Toward a Universal Unit for Quantification of Antibiotic Resistance Genes in Environmental Samples


Cite This: Environ. Sci. Technol. 2023, 57, 9713−9721

ABSTRACT: Surveillance of antibiotic resistance genes (ARGs) has been increasingly conducted in environmental sectors to complement the surveys in human and animal sectors under the “One-Health” framework. However, there are substantial challenges in comparing and synthesizing the results of multiple studies that employ different test methods and approaches in bioinformatic analysis. In this article, we consider the commonly used quantification units (ARG copy per cell, ARG copy per genome, ARG density, ARG copy per 16S rRNA gene, RPKM, coverage, PPM, etc.) for profiling ARGs and suggest a universal unit (ARG copy per cell) for reporting such biological measurements of samples and improving the comparability of different surveillance efforts.

KEYWORDS: one-health, metagenomics, long-read sequencing, ARG copy per cell, standardization

INTRODUCTION

Antimicrobial resistance (AMR) has emerged as a leading public health threat in recent decades. The global burden of antimicrobial resistant bacterial infections was associated with 4.95 million deaths in 2019, highlighting the urgency to combat this public health challenge. Various national and international programs for surveillance of antibiotic resistant bacteria (ARB) and antibiotic resistance genes (ARGs) have been conducted, and more efforts are underway to better understand the scope and drivers of AMR.

To complement the surveys in human and animal sectors under the “One-Health” framework, surveillance has been increasingly conducted in environmental sectors, including manure, soil, freshwater, marine water, and wastewater treatment plants. Conventional cultivation-based methods for characterizing ARB provide comprehensive information in terms of ARB phylogeny, resistance phenotype, and transmissibility of the resistance determinants. However, cultivation is highly exhaustive, and frequently only a small fraction of environmental bacteria can be cultivated. Cultivation-independent methods based on nucleic acids directly extracted from environmental samples mirror and complement cultivation-based methods by overcoming the culture bias and are therefore increasingly applied as a monitoring tool.

Nucleic acid-based methods for environmental surveillance of ARGs are versatile approaches for flagging critical AMR dissemination routes to be prioritized for mitigation. However, there are substantial challenges in comparing and synthesizing the results of multiple studies that employ different test methods.
Table 1. Summary of Different Methods Used for Normalizing ARGs Quantity in Metagenomes

<table>
<thead>
<tr>
<th>terms used</th>
<th>definition/other names</th>
<th>equations(^a)</th>
<th>normalization</th>
<th>reads(^b)</th>
<th>refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PPM</td>
<td>part per million, number of ARG-like reads per million reads</td>
<td>[ \sum_{i=1}^{n} \frac{N_{\text{ARG-like reads}}}{S/10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}})</td>
<td>only short</td>
<td>28</td>
</tr>
<tr>
<td>RPKM</td>
<td>read per kilobase million, number of ARG-like reads per kilobase per million bacterial (or prokaryotic) reads. Another similar unit is FPKM, which treats paired-end reads as fragments</td>
<td>[ \sum_{i=1}^{n} \frac{N_{\text{ARG-like reads}}}{L_{\text{ARG reference sequence}}/1000} \times \frac{S}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{L_{\text{ARG reference sequence}}/1000} \times \frac{S}{10^6} )</td>
<td>only short</td>
<td>29–32</td>
</tr>
<tr>
<td>coverage</td>
<td>ARG copy of mapped nucleotide bases per Giga bases, also termed as Genecopy/Gb</td>
<td>[ \sum_{i=1}^{n} \frac{R_{\text{ARG-like reads}}}{L_{\text{ARG reference sequence}}} \times \frac{B}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{L_{\text{ARG reference sequence}}} \times \frac{B}{10^6} )</td>
<td>both short and long</td>
<td>26, 33, 34</td>
</tr>
<tr>
<td>ARG copy per 16S</td>
<td>copy of ARGs per copy of 16S rRNA genes</td>
<td>[ \sum_{i=1}^{n} \frac{R_{\text{ARG-like reads}}}{16S_{\text{reference sequence}}} \times \frac{B}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{16S_{\text{reference sequence}}} \times \frac{B}{10^6} )</td>
<td>both short and long</td>
<td>34</td>
</tr>
<tr>
<td>ARG copy per genome</td>
<td>copy of ARGs normalized against the estimated bacterial (or prokaryotic) genome number (one option is to base on the number of sequenced bases divided by the average or median bacterial (or prokaryotic) genome size)</td>
<td>[ \sum_{i=1}^{n} \frac{R_{\text{ARG-like reads}}}{N_{\text{marker genes}}} \times \frac{B}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{N_{\text{marker genes}}} \times \frac{B}{10^6} )</td>
<td>both short and long</td>
<td>35</td>
</tr>
<tr>
<td>ARG copy per cell</td>
<td>copy of ARGs normalized against the cell number [one option is to apply the mean copies of 30 essential bacterial (or prokaryotic) single-copy marker genes (ESCMGs)]</td>
<td>[ \sum_{i=1}^{n} \frac{R_{\text{ARG-like reads}}}{S_{\text{marker genes}}} \times \frac{B}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{S_{\text{marker genes}}} \times \frac{B}{10^6} )</td>
<td>both short and long</td>
<td>36</td>
</tr>
<tr>
<td>ARG density</td>
<td>RPKM of ARG normalized against the mean RPKM of 40 single copy genes (SCGs)</td>
<td>[ \frac{1}{30} \times \sum_{j=1}^{30} \sum_{k=1}^{n} \frac{R_{\text{marker reads}}}{L_{j_{\text{marker gene}}}} \times \frac{S_{\text{marker genes}}}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{L_{j_{\text{marker gene}}}} \times \frac{S_{\text{marker genes}}}{10^6} )</td>
<td>both short and long</td>
<td>37</td>
</tr>
<tr>
<td>absolute quantification</td>
<td>copy per unit mass or volume (e.g., per gram or mL)</td>
<td>[ \sum_{i=1}^{n} \frac{R_{\text{ARG-like reads}}}{L_{\text{ARG reference sequence}}} \times \frac{\text{SF}}{S_{\text{marker genes}}} \times \frac{B}{10^6} ]</td>
<td>(\sqrt{\text{sequencing depth}} \times \frac{\text{ARG length}}{L_{\text{ARG reference sequence}}} \times \frac{\text{SF}}{S_{\text{marker genes}}} \times \frac{B}{10^6} )</td>
<td>both short and long</td>
<td>38</td>
</tr>
</tbody>
</table>

\(^a\) Equations in Table 1 may not be the exact original equations in the corresponding literature, but their definitions have been strictly followed. Here, \(n\) is the total number of reference sequences in the ARG database; \(N_{\text{ARG-like reads}}\) is the number of ARG-like reads assigned to the \(i\)-th ARG reference sequence based on a cut-off criteria; \(S\) is the total number of reads in the dataset; \(S\) is the total number of reads assigned to bacteria (or prokaryote) in the dataset; \(L_{\text{ARG reference sequence}}\) is the nucleotide sequence length of the \(i\)-th reference sequence in the ARG database; \(R_{\text{ARG-like reads}}\) is the sum of aligned lengths of all reads assigned to the \(i\)-th ARG reference sequence; \(B\) is the number of bases in the dataset; \(m\) is the total number of reference sequences of 16S rRNA gene (or other marker genes); \(j\) represents the index of a specific reference sequence in the 16S rRNA gene (or other marker genes) database; \(R_{\text{marker reads}}\) is the sum of aligned lengths of all reads that mapped to the \(j\)-th reference sequence of 16S rRNA gene database; \(L_{j_{\text{marker gene}}}\) is the nucleotide sequence length of the \(j\)-th 16S rRNA reference sequence; \(N_{\text{marker genes}}\) is the number of reads aligned to the \(j\)-th reference sequence in the \(k\)-th marker gene cluster of ESCMGs, where currently 30 genes are usually used; \(L_{j_{\text{marker gene}}}\) is the nucleotide sequence length of the \(j\)-th reference sequence in the \(k\)-th marker gene cluster of ESCMGs or SCGs; \(N_{\text{marker reads}}\) is the number of reads aligned to the \(j\)-th reference sequence in the \(k\)-th marker gene cluster of SCGs; \(\text{SF}\) (scaling factor) is the ratio between the known gene copy (or cell number) of spike-in cells and estimated gene copy (or cell number) in the dataset. \(^b\) This means short or long sequencing reads obtained by different sequencing platforms, i.e., short reads by Illumina and long reads by PacBio or Nanopore.
and approaches in bioinformatic analysis. Harmonizing such measures would help the communication and consolidation of knowledge in this research area, including more precisely linking ARGs to ARB for public-health decision-makers. Thus, efforts toward standardization are needed. Here, we consider the commonly used quantification units for profiling ARGs and suggest a universal unit for reporting such biological measurements of samples and improving the comparability of different surveillance efforts.

**QUANTIFICATION UNIT USING QPCR METHODS**

To detect and quantify targeted ARGs, real-time quantitative polymerase chain reaction (qPCR) methods have been widely applied using ARG primer sets. High-throughput qPCR (HT-qPCR) platforms integrate multiple ARG primers to simultaneously measure hundreds of targeted ARGs. Both methods usually express ARG quantities as relative abundance using gene copy number per 16S rRNA gene copy number (copy per 16S), which helps account for variable biomass and bacterial recovery across samples and provides a proxy for the proportion of bacteria carrying ARGs. Sometimes, the absolute abundance using gene copy number per unit mass or volume (copy per unit mass or volume) is also derived from qPCR measurements, with a general assumption of consistent DNA extraction efficiencies, for example, 100%.

**QUANTIFICATION UNIT USING METAGENOMICS**

Metagenomics is becoming a widely accepted broad-spectrum tool for quantitative ARG analysis, avoiding qPCR limitations such as the availability of primers/probes. Although the current metagenomic methods are still less sensitive than qPCR, they have much higher throughput in terms of the number of ARGs covered, bioinformatic databases/tools/procedure and data-sharing platforms, and have been applied extensively for the

---

Figure 1. Total abundances of ARGs in representative samples (number of datasets: 1809) expressed in four units: (a) ARG copy per cell; (b) ARG copy per 16S rRNA gene; (c) RPKM (# of ARG-like read per kilobase million); and (d) PPM (# of ARG-like reads part per million). The X-axis order of sample type names of the four panels is identical and ordered by ascending total abundance in the unit of ARG copy per cell, as shown in panel (a). All metagenomes analyzed in this study were from the ARGs Online Searching Platform v2.0.
surveillance of ARGs. Metagenomic analysis has also enabled the discovery of new ARGs when combined with functional screening. Additionally, existing DNA sequence datasets can be retrospectively explored with expanded ARG databases after adding the novel ARG sequences. The increasing capacity and decreasing costs of long-read sequencing make them an accessible tool to link ARGs to mobile genetic elements (MGEs), providing valuable information to track its evolution and hosts for a more comprehensive risk assessment.

Currently, the quantity of ARGs obtained using metagenomics is expressed in different ways, making it difficult to directly compare the results of different studies (Table 1). For example, RPKM normalizes the data by considering both ARG lengths and the sequencing depth in the normalization process. However, the biological and ecological meanings of RPKM are not straightforward, although they could be transformed into more meaningful units of biological and/or clinical significance by making some assumptions. Copy numbers of ARGs per copy of 16S rRNA genes (ARG copy per 16S) provide a useful unit for comparison with qPCR data. However, different species vary in the number of 16S rRNA gene copy per genome, and it is still not that easy to compare the results of different studies.

We recommend the unit of ARG copy per cell (also equivalent to copy per genome if we assume one genome per cell), which normalizes sequencing depth, ARG lengths, and prokaryotic cells-genomes in the dataset since it is more suited to present the ARG relative abundance in samples. The unit of ARG copy per cell has a straightforward biological clinical meaning, allowing for better communication and comparison, and can be further developed into absolute quantification when actual cell count data are available by flow cytometry, microscopy, or spike-in methods. A unit similar to ARG copy per cell is “ARG density,” which estimates RPKM for ARGs and normalizes by the average RPKM of 40 single copy marker genes. Density has the same biological meaning as the average ARG copy carried by a single bacterial cell, but it is not applicable to long-read metagenomics since it is based on the read number. Overall, specific methodological scenarios may need to adopt different units; however, the use of a universal unit with biological meaning may facilitate systematic and direct comparisons of ARG levels across different samples or studies.

**BENCHMARKS OF ARGs IN DIFFERENT UNITS IN REPRESENTATIVE SAMPLE TYPES**

In addition to the comparability of different studies, another obvious question is whether using different units will lead to different conclusions. To address this question, we quantified ARGs in representative sample types and applied four of the units described above, including PPM, RPKM, ARG copy per 16S rRNA gene, and ARG copy per cell. The metagenomic datasets were first trimmed by removing adapters and low-quality sequences using Trimomatic. Then, clean reads were aligned against the SARG database using the ARGs-OAP v3.2. This tool can simultaneously estimate the copy of 16S rRNA genes by referring to the Greengenes database and estimate cell numbers by referring to bacterial essential single copy marker genes, followed by normalization of ARG quantitation into different units, including PPM, RPKM, ARG copy per 16S rRNA gene, and ARG copy per cell.

Compared with ARG copy per cell, the choice of other units for quantifying ARGs can lead to different ranking, as demonstrated in Figure 1: ARG copy per 16S rRNA gene, RPKM, and PPM do not result in the same order of ARG copy per cell. For example, ARG abundance in cattle feces was lower than in wastewater treatment plants (WWTPs) samples when expressed in ARG copy per 16S rRNA gene (Mann-Whitney U test, $P = 0.98$) but was significantly higher when the result was expressed in ARG copy per cell (Mann-Whitney U test, $P = 2.4 \times 10^{-3}$). The differences can be explained because cattle fecal bacteria have higher average 16S rRNA gene copy numbers than bacteria in the WWTPs.

Moreover, the choice of PPM or RPKM can flatten differences among samples. For example, the total abundance of ARGs in marine water and natural sediments showed no discernible difference when assessed with the unit of PPM (Mann-Whitney U test, $P = 0.31$) or RPKM (Mann-Whitney U test, $P = 0.55$), but there was a significant difference in the unit of ARG copy number per cell (Mann-Whitney U, $P = 1.6 \times 10^{-3}$).

This discrepancy occurs because RPKM and PPM normalize the sequencing depth but do not capture variations in bacterial genome sizes. Previous reports on Earth’s microbiomes have predicted larger genome sizes in soils and sediments than in marine environments and explain the greater variability in terrestrial environments. We found that when the abundance values in RPKM were similar, samples with larger mean genome sizes resulted in higher abundance values of ARG copy per cell.

**IMPLICATIONS**

Since AMR is an increasing global concern, scientific research and environmental management require systematic, quantitative, and comparable ARG datasets. The consensus around ARG units will help identify important sources and natural reservoirs of AMR, inform effective mitigation strategies, and identify key trends across countries, world regions, and One Health compartments. This goal requires a standardized approach, including universally adopted quantification units. Certainly, different units used to quantify ARGs have individual strengths, can be applied in different scenarios, and complement each other. However, to study the resistome across a wide range of complex environmental systems, we recommend the unit of “ARG copy per cell” as the most pragmatic option, considering its strengths discussed above, while other units are also useful for different scenarios and could be reported together with it.

Unlike metagenomic analysis, which offers numerous ways to obtain cell numbers (genome counts) for analysis, qPCR currently estimates cell numbers by measuring 16S rRNA gene abundance and converting it to cell numbers by dividing the average copy numbers of 16S rRNA genes per cell. In the future, we suggest designing primers that target multiple single-copy marker genes. This would allow cell numbers to be estimated directly from qPCR, thus providing the ARG copy per cell estimate.

To normalize ARG quantification in the unit of ARG copy per cell, the denominator (cell numbers) and numerator (ARG copy number) are estimated separately following the formula in Table 1. It is worth pointing out that the same value using the same unit “ARG copy per cell” may still have different implications since ARGs may vary in the detection of genes by the different method employed and also in their associated risks. Other factors, including variance in the measuring process from sampling to wet-lab differences, can also matter.

The unit of ARG copy per cell is still a relative rather than an absolute measure and requires measurement of the cell number per volume or mass of sample to become truly quantitative. Both
absolute and relative measures are essential for a comprehensive evaluation of antibiotic resistance in environmental settings and provide complimentary insights, e.g., regarding mass flows and exposure risks, on one hand, and processes of resistance selection and comparison of widely different environments (e.g., dense gut microbiomes versus water), on the other. The unit “ARG copy per cell” also has an additional advantage that it can be transformed to absolute abundance data with ease if cell abundance data are available.34 This unit can be further developed into absolute quantification of viable and membrane-compromised dead ARB cells if combined with the pretreatment using a cell-impermeant dye like propidium monoazide (PMA).35

Overall, we recommend that the research community adopt “ARG copy per cell” as a consensus unit as the first step in the adoption of a standardized analytical approach to ensure a comparable and biologically relevant interpretation of the results in order to combat antibiotic resistance. Such an approach could be readily applied to other functional genes that potentially confer phenotypes of concern, including metal resistance, disinfectant resistance, pathogenicity, virulence genes, and the MGEs that disseminate all these determinants.

## AUTHOR INFORMATION

### Corresponding Author

Tong Zhang — Environmental Microbiome Engineering and Biotechnology Laboratory, Center for Environmental Engineering Research, Department of Civil Engineering, The University of Hong Kong, Pokfulam 99077 Hong Kong, China; orcid.org/0000-0003-1148-4322; Email: zhangt@hku.hk

### Authors

Xiaole Yin — Environmental Microbiome Engineering and Biotechnology Laboratory, Center for Environmental Engineering Research, Department of Civil Engineering, The University of Hong Kong, Pokfulam 99077 Hong Kong, China

Xi Chen — Environmental Microbiome Engineering and Biotechnology Laboratory, Center for Environmental Engineering Research, Department of Civil Engineering, The University of Hong Kong, Pokfulam 99077 Hong Kong, China

Xiao-Tao Jiang — Microbiome Research Centre, St George and Sutherland Clinical School, University of New South Wales, 2052 Sydney, Australia

Ying Yang — School of Marine Sciences, Sun Yat-sen University, 519082 Zhuhai, China

Bing Li — State Environmental Protection Key Laboratory of Microorganism Application and Risk Control, Tsinghua Shenzhen International Graduate School, Tsinghua University, FS18055 Shenzhen, China; orcid.org/0000-0002-7161-0477

Marcus Ho-Hin Shum — State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Pokfulam 99077 Hong Kong, China

Tommy T. Y. Lam — State Key Laboratory of Emerging Infectious Diseases, School of Public Health, The University of Hong Kong, Pokfulam 99077 Hong Kong, China

Gabriel M. Leung — Laboratory of Data Discovery for Health, Hong Kong Science & Technology Parks, New Territories 99077 Hong Kong, China

Joan Rose — Department of Fisheries and Wildlife, Michigan State University, East Lansing 48824 Michigan, United States

Concepcion Sanchez-Cid — Environmental Microbial Genomics, CNRS UMR 5005 Laboratoire Ampère, École Centrale de Lyon, Université Claude Bernard Lyon 1, Université de Lyon, 69130 Écully, France

Timothy M. Vogel — Environmental Microbial Genomics, CNRS UMR 5005 Laboratoire Ampère, École Centrale de Lyon, Université Claude Bernard Lyon 1, Université de Lyon, 69130 Écully, France

Fiona Walsh — Department of Biology, Maynooth University, RS1 Co. Kildare, Ireland

Thomas U. Berendonk — Faculty of Environmental Sciences, Technische Universität Dresden, Institute for Hydrobiology, 01217 Dresden, Germany

Janet Midega — Wellcome Trust, NW1 2BE London, U.K.

Chibuzor Uchea — Welcome Trust, NW1 2BE London, U.K.

Dominic Frigon — Department of Civil Engineering and Applied Mechanics, McGill University, Montreal H3A 0C3 Québec, Canada; orcid.org/0000-0003-1587-8943

Gerard D. Wright — Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton L8N 3Z5 Ontario, Canada; orcid.org/0000-0002-9129-7131

Carlos Bezuendehout — Unit for Environmental Sciences and Management (UESM)-Microbiology, North-West University, 2531 Potchefstroom, South Africa

Renata C. Picão — Medical Microbiology Department, Paulo de Góes Microbiology Institute of the Federal University of Rio de Janeiro, 21941-902 Rio de Janeiro, Brazil

Shaikh Z. Ahammad — Department of Biochemical Engineering & Biotechnology, Indian Institute of Technology Delhi, 110016 New Delhi, India; orcid.org/0000-0002-4459-0239

Per Halkjær Nielsen — Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University, 9210 Aalborg, Denmark

Philip Hugenholtz — School of Chemistry and Molecular Biosciences, Australian Centre for Ecogenomics, The University of Queensland, Brisbane 4072 Queensland, Australia

Nicholas J. Ashbolt — Faculty of Science and Engineering, Southern Cross University, Bilinga 4225 Queensland, Australia; orcid.org/0000-0002-3853-0096

Gianluca Corno — Molecular Ecology Group (MEG), Water Research Institute, National Research Council of Italy (CNR-IRSA), 28922 Verbania, Italy; orcid.org/0000-0002-7423-8797

Despo Fatta-Kassinos — Department of Civil and Environmental Engineering and Nireas International Water Research Center, University of Cyprus, 1678 Nicosia, Cyprus

Helmut Bürgmann — Eawag: Swiss Federal Institute of Aquatic Science and Technology, 6047 Kastanienbaum, Switzerland; orcid.org/0000-0002-5651-5906

Heike Schmitt — Centre for Zoonoses and Environmental Microbiology-Centre for Infectious Disease Control, National Institute for Public Health and the Environment (RIVM), 3721 Bilthoven, The Netherlands; Department of Biotechnology, Delft University of Technology, 2628 Delft, the Netherlands

Chang-Jun Cha — Department of Systems Biotechnology and Center for Antibiotic Resistance, Chung-Ang University, 17546 Anseong, Republic of Korea; orcid.org/0000-0002-2210-2898

Amy Pruden — The Charles Edward Via, Jr., Department of Civil and Environmental Engineering, Virginia Tech,
ACKNOWLEDGMENTS

The work was substantially supported by a Theme-based Research Scheme grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project no. T21-705-20-N).

REFERENCES


Notes

The authors declare no competing financial interest.


Environmental Science & Technology


(51) Martínez-Gutierrez, C. A.; Aylward, F. O. Genome size distributions in bacteria and archaea are strongly linked to evolutionary history at broad phylogenetic scales. *PLoS Genet.* 2022, 18, e1010220.


Recommended by ACS

Emergence of Mobile Colistin Resistance Genes within Los Angeles County Wastewater

Philip Wang and Adam L. Smith

MARCH 17, 2023

ENVIRONMENTAL SCIENCE & TECHNOLOGY LETTERS

Uncovering Bacterial Hosts of Class 1 Integrons in an Urban Coastal Aquatic Environment with a Single-Cell Fusion-Polymerase Chain Reaction Technology

Qin Qi, Michael R Gillings, et al.

MARCH 13, 2023

ENVIRONMENTAL SCIENCE & TECHNOLOGY

Snapshot ARG Removal Rates across Wastewater Treatment Plants Are Not Representative Due to Diurnal Variations

Esther G. Lou, Lauren B. Stadler, et al.

DECEMBER 15, 2022

ACS ES&T WATER

Effects of Antibiotic Resistance Genes and Antibiotics on the Transport and Deposition Behaviors of Bacteria in Porous Media

Shuai Wang, Meiping Tong, et al.

JULY 05, 2023

ENVIRONMENTAL SCIENCE & TECHNOLOGY

Get More Suggestions >