# Fast and accurate detection of metal resistance genes using MetHMMDb

**Karol Ciuchcinski[1,*], Mikolaj Dziurzynski[2]**

[1] Department of Environmental Microbiology and Biotechnology, Institute of Microbiology, Faculty of Biology, University of Warsaw

[2] Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw

* Corresponding author:

Karol Ciuchcinski; phone: +48 225542205; e-mail: k.ciuchcinski@uw.edu.pl

**ABSTRACT**

Heavy metal pollution, driven by industrialization, is a growing environmental challenge with severe ecological impacts. One of key factors in fighting this issue is microbial resistance to heavy metals, and its potential for use in bioremediation and environmental monitoring. Unfortunately, existing detection methods often fail to identify distant homologs or to provide insight into the function of microbial metal resistance genes (MMRGs). To overcome these limitations, we developed MetHMMDB - a novel database containing 254 profile Hidden Markov Models (HMMs) representing 121 microbial MMRGs and functions. Unlike traditional sequence similarity-based tools, MetHMMDB leverages HMMs to enhance detection sensitivity and functional specificity across diverse microbial communities. MetHMMDB was created through iterative database searches, combined with sequence

clustering and structural prediction, combining both recent advances in bioinformatics and manal curation of data. In its core, the database prioritises functional annotation, ensuring accurate detection and classification of genes responsible for metal resistance. Our results indicate that MetHMMDB outperforms sequence-based approaches, identifying more than twice as many MMRGs in metagenomic datasets, including extreme environments. Additionally, application to agricultural soil revealed distinct resistance profiles linked to soil quality. By improving the identification and functional characterization of metal resistance genes, MetHMMDB facilitates interdisciplinary research in microbial ecology, environmental monitoring, and bioremediation. This robust and versatile resource represents a significant advancement in understanding microbial adaptation to heavy metal contamination and supports the development of effective environmental management strategies. The database is available at https://github.com/Haelmorn/MetHMMDB.

Heavy metal pollution, driven by industrialization and human activities, is a growing environmental challenge with severe consequences for ecosystems and biodiversity [1–3]. As heavy metal contamination continues to escalate, there is a need for effective solutions for monitoring and mitigating these pollutants. One promising avenue lies in the study of microbial resistance mechanisms, which have the potential to be harnessed for bioremediation and environmental monitoring [4]. Microbial metal resistance is mediated by a diverse set of genes and pathways that enable microorganisms to tolerate or detoxify metal ions in contaminated environments. Identifying these microbial metal resistance genes (MMRGs) is crucial for understanding microbial adaptation, ecological impacts, and for developing biotechnological applications. Current detection methods, however, largely rely on sequence similarity searches against sequence databases like BacMet [5–7], which, while being a valuable source of manually curated data, has not been updated since 2018 [8]. Additionally, sequence-based methods often fail to detect distant homologs or provide functional context, limiting their applicability for analyzing diverse and complex microbial communities. Here we present MetHMMDB, a comprehensive, manually annotated database of 254 profile Hidden Markov Models (HMMs), uniquely designed to detect and characterize bacterial metal resistance genes with enhanced sensitivity and functional specificity in diverse datasets.

The MetHMMDB database was constructed through a systematic, multi-step process starting with 374 experimentally confirmed metal resistance genes from the MetGeneDB database, previously described in our work [9] (Fig 1A). Those sequences served as queries in DIAMOND BLASTP searches against SwissProt (obtained 07.08.2024), using stringent criteria (>90% identity and coverage), yielding 315 new sequences [10]. The combined set of 681 sequences was then used to search NCBI NR (obtained 04.04.2024), identifying 10,344 additional unique sequences and bringing the final dataset to 11,025 proteins. These sequences were then clustered using MMseqs2 [11] with clustering parameters optimized through a

gradient of identity and coverage values (0.5-0.99, 0.02 increments). The 0.90 threshold was selected, balancing mean cluster size and singleton count (Fig 1B). This resulted in 341 non-singleton clusters with a mean size of 27.98 proteins. The largest clusters included sequences encoding genes *mdtC* (811 sequences, represented by protein HCR6247170.1), *mdtB* (632 sequences; P69340), *cutC* (242 sequences; 1X7I), *fieF* (211 sequences; WP_085449461.1), and *zupT* (195 sequences; Q32BU5).

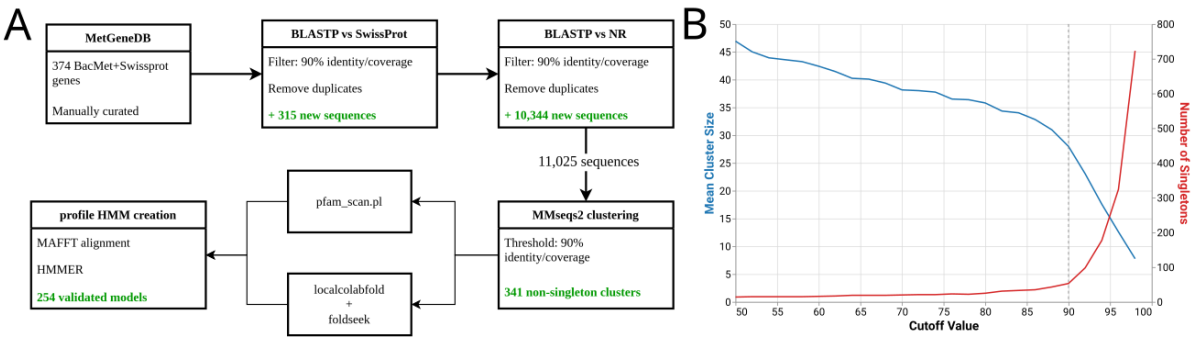


Figure 1. A) Database creation pipeline. Each box represents a separate step in the pipeline. B) Results of clustering parameters optimisation process. Coverage and identity value cutoffs are present on X-axis, while mean cluster size and number of singleton clusters can be seen on Y-axis.

For each non-singleton cluster, a comprehensive functional annotation using multiple complementary approaches was performed. The analysis combined domain detection using Pfam_scan.pl (Pfam v.37.1), protein structure prediction via LocalColabFold, and structural similarity searches against SwissProt using FoldSeek (E-value threshold ≤ 1e-3) [12–14]. Based on the resulting domain architectures and structural homologies to known SwissProt proteins, we determined the general function, metal specificity, and appropriate gene nomenclature for each cluster. This process identified 254 clusters associated with metal resistance mechanisms. Among these, copper-related clusters were most abundant (47),

84 followed by clusters involved in arsenic and mercury resistance (32 each). For annotated
85 clusters, multiple sequence alignments were generated using MAFFT (v7.525) and converted
86 to profile HMMs using HMMER3.4 (http://hmmer.org/) [15]. Each cluster was then assigned
87 a descriptive name containing the metal specificity, general function, and gene name when
88 possible. The database, alongside reference data and individual models is available at
89 https://github.com/Haelmorn/MetHMMDB.

90 To evaluate the practical utility of MetHMMDB, we compared its search speed to
91 DIAMOND BLASTP. Although hmmsearch exhibited slower processing times than
92 DIAMOND, its performance remained within acceptable limits for practical applications
93 (Supplementary Table 1). To evaluate detection capabilities, we analyzed a metagenome from
94 a metal-rich Alpine stream (PRJNA689378) containing 215,000+ protein sequences [16].
95 Using an E-value cutoff of 1e-5, hmmscan with MetHMMDB identified 3,846 metal resistance
96 genes compared to 1,786 found by DIAMOND blastp with MetGeneDB. Both methods
97 revealed similar distributions of resistance types, with *czcA*-like cobalt-zinc-cadmium
98 resistance proteins and *fbpC*-like iron transporters being predominant. The HMM-based
99 approach demonstrated enhanced sensitivity, successfully detecting previously reported *Acr3*
100 and *ArsACHM* arsenic resistance genes, notably revealing the presence of multiple copies of
101 the aioAB arsenite oxidase—a finding not reported in the original study. The methods showed
102 substantial agreement, with 1,636 proteins (91.6% of blastp results) detected by both
103 approaches. The HMM-based method primarily missed genes with low sequence conservation
104 (mean identity 50.1%, coverage 26.3%), including *acn* (aconitate hydratase), *nrsD* (MFS
105 superfamily transporter), and *pitA* (inorganic phosphate transporter) genes.

106 Finally, we checked the performance of the database on agricultural soil metagenomes
107 from BioProject PRJEB22376 (ERR2486626, ERR2486627) to evaluate its ability to
108 differentiate healthy and unhealthy soil samples. The analysis revealed distinct metal resistance

profiles between metagenomes, with multidrug *mdtBC* genes and iron transport related genes enriched in healthy soil, while nickel transporters and *fecE*-like iron transporter were more abundant in unhealthy soil (Supplementary Figure 1). Notably, two *znuC* zinc transporter variants displayed opposite patterns: a *Bacillus*-derived variant dominated in healthy soil, while a *Salmonella*-derived variant was more common in unhealthy soil, which might contribute to perceived soil quality differences (Fig 2).

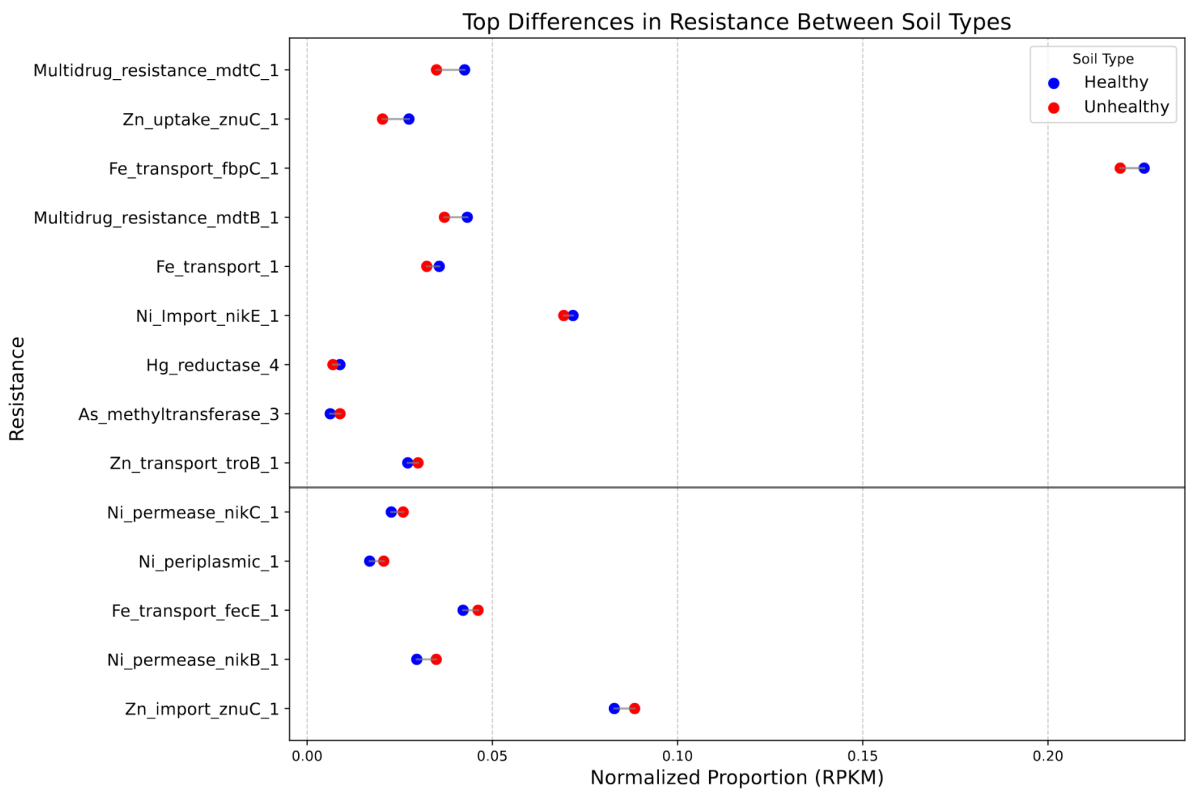


Figure 2. Comparison of resistance mechanisms between healthy and unhealthy soil samples. Top half of the plots shows 5 hits highest difference

In summary, MetHMMDB represents the first profile HMM-based database, specifically designed for MMRGs. By integrating sequence analysis, structural predictions, and manual curation, we developed 254 functionally annotated models that prioritize molecular function over traditional gene nomenclature. This approach improves identification accuracy and enables the detection of resistance determinants across diverse environmental contexts. The

database's utility was demonstrated through applications to both extreme environments and agricultural soils. These analyses showed that MetHMMDB offers superior detection sensitivity and functional specificity compared to traditional sequence-based methods, identifying a broader range of MMRGs while providing ecological insights. Though computationally more intensive, the database remains practical for large-scale analyses and is particularly well-suited for environmental monitoring and bioremediation research. With its focus on functional annotation rather than sequence similarity, MetHMMDB provides a robust tool for ecological and environmental studies.

## DATA AVAILABILITY

The datasets generated during and/or analysed during the current study are available in the GitHub repository, https://github.com/Haelmorn/MetHMMDB.

136 **REFERENCES**


137 1. Das S et al. Heavy Metal Pollution in the Environment and Its Impact on Health: Exploring
138 Green Technology for Remediation. *Environ Health Insights* 2023; **17**: 11786302231201259.
139 https://doi.org/10.1177/11786302231201259
140 2. Darham S et al. Antarctic heavy metal pollution and remediation efforts: state of the art of
141 research and scientific publications. *Braz J Microbiol* 2023; **54**: 2011–2026.
142 https://doi.org/10.1007/s42770-023-00949-9
143 3. Briffa J, Sinagra E, Blundell R. Heavy metal pollution in the environment and their
144 toxicological effects on humans. *Heliyon* 2020; **6**. https://doi.org/10.1016/j.heliyon.2020.e04691
145 4. Roy R et al. An Overview of Bacteria-Mediated Heavy Metal Bioremediation Strategies. *Appl*
146 *Biochem Biotechnol* 2024; **196**: 1712–1751. https://doi.org/10.1007/s12010-023-04614-7
147 5. Li L et al. Genome-resolved metagenomics provides insights into the ecological roles of the
148 keystone taxa in heavy-metal-contaminated soils. *Front Microbiol* 2023; **14**.
149 https://doi.org/10.3389/fmicb.2023.1203164
150 6. Ekhlas D et al. Metagenomic comparison of the faecal and environmental resistome on Irish
151 commercial pig farms with and without zinc oxide and antimicrobial usage. *Animal Microbiome*
152 2023; **5**: 62. https://doi.org/10.1186/s42523-023-00283-3
153 7. Balcha ES et al. Shotgun Metagenomics-Guided Prediction Reveals the Metal Tolerance and
154 Antibiotic Resistance of Microbes in Poly-Extreme Environments in the Danakil Depression,
155 Afar Region. *Antibiotics* 2023; **12**: 1697. https://doi.org/10.3390/antibiotics12121697
156 8. Pal C et al. BacMet: antibacterial biocide and metal resistance genes database. *Nucleic Acids*
157 *Res* 2014; **42**: D737–D743. https://doi.org/10.1093/nar/gkt1252
158 9. Dziurzynski M et al. Development of the LCPDb-MET database facilitating selection of PCR
159 primers for the detection of metal metabolism and resistance genes in bacteria. *Ecological*
160 *Indicators* 2022; **145**: 109606. https://doi.org/10.1016/j.ecolind.2022.109606
161 10. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using
162 DIAMOND. *Nat Methods* 2021; **18**: 366–368. https://doi.org/10.1038/s41592-021-01101-x

11. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017; **35**: 1026–1028. https://doi.org/10.1038/nbt.3988

12. Mistry J et al. Pfam: The protein families database in 2021. *Nucleic Acids Research* 2021; **49**: D412–D419. https://doi.org/10.1093/nar/gkaa913

13. Mirdita M et al. ColabFold: making protein folding accessible to all. *Nat Methods* 2022; **19**: 679–682. https://doi.org/10.1038/s41592-022-01488-1

14. Van Kempen M et al. Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* 2024; **42**: 243–246. https://doi.org/10.1038/s41587-023-01773-0

15. Eddy SR. Accelerated profile HMM searches. *PLoS Computational Biology* 2011; **7**. https://doi.org/10.1371/journal.pcbi.1002195

16. Buetti-Dinh A et al. Geochemical and metagenomics study of a metal-rich, green-turquoise-coloured stream in the southern Swiss Alps. *PLOS ONE* 2021; **16**: e0248877. https://doi.org/10.1371/journal.pone.0248877

## AUTHOR CONTRIBUTIONS

KC and MD conceptualized the study. KC performed data curation, formal analysis, validated and visualised the data. KC and MD conducted the investigations. KC wrote the original draft. KC and MD reviewed and edited the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.