNA, on the way to protein making. A TF is a sequence-specific DNA-binding protein that controls the transcription of a set of genes, and thus regulates gene expression in a cell.²⁷ Thus, for this study, about nine TFs were identified from TOMTOM databases that handle the regulation of gene expression in the bacterium of HMR. *AmrZ*, *ArgR*, *CcpA*, *Fur*, *HexR*, *HrpX*, *IhfA*, *Lrp*, *PhhR*, *PvdS*, *RpoN*, and *VqsM* are TFs with diverse functions in gene regulation and expression. The primary functions of these TFs were activation and repression as seen in Table 3. *P aeruginosa* PAO1, *Streptococcus pneumoniae*, *P putida* KT2440, *Xanthomonas oryzae*, *Vibrio cholerae*, *Shewanella oneidensis*, *Salmonella enterica*, *Helicobacter pylori*, *Lactococcus lactis*, and *E coli* as revealed in Table 3. The sequences of all motif-binding sites were obtained from prokaryotic organisms.

Table 3 shows that *HrpX*, *AmrZ*, *RpoN*, *HexR*, *PvdS*, *Lrp*, *CcpA*, *ArgR*, *PhhR*, and *IhfA* were predicted TFs that activate transcription by 100%. The rest of the TFs that act as activators, repressors, and others have yet to be identified. As shown in Table 3 above, none of the identified TFs was used in dual. It was found that several TFs exist, but their functions have not

yet been revealed, as in the case of *Lrp* and *VqsM*. Additional research using various tools and wet labs may be needed for further investigation. Only *Fur* and *HexR* TFs had the highest functions of repression on 78% and 69% transcriptional gene expressions, in Table 3. This shows that these TFs take part in greater transcription repression than activation. In general, the data show that most identified TFs are used in activation rather than repression. From an environmental perspective, the *PhhR* TF has been exploited in a variety of biotechnological applications. It was employed in different technological areas in addition to agriculture, biocatalysts, bioremediation potential applications, and bioplastic production.¹⁰

Table 4 also shows the TF conformational mode of application of the identified TFs as monomers, dimers, tetramers. For instance, about 35.71% of the identified TFs' conformational mode were not known while 7.14% and 21.43% were identified as monomer and dimer TF conformation mode applications. As shown in Table 4, no TFs were associated with the tetramer TF conformational mode.

Figure 2 revealed the open reading frame (ORF) in which the predicted genes were located. It is usually located at the 5'-UTR region of the upstream of the transcripts. It was known

Table 4. Lists of matching candidates from EXPREG transcription confirmation factor (TCF).

S. NO.	CANDIDATE OF TF	STRAINS THAT SHOW MOTIF SEQUENCE BINDING	GC (%)	TF CONFIRMATION MODE				NOT SPECIFIED (%)	STATISTICAL SIGNIFICANCE
				MONOMER (%)	DIMER (%)	TETRAMER (%)	OTHER (%)		
1	<i>HrpX</i>	<i>Xanthomonas oryzae</i>	53.20	80	20	0	0	0	6.42e-02
2	<i>AmrZ</i>	<i>Pseudomonas aeruginosa</i> PAO1	60.21	0	0	0	0	99	4.11e-01
3	<i>RpoN</i>	<i>Vibrio cholerae</i>	47.75	100	0	0	0	0	9.60e-01
4	<i>HexR</i>	<i>Shewanella oneidensis</i>	21.72	0	0	0	0	100	1.69e + 00
5	<i>VqsM</i>	<i>P aeruginosa</i> PAO1	59.33	0	0	0	0	100	2.02e + 00
6	<i>PvdS</i>	<i>P aeruginosa</i> PAO1	48.33	0	0	0	0	100	2.23e + 00
7	<i>Lrp</i>	<i>Escherichia coli</i>	40.00	0	96	0	0	3	5.34e + 00
8	<i>IhfA</i>	<i>Pseudomonas putida</i>	10.00	0	50	0	0	50	5.82e + 00
9	<i>Fur</i>	<i>Helicobacter pylori</i>	14.52	0	82	0	0	17	7.33e + 00
10	<i>ArgR</i>	<i>P aeruginosa</i> PAO1	50.20	0	16	0	0	83	7.57e + 00
11	<i>Fur</i>	<i>Salmonella enterica</i>	28.33	0	100	0	0	0	8.08e + 00
12	<i>CcpA</i>	<i>Lactococcus lactis</i>	33.46	0	0	0	0	100	8.20e + 00
13	<i>PhhR</i>	<i>P putida</i>	44.67	0	100	0	0	0	8.27e + 00
14	<i>CcpA</i>	<i>Streptococcus pneumoniae</i>	32.21	0	100	0	0	0	9.28e + 00

Abbreviations: *AmrZ*, Transcriptional activation and biofilm formation; *ArgR*, arginine biosynthesis; *CcpA*, Catabolite control protein A; *Fur*, Ferric uptake regulation protein; *HexR*, regulation glucose metabolism; *HrpX*, transcriptional regulator; *IhfA*, Integration host factor subunit alpha; *Lrp*, Leucine-responsive regulatory protein; *PhhR*, phenylalanine hydroxylase regulator and *VqsM* is virulence and QS modulator; *PvdS*, encode putative sigma factor; *RpoN*, gene expression.

that six ORFs are found in molecular biology, three on the positive, and three on the negative strands. This refers to reading frames in the RNA code used by ribosomes in protein synthesis. As a result, it may serve as a repository for genes with functional elements that could be targeted by heavy metal bioremediation in current studies. In Figure 2, the red color indicates that genes with similar sequences are grouped together in the same number and color. Genes whose relative position is conserved in other species are functionally coupled and share gray background boxes. The focus gene always points to the right, even if it is located on the minus strand. For instance, the *czcA* and *merR* genes were pinpointed out of the reading frame in red color in the same direction. The same reading frame varied in size, sequence, and location within the genome.

Genome re-annotation and genome mapping of *C. gilardii* CR3 genome

Circular genome mapping for *C. gilardii* CR3 genomic DNA was performed using a Prokka online server that can be accessed at <https://proksee.ca/>, which replaced CGView. The GC content, ORF, GC skew (+/-), rRNA, starting codon, stop codon,

regulatory elements and CDS were predicted on the genome mapping (see Figure 3A to C). The sequence of *C. gilardii* CR3 was obtained via the NCBI database and re-annotated utilizing subsystems category distribution technology. An analysis was performed on the circular genome for *C. gilardii* CR3, which showed the position of GC content, sequence length and/or contigs, stop, start, ORF, GC skew⁺ and GC skew⁻ (see Figure 3A to C). Several genes have been thought to be associated with HMR mechanisms in environments with contamination, following this investigation. Genes such as *chrA*, *chrB*, *chrR*, *arsA*, *arsB*, *copC/D*, *czcA*, *czcB*, *czcD*, *czcABC*, *cadiC*, and *cadiC* play a significant role in the polluted environment to clean or decrease pollutant levels after it had been expressed (see Figure 3A to C). Similar investigation of circular genome analysis for the *trehalase* gene from the complete genome of *Shigella* sp. PAMC28760 from Antarctica were predicted with CDS, GC contents, rRNA, tRNA, GC skew⁺ and GC skew⁻ and ORFs have been predicted.⁸ The RAST server identified genes with their functions, locations, and levels of protein expression for *C. gilardii* CR3 (CP010516.1) and other related associated strains were shown in Figure 3A. For this study, the number of coding sequences (3399), GC contents (67.7), contigs (1), shortest contig size (3539530), media sequence sizes

(3539530), mean sequence sizes (3539530.0), longest contig sizes (3539530), the sequence length in base pairs (N_{50} values; not predicted) and the smallest number of contigs (L_{50}) (1) whose length sum makes up half of genome size were predicted.

The prediction revealed that the genome contains multiple genes involved in co-factor production, chromosomal replication initiator protein DnaA, vitamin biosynthesis, protein metabolism, nitrogen metabolism, RNA metabolism, nucleosides and nucleotide biosynthesis, stress response and aromatic compound metabolism, copper-sensing two-component system response regulator, mercuric resistance operon regulatory protein, lead, cadmium, zinc and mercury transporting ATPase, and chromate transport protein in Figure 3A.

Comparative genome analysis by using OAT software tools

The 16S rRNA sequences were obtained from WGS *C. gilardii* CR3 (CP010516.1) and related other strains such as *C. metallidurans* CH34 (NC_007973.1), *C. nantongensis* X1 (CP014844.1), *Cupriavidus* sp. ISTL7 (CP066227.10156), *Cupriavidus* sp. MP-37 (NZ_CP085344). The phylogenetic tree was performed using the OAT offline server. The sequences were aligned with ClustalW. *P. putida* PD584 (CP115665.1) was used as an out-group in Figure 4A and B.

The neighbor-joining method was used to infer evolutionary history.²⁸ The best tree was shown with branch lengths. Based on 1000 bootstrap replications through confidence points, evolutionary distances were calculated using neighbor-joining. A bar, 0.01 substitutions per nucleotide location.²⁹ The evolutionary distance was calculated by p-distance techniques.³⁰ The investigation connected eight nucleotide sequences as one nucleotide sequence was used as an out-group. All locations having gaps and missing data were eliminated. Finally, the evolutionary relationship investigation was performed with OAT offline server.²² *P. putida* PD584 (CP115665.1) was selected and used as an out-group in the evaluation; other related sequences were obtained from the EzTaxon-e server and annotated with the RAST server before tree construction (b) UPGMA dendrogram Heatmap for OrthoANI.²² The OrthoANI values of $\geq 96\%$ show grouped genomes originating from strains of the same species together (see Figure 4A and B).

A circular phylogenetic analysis of the complete genome of *C. gilardii* CR3 (CP010516.1) was studied. The phylogenetic tree shows heavy metal resistance genes location in the complete genome of the organism in Figure 4C. The data also show that *C. gilardii* CR3 is 99.41% related to *Cupriavidus* sp. ISTL7. Figure 4A and B shows *C. gilardii* CR3 (CP010516.1) has a close relationship to *Cupriavidus* sp. ISTL7 (CP066227.10156) and *C. gilardii* WM02 (CP104386.1) and less related to that of *C. nantongensis* X1 (CP014844.1) and *C. metallidurans* CH34 (NC_007973.1). The 16S rRNA gene sequence results of

C. gilardii CR3 (CP010516.1) were highly evolutionary related to *Cupriavidus* sp. ISTL7 (CP066227.10156), *C. gilardii*, *C. gilardii* WM02 (CP104386.1), and *Cupriavidus* sp. MP-37 (NZ_CP085344.1) as seen in Table 5. The data also revealed that it was less evolutionary related to that of *C. metallidurans* CH34 (NC_007973.1). Thus, the evolutionary relationship between the organisms was 99.41%, 98.85%, 98.41%, and 94.39%, respectively (see Table 5). The OrthoANI again revealed a weak correlation among *C. metallidurans* CH34 (NC_007973.1) compared with *C. gilardii* CR3 (see Figure 4A and B). As shown in Figure 4A and B, the sequence similarity of OrthoANI between *C. gilardii* CR3 (CP010516.1) and *Cupriavidus* sp. ISTL7 (CP066227.10156), *C. gilardii* CR3 (CP010516.1), *C. gilardii*, *C. gilardii* CR3 (CP010516.1) and *C. gilardii* WM02 (CP104386.1), *C. gilardii* CR3 (CP010516.1) and *C. nantongensis* X1 (CP014844.1), *C. gilardii* CR3 (CP010516.1) and *Cupriavidus* sp. MP-37 (NZ_CP085344.1), *C. gilardii* CR3 (CP010516.1) and *C. metallidurans* CH34 (NC_007973.1) were 99.41%, 98.85%, 98.41%, 84.10%, 84.21%, and 80.85% (see Figure 4A and B). The larger sequence similarities among the current genome annotated sequences of our strains were calculated between *C. gilardii* CR3 (CP010516.1) and *Cupriavidus* sp. ISTL7 (CP066227.10156), OrthoANI is a genomic similarity measuring technique. It was utilized to evaluate the same species of organism with a demarcation cut-off value of 95% to 96%, and massive comparison tests showed that both algorithms produced identical reciprocal similarities (see Figure 4A and B) which confirms our phylogenetic tree. Accordingly, the OrthoANI obtained between *C. metallidurans* CH34 (NC_007973.1) and *P. putida* PD584 (CP115665.1) as well as other related genera in the present investigation were lower of the demarcation cut-off values.

Discussions

The objective of this work was regulatory elements analysis and comparative genome analysis of HMR bacteria extracted from *C. gilardii* CR3 as a means of bioremediation. Sequence-specific DNA-binding TFs are usually revealed as “main regulators” because they bind to DNA and either activate or suppress gene expression. The identification of a gene’s TSS precedes the pointing out of the upstream promoter region, making gene expression analysis easier. Most of the sequences in this study showed multiple TSSs, which accounted for about 86.67% in Table 1. As a result, the gene sequence in consideration may have alternate transcription possibilities. However, TSS with a higher threshold value were evaluated for better prediction that related to earlier studies.³¹ Genes that have multiple TSS enhance transcription initiation and are more likely to respond to environmental changes, which agreed with the findings of Dai et al.³²

Common motif identification was a core element of TF identification processes. Motifs recruit TFs, or regulatory elements to properly target genes in gene expression processes.³³ The MEME tools were utilized to identify a conserved

Table 5. Features of re-annotated genomic DNA for *Cupriavidus gilardii* CR3 genome, its allied reference strains, and out-group.

S. NO.	PART OF GENOMIC INFORMATION	AS UPLOADED AND AFTER SPLITTING INTO SCAFFOLDS						
		<i>CUPRIAVIDUS GILARDII</i> CR3	<i>CUPRIAVIDUS METALLIDURANS</i> CH34	<i>CUPRIAVIDUS NANTONGENSIS</i> X1	<i>CUPRIAVIDUS</i> SP. ISTL7	<i>CUPRIAVIDUS</i> SP. MP-37	<i>CUPRIAVIDUS GILARDII</i> WM02	<i>PSEUDOMONAS PUTIDA</i> PD584
1	Accession No	CP010516.1	NC_007973.1	CP014844.1	CP066227.1	NZ_CP085344.1	CP104386.1	NZ_CP115665.1
2	Taxonomy ID	1267562	266264	1796606	2771360	2884455	82541	303
3	Sequence size	3539530	6913352	4619400	3314946	3615976	3511514	5913415
4	No. of coding sequences	3399	6603	4525	3221	3354	3323	5375
5	Number of contigs	1	4	3	207	1	1	1
6	GC content (%)	67.4	63.5	66.2	67.7	67.5	67.4	61.6
7	Shortest contig size	3539530	171459	1150655	19	3615976	3511514	5913415
8	Media sequence size	3539530	2580084	1728946	564	3615976	3411514	5913415
9	Mean sequence size	3539530	172833.0	1539800.0	16014.2	3615976.0	3511514.0	5913415.0
10	Longest contig size	3539530	3928089	1739799	169580	3615976	3511514	5913415
11	N ₅₀ values	Not predicted	3928089	17289946	71905	Not predicted	Not predicted	Not predicted
12	L ₅₀ values	1	1	1	16	1	1	1
13	Number of RNA	61	74	63	60	67	61	96
14	No. of subsystems	285	366	316	275	302	289	364

The data were predicted and generated by RAST (Rapid Annotation utilizing Subsystem Technology) and SEED (The database and infrastructure for comparative genomics) technologies.

candidate motif where these 14 TFs bind and motif sites were discovered (see Figure 5). In this study, common candidate motifs with different information content were identified. The densely populated motifs with their locations were found upstream between -100 and -600 (see Figure 5). These motifs might take part in the regulation of recognized promoters with contributors in gene expression. In this investigation, several binding sites were discovered in the promoter region of candidate motifs. These locations might enhance binding interactions and produce different regulatory effects.^{34,35}

The data also show that the gene *arsB* at locus_tag=CR3_1877 with CDS located at complement (2282267. . . 2283709) was a membrane product protein extracted from *C gilardii* CR3. In addition, the *arsC* and *arsH* genes at locus_tag=CR3_2746 and locus_tag=CR3_1262 were used as arsenate reductase and arsenical resistance functions, respectively. Moreover, the *arsR* gene with CDS complement (1487862. . . 1488218) is located in the genome of *C gilardii* CR3, a family transcriptional regulator protein widely used for arsenate bioremediation capacities. Similar studies were

performed in the complete genome of *C metallidurans* CH34 as a model bacterium of HMR. Resisting heavy metals by tri-component efflux systems were identified in pMOL30, with *czcCBA* present with resistance to Cd²⁺, Zn²⁺, and Co²⁺.³⁶ Environmental components such as soil and water contaminated areas with heavy metals are potential sources of HMR bacteria. These microbes could reduce, oxidize bio-accumulate, and efflux heavy metals through cell membranes and are widely used in bioremediation processes.³⁷ According to the comparative analysis, the genome of *Cupriavidus* sp. ISTL37 is 99.41% identical to the genomes of *C gilardii* CR3 as shown in Figure 5. A similar finding was observed on the *C metallidurans* MSR33;³⁸ however, it was less related to *C gilardii* CR3 as seen in Figure 5.

The *MerR* families include *merP*, *MerT* protein that is widely distributed in the bacterium of *C gilardii* CR3 and mercury resistance has also been investigated in this study (see Figure 3A). The finding is similar to the studied identified genes from Eubacteria.³⁹ The *C gilardii* CR3 regulates heavy metals through TFs controlled by the *MerR*, *CopR*, *CadR*, and *ChrR* protein

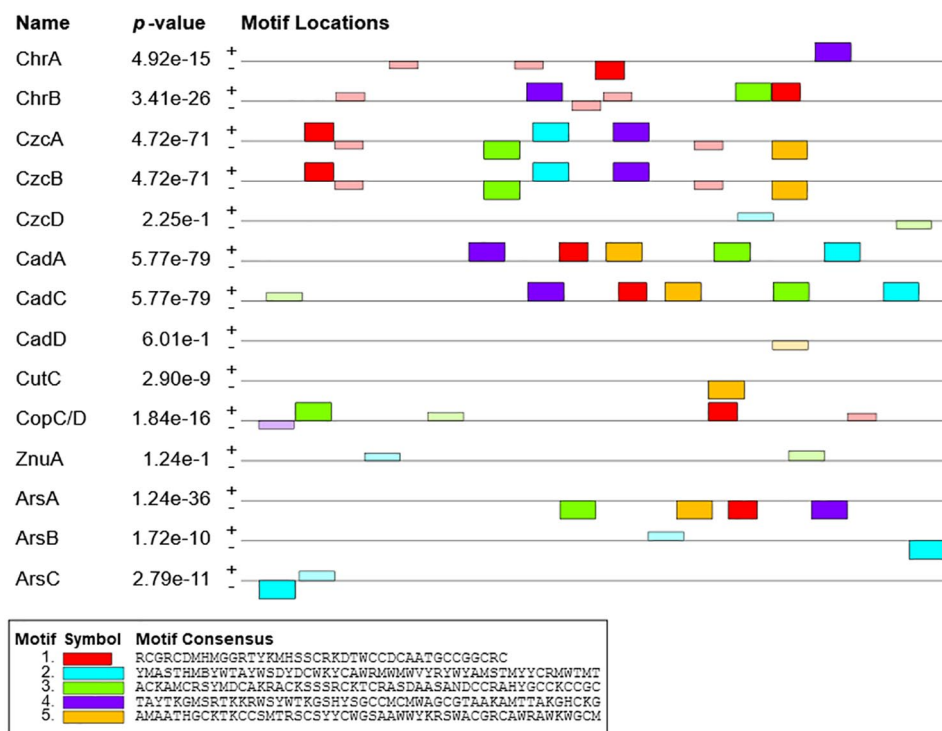


Figure 5. Block diagrams illustrate the relative location of potential motif + scanned sites in the promoter region relative to TSSs (transcription start sites). The bottom of the graph indicates the locations of nucleotides in the promoter region of heavy metal resistance genes. The locations range from the beginning of the TSS (+1) to upward of 1 kb (–1 kb) of the MEME suite outcome.

families. We also studied the cadmium, copper, arsenate, zinc, and chromium resistance genes of *CopB*, *CopC*, *CopCD*, *CopR*, *CadA*, *CadB*, *CadR*, *ArsA*, *ArsB*, *ArsC*, *ArsH*, *ArsR*, *CzcA*, *CzcB*, *ChrA*, *ChrB* (see Figure 4C) resistant systems and demonstrated that the combined protein sequence analysis and analysis of DNA regulatory elements was possible to differentiate the genes involved in the metal resistant, in particular, mercury, copper, cadmium, zinc, and arsenate.^{40–43} Several enzymes contribute to the efficient functions of the genes, including a multicopper oxidase, a cadmium transporter, a zinc transporter, and a copper/zinc/cadmium efflux system.⁴⁴

The RAST server identified gene calls based on their function, location, and degree of protein expression. In addition, sequence size, number of coding sequences, number of contigs, GC content percentage, shortest contig sequence, mean sequence size, longest contig size, No₅₀ values, L₅₀ values, number of RNA, and a number of subsystems were predicted by the RAST server. The GC content percentage of *C. gilardii* CR3 is similar to that of *C. gilardii* WM02, which is 67.4%. However, the highest GC% is documented for *Cupriavidus* sp. ISTL7 (67.7%) while the lowest GC% is observed for *P. putida* PD584 (61.6%). Similar findings were found in the study on *Bacillus sereus*.⁴⁵ Following *C. gilardii* CR3 genomic DNA re-annotation with the RAST online server, subsystem category distribution and subsystem feature counts were anticipated. These subsystems are supposed to support cellular activity and

metabolism. Strengthening our suggestion, it was projected that these subsystems would be used to provide broad metabolic information, increase annotation quality, and provide a framework for determining the statistical qualities required to properly exploit these tendencies. Our results anticipate carbohydrates and other chemical molecules. The OrthoANI, a tool that predicts sequence similarity between *C. gilardii* CR3 and related strains, is an example of an OAT tool offline server.²² When tested against OrthoANI, *C. gilardii* CR3 (CP010516.1) displayed a genome sequence similarity of 99.16%. The *C. gilardii* CR3 genomic DNA was found to be most similar to *Cupriavidus* ISTL7 with 98.85% OrthoANI (see Figure 4A), using the OAT tool.²² Followed by *C. gilardii* MW5 (98.80% OrthoANI), *C. gilardii* strain WM02 (98.32% OrthoANI), *Cupriavidus* sp. MP37 (94.39% OrthoANI), *C. nantongensis* X1 (CP014844.1; 83.51% OrthoANI) and *C. metallidurans* CH4.

Conclusions

In this study, 15 genes from *C. gilardii* CR3 genomes were examined for HMR applications. To determine the functional conformation of the predicted genes, the promoter region, transcriptional regulatory elements, phylogenetic tree analysis and comparative genomic by OAT analysis, and RAST and SEED tools were utilized. Five common candidate motifs for binding sites related to regulatory elements were investigated.

UniProtKB data showed additional research for a better understanding of the function of the TFs. The overall results revealed that identification of the common candidate motifs and their regulatory elements such as TFs, comparative genome analysis by OAT, and genome re-annotation by RAST with SEED technologies were employed for the investigation. The data show that 14 TFs were identified. In both positive and negative regulations, these TFs were widely used to regulate enzymatic biosynthesis and metabolism applications. In addition, comparative genome analysis, annotation, bioinformatics, and omics tools were expected to contribute significantly to our understanding of gene expression in prokaryotic. Therefore, more studies could be done to identify the screened TFs and their common binding sites in gene expression mechanisms in microorganisms with the help of a wet lab.

Acknowledgements

The authors are grateful to Adama Science and Technology University for the opportunities.

Author Contributions

DD, TD, and SM conceptualized and designed the study, analyzed, and interpreted the data. TD and SM analyzed, revised, and edited the article. All authors read and approved the final article to the published.

Ethical Approval and Consent of Participation

Not applicable.

Consent of Publication

Not applicable.

Availability of Data and Materials

The datasets used in this study were extracted from the NCBI database. Where necessary, it can be obtained from the corresponding author.

ORCID iD

Duguma Dibbisa  <https://orcid.org/0000-0002-2534-6337>

REFERENCES

- Girma B. The effects of environmental pollution of eastern industrial park on human health and comfort of Dukem Town, Ethiopia. *J Ind Pollut Control*. 2021;37:1-14.
- Luo Y. Environmental problems in the mining of metal minerals. *IOP Conf Ser Earth Environ Sci*. 2019;384:012195. doi:10.1088/1755-1315/384/1/012195
- Rothschild LJ, Mancinelli RL. Life in extreme environments. *Nature*. 2001;409:1092-1101. doi:10.1038/npg.els.0000391
- Tahir U, Yasmin A. Role of bacterial extracellular polymeric substances (EPS) in uptake and accumulation of co-contaminants. *Int J Environ Sci Technol*. 2019;16:8081-8092. doi:10.1007/s13762-019-02360-0
- Vélez JMB, Martínez JG, Ospina JT, Agudelo SO. Bioremediation potential of *Pseudomonas* genus isolates from residual water, capable of tolerating lead through mechanisms of exopolysaccharide production and biosorption. *Biotechnol Rep (Amst)*. 2021;32:e00685. doi:10.1016/j.btre.2021.e00685
- Aransiola EF, Ige OA, Ehinmitola EO, Layokun SK. Heavy metals bioremediation potential of *Klebsiella* species isolated from diesel polluted soil. *Afr J Biotechnol*. 2017;16:1098-1105. doi:10.5897/ajb2016.15823
- Lu CW, Ho HC, Yao CL, Tseng TY, Kao CM, Chen SC. Bioremediation potential of cadmium by recombinant *Escherichia coli* surface expressing metallothionein MTT5 from *Tetrahymena thermophila*. *Chemosphere*. 2023;310:136850. doi:10.1016/j.chemosphere.2022.136850
- Shrestha P, Karmacharya J, Han SR, Park H, Oh TJ. In silico analysis and a comparative genomics approach to predict pathogenic trehalase genes in the complete genome of Antarctic *Shigella* sp. *Virulence*. 2022;13:1502-1514. doi:10.1080/21505594.2022.2117679
- Sharma V. Bioinformatics and its applications in environmental science and health and its applications in other disciplines. *Sambodhi J*. 2020;4:88-93.
- Nelson KE, Weinel C, Paulsen IT, et al. Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol*. 2002;4:799-808. doi:10.1046/j.1462-2920.2002.00366.x
- Sivashankari S, Shanmughavel P. Comparative genomics—a perspective. *Bioinformatic*. 2007;1:376-378. doi:10.6026/97320630001376
- Koo HJ, Ahn S, Chung HY, et al. Comparative genomic analysis reveals genetic features related to the virulence of *Bacillus cereus* FORC-013. *Gut Pathog*. 2017;9:1-8. doi:10.1186/s13099-017-0175-z
- Benmalek Y, Fardeau ML. Isolation and characterization of metal-resistant bacterial strain from wastewater and evaluation of its capacity in metal-ions removal using living and dry bacterial cells. *Int J Environ Sci Technol*. 2016;13:2153-2162. doi:10.1007/s13762-016-1048-6
- Pham VHT, Kim J, Chang S, Chung W. Bacterial biosorbents, an efficient heavy metals green clean-up strategy: prospects, challenges, and opportunities. *Microorganisms*. 2022;10:1-16. doi:10.3390/microorganisms10030610
- Reese MG. Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput Chem*. 2001;26:51-56. doi:10.1016/S0097-8485(01)00099-7
- Dinka H, Milkesa A. Unfolding SARS-CoV-2 viral genome to understand its gene expression regulation. *Infect Genet Evol*. 2020;84:104386. doi:10.1016/j.meegid.2020.104386
- Grant JR, Stothard P. The CGView server: a comparative genomics tool for circular genomes. *Nucleic Acids Res*. 2008;36:181-184. doi:10.1093/nar/gkn179
- Aziz RK, Bartels D, Best A, et al. The RAST server: rapid annotations using subsystems technology. *BMC Genomics*. 2008;9:1-15. doi:10.1186/1471-2164-9-75
- Overbeek R, Begley T, Butler RM, et al. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res*. 2005;33:5691-5702. doi:10.1093/nar/gki866
- Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*. 2004;5:1-19. doi:10.1186/1471-2105-5-113
- Kumar S, Tamura K, Nei M. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief Bioinform*. 2004;5:150-163. doi:10.1093/bib/5.2.150
- Lee I, Ouk Kim Y, Park SC, Chun J. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. *Int J Syst Evol Microbiol*. 2016;66:1100-1103. doi:10.1099/ijsem.0.000760
- Yoon SH, Ha SM, Kwon S, et al. Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. *Int J Syst Evol Microbiol*. 2017;67:1613-1617. doi:10.1099/ijsem.0.001755
- Lai HY, Zhang ZY, Su ZD, et al. IProEP: a computational predictor for predicting promoter. *Mol Ther Nucleic Acids*. 2019;17:337-346. doi:10.1016/j.omtn.2019.05.028
- Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 1994;2:28-36
- Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. *Nucleic Acids Res*. 2015;43:W39-W49. doi:10.1093/nar/gkv416
- Kim GB, Gao Y, Palsson BO, Lee SY. DeepTFactor: a deep learning-based tool for the prediction of transcription factors. *Proc Natl Acad Sci USA*. 2021;118:1-5. doi:10.1073/pnas.2021171118
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406-425. doi:10.1093/oxfordjournals.molbev.a040454
- Felsenstein J. Phylogenies and the comparative method. *Am Nat*. 1985;125:1-15. doi:10.1086/284325
- Tamura K, Stecher G, Kumar S. MEGA11: molecular evolutionary genetics analysis version 11. *Mol Biol Evol*. 2021;38:3022-3027. doi:10.1093/molbev/msab120
- Dibbisa D, Wagari G. In silico study of mercury resistance genes extracted from *Pseudomonas* spp. involved in bioremediation: understanding the promoter regions and regulatory elements. *Int J Genomics*. 2022;2022:6185615. doi:10.1155/2022/6185615
- Dai Z, Xiong Y, Dai X. DNA signals at isoform promoters. *Sci Rep*. 2016;6:1-8. doi:10.1038/srep28977
- Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. *Molecular Biology of the Cell*. W.W. Norton & Company; 2002. doi:10.1201/9780203833445

34. Aman Beshir J, Kebede M. In silico analysis of promoter regions and regulatory elements (motifs and CpG islands) of the genes encoding for alcohol production in *Saccharomyces cerevisiae* S288C and *Schizosaccharomyces pombe* 972h. *J Genet Eng Biotechnol.* 2021;19:8. doi:10.1186/s43141-020-00097-9
35. Bantihun G, Kebede M. In silico analysis of promoter region and regulatory elements of mitogenome co-expressed *trn* gene clusters encoding for bio-pesticide in entomopathogenic fungus, *Metarhizium anisopliae*: strain ME1. *J Genet Eng Biotechnol.* 2021;19:94. doi:10.1186/s43141-021-00191-6
36. Mergeay M, Van Houdt R. *Cupriavidus metallidurans* CH34, a historical perspective on its discovery, characterization and metal resistance. *FEMS Microbiol Ecol.* 2021;97:1-7. doi:10.1093/femsec/fiaa247
37. Razzak SA, Faruque MO, Alsheikh Z, et al. A comprehensive review on conventional and biological-driven heavy metals removal from industrial wastewater. *Environ Adv.* 2022;7:100168. doi:10.1016/j.envadv.2022.100168
38. Rojas LA, Yáñez C, González M, Lobos S, Smalla K, Seeger M. Characterization of the metabolically modified heavy metal-resistant *Cupriavidus metallidurans* strain MSR33 generated for mercury bioremediation. *PLoS ONE.* 2011;6:e17555. doi:10.1371/journal.pone.0017555
39. Permina EA, Kazakov AE, Kalinina OV., Gelfand MS. Comparative genomics of regulation of heavy metal resistance in Eubacteria. *BMC Microbiol.* 2006;6:49. doi:10.1186/1471-2180-6-49
40. Haque F, Jabeen I, Keya CA, Shuvo SR. Whole-genome sequencing and comparative analysis of heavy metals tolerant *Bacillus anthracis* FHq strain isolated from tannery effluents in Bangladesh. *AIMS Microbiol.* 2022;8:227-239. doi:10.3934/MICROBIOL.2022018
41. Yetunde Mutiat FB, Gbolahan B, Olu O. A comparative study of the wild and mutated heavy metal resistant *Klebsiella variicola* generated for cadmium bioremediation. *Bioremediat J.* 2018;22:1-15. doi:10.1080/10889868.2018.1445695
42. Jayanti S, Sarkar M, Mandal P, Pal A. Comparative study of heavy metal uptake and analysis of plant growth promotion potential of multiple heavy metal-resistant bacteria isolated from arable land. *Curr Microbiol.* 2022;79:7. doi:10.1007/s00284-021-02704-5
43. Saha J, Dey S, Pal A. Whole genome sequencing and comparative genomic analyses of *Pseudomonas aeruginosa* strain isolated from arable soil reveal novel insights into heavy metal resistance and codon biology. *Curr Genet.* 2022;68:481-503. doi:10.1007/s00294-022-01245-z
44. Learman DR, Ahmad Z, Brookshier A, et al. Comparative genomics of 16 *Microbacterium* spp. that tolerate multiple heavy metals and antibiotics. *PeerJ.* 2019;6:e6258. doi:10.7717/peerj.6258
45. Zeng Q, Xie J, Li Y, Gao T, Xu C, Wang Q. Comparative genomic and functional analyses of four sequenced *Bacillus cereus* genomes reveal conservation of genes relevant to plant-growth-promoting traits. *Sci Rep.* 2018;8:17009. doi:10.1038/s41598-018-35300-y