Global distribution of mcr gene variants in 214,095 metagenomic samples

Hannah-Marie Martiny (✉ hanmar@food.dtu.dk)

National Food Institute, Technical University of Denmark <https://orcid.org/0000-0001-6733-7888>

Patrick Munk

National Food Institute, Technical University of Denmark <https://orcid.org/0000-0001-8813-4019>

Christian Brinch

National Food Institute, Technical University of Denmark

Judit Szarvas

National Food Institute, Technical University of Denmark <https://orcid.org/0000-0002-0720-3017>

Frank Aarestrup

National Food Institute, Technical University of Denmark <https://orcid.org/0000-0002-7116-2723>

Thomas Petersen

National Food Institute, Technical University of Denmark

Article

Keywords: mobilized colistin resistance gene (mcr-1), antimicrobial resistance

Posted Date: January 12th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-640935/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Global distribution of *mcr* gene variants in 214,095 metagenomic samples

Authors:

Hannah-Marie Martiny^{1#},
Patrick Munk¹,
Christian Brinch¹,
Judit Szarvas¹,
Frank M. Aarestrup¹,
Thomas Nordahl Petersen¹

Affiliations:

¹Research Group for Genomic Epidemiology, Technical University of Denmark, Kgs. Lyngby,
Denmark

#Address correspondence to Hannah-Marie Martiny, hanmar@food.dtu.dk

Abstract

Since the initial discovery of a mobilized colistin resistance gene (*mcr-1*), several other variants have been reported, some of which might have circulated a while before being discovered. Metagenomic data provides an opportunity to re-analyze available older data to understand the evolutionary history of recently discovered antimicrobial resistance genes (ARGs). Here, we present a large-scale metagenomic study of 442 Tbp of sequencing reads from 214,095 samples to identify the host and geographical distribution and genomic context of nine *mcr* gene variants (*mcr-1* to *mcr-9*). Our results show that the dissemination of each variant is not uniform. Instead, the source and location play a role in the spread. Flanking region analyses of metagenomic assemblies suggest that despite the very diverse distribution, the genomic background of the *mcr* genes remained unchanged as the same mobile genetic elements and plasmid replicons occur. This work emphasizes the importance of sharing genomic data for the surveillance of ARGs in our understanding of antimicrobial resistance.

Importance

The ever-growing collection of metagenomic samples available in public data repositories have the potential to reveal new details on the emergence and dissemination of mobilized colistin resistance genes. Our analysis of metagenomes deposited online in the last 10 years show that the environmental distribution of *mcr* gene variants depends on sampling source and location, although the genetic context of each variant remained consistent.

Introduction

Antimicrobial resistance (AMR) is considered one of the most significant threats against human and animal health¹. Over the years, we have observed the emergence of a multitude of novel antimicrobial resistance genes (ARGs), and it is generally believed that such genes have emerged and evolved in the commensal flora for a long time prior to being detected in pathogenic isolates².

Colistin is an important antibiotic used as a last-resort choice to treat multidrug-resistant (MDR) and carbapenem-resistant bacteria³. Before 2015, colistin resistance was believed to be only due to mutational and regulatory changes in chromosomal genes. A mobilized colistin resistance gene *mcr-1* was discovered in 2015 on a plasmid in *Escherichia coli* isolates from China⁴, raising concern in the scientific community about the possibility of resistance spreading more rapidly by horizontal gene transfer by mobile genetic elements (MGEs)^{4,5}. Immediately following the first report, a large number of studies were initiated in several countries around the world, and it was soon determined that *mcr-1* was already widespread and has now been detected in all continents⁶⁻⁸. In initial reports, the most frequent isolates were sampled from livestock sources, followed by humans, meat, and food products⁹. Since then, several new variants of *mcr* genes have also been identified, named *mcr-2* to *mcr-10* sharing 81%, 32.5%, 34%, 36%, 83%, 35%, 31%, 36% and 29.31% amino acid sequence identity to *mcr-1*, respectively¹⁰⁻¹⁷. Retrospective screening of bacterial isolates and available sequences of mainly pathogenic isolates showed a more widespread occurrence and prior evolution of *mcr* before its initial discovery^{8,18,19}.

However, investigating only pathogenic strains or cultivable bacteria will only provide limited insight into the potential reservoirs of such novel ARGs. As documented by our research group and others, investigating the entire microbiome provides additional information on the presence and diversity of ARGs^{7,20–22}. Today most researchers conducting microbiome studies are depositing their raw data in global repositories allowing other researchers to re-analyze the data and provide novel insight.

This study was conducted to investigate the occurrence and global dissemination of known *mcr*-gene variants in publicly available metagenomic data sets. We downloaded 442 Tbp of raw reads from 214,095 metagenomic datasets and determined the presence and abundance of nine *mcr* gene variants. We find that only a small subset of the metagenomic datasets were positive for at least one of the *mcr* genes but that the abundance gradually increases as a function of time. The distribution of each variant varied by region and sampling source, but the genomic background of each gene was consistent across different environments. Despite the sparsity of the data once stratified by the presence of *mcr* genes, our analysis suggest that multiple factors have likely influenced the dissemination of colistin resistance and that screening publicly available metagenomic samples can, together with single isolates, further deepen our understanding of the distribution of *mcr* gene variants.

Methods

Data collection

Metagenomic datasets were collected from the public data repository, the European Nucleotide Archive (ENA)²³. We queried the API for samples uploaded to ENA between 2010-01-01 and 2020-01-01 that were shotgun sequenced and had at least 100,000 sequencing reads. In total, we downloaded 214,095 sequencing runs from 146,732 samples from 6,307 projects corresponding to 442 Tbp of raw reads.

Reference sequence databases

The AMR gene database ResFinder²⁴ (downloaded 25-01-2020) contains 3,085 sequences. The 16S rRNA Silva²⁵ gene database (version 138, downloaded 25-01-2020) contains 2,225,272 sequences.

Preprocessing and mapping sequencing reads

The raw FASTQ reads were quality checked using FASTQC v.0.11.15 (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) and trimmed with BBduk2 36.49²⁶ to remove low-quality sequence and adaptors. BBduk2 settings were set as follows: minimum read length set to 50 bp, k=19, kmin=11, tbo flag on 11, the Phred quality threshold at 20 (99% accuracy) and only right trimming (ktrim=r). Assignment of trimmed reads to reference sequences was done with global alignment using KMA²⁷ 1.2.21 with an alignment threshold of 0.5, which results in an effective identity of 90%. We used ResFinder to assess the number of acquired AMR reads in each sample, and Silva to determine the bacterial content. On average, it took 5.7 seconds per metagenome for ResFinder mapping and 232.7 seconds for Silva mapping on a node equipped with dual 20 core Xeon Gold 6230 CPUs clocked at 2.1Ghz using the Danish National Supercomputer for Life Sciences (<https://www.computerome.dk>).

Compositional data analysis

Our collected data have large variability in how the samples were collected, how DNA was extracted, and how it was sequenced. Furthermore, the probability of observing a reference depends on the sequencing depth per run. To account for some of the variability, we use read fragment counts as the gene counts for mapping against ResFinder genes and they were adjusted by individual gene lengths. Bacterial read fragment counts from Silva mapping were aggregated to a total sum for each sample and divided by a million.

The observed counts are relative to each sample, and if we fail to consider that downstream analyses can produce faulty results. By treating the data as compositional, we avoid creating false results. As proposed by Aitchison²⁸, we log-ratio transform the counts to make the data symmetric, linear, and in a log-ratio coordinate space.

Since a zero does not necessarily mean that a gene is absent from a sample and the logarithm of zero is an undefined value, we infer the proportion p_i of reads of a gene i within a sequenced sample directly from the observed read count n_i . If we assume that each n_i was sampled from a Poisson process $n_i \sim \text{Poisson}(\lambda_i)$ and the vector of counts follows a multinomial distribution $\{[n_1, n_2, \dots] | n\} \sim \text{Multinomial}(p_1, p_2, \dots | n)$ where $n = \sum_i n_i$ and $p_i = \frac{\lambda_i}{\sum_k \lambda_k}$. The posterior distribution of $[p_1, p_2, \dots]$ is given as the product of the multinomial likelihood with a Dirichlet $\left(\frac{1}{2}, \frac{1}{2}, \dots\right)$ prior. These inferred proportions will never be precisely zero, even if the observed count is zero because of the multivariate distribution.²⁹

Abundance tables of *mcr* genes were created by transforming the composition $x = \{p_1, p_2, \dots, p_D\}$ of read proportions with the additive log-ratio (ALR), in which the bacterial component is used as the reference p_D :

$$\text{ALR}(x) = \left(\log \frac{p_1}{p_D}, \dots, \log \frac{p_{D-1}}{p_D} \right)$$

For other downstream analyses, we use the centered log-ratio (CLR) transformation instead of ALR:

$$\text{CLR}(x) = \left(\log \frac{p_1}{g_m(p)}, \dots, \log \frac{p_D}{g_m(p)} \right)$$

where $g_m(p) = (\prod_{i=1}^D p_i)^{\frac{1}{D}}$ is the geometric mean of the composition.

Data visualization

Graphics visualizing abundance and relative abundances were created with Python3.8 with Matplotlib 3.3.2³⁰ and Seaborn 0.11.0³¹. Bar plots showing relative abundances are created by closing the composition to 100. Geographical maps showing gene abundances were created using shapely 1.3166.7.1³² and cartopy 0.18.0³³ to translate labels from the metadata into geographical shapes with the Natural Earth Dataset.

Statistical analysis

We carried out a differential abundance analysis on samples containing *mcr* fragments with ALDEx2³⁴ 1.18.0 in R. We aimed to identify which experimental groups showed a difference in their abundance of *mcr* genes compared to other groups. ALDEx2 tests for significant differences

of CLR abundance between categorical sample groups by using Welch's t-test followed by a Benjamini-Hochberg false discovery rate (FDR) correction³⁵. We report significant groups of either sample locations, host, or collection year where FDR < 0.05 and differential abundance were represented in an effect plot³⁶ displaying the within and between-group variation in CLR values.

Principal component analysis (PCA) was applied to the centered, scaled by total variance, and CLR transformed dataset³⁷ to reduce the dimensionality of the data. The eigenvectors and eigenvalues from PCA were used to create a biplot, highlighting the significant sample groups found in the differential abundance analysis.

Comparison of metagenomic abundance levels to occurrence in pathogen isolates

The NCBI Pathogen Detection Project³⁸ routinely screens new isolates to identify AMR genes with the tool AMRFinderPlus³⁹, which reports whether a gene is found or not in an assembled genome. We downloaded the annotation results of 912,469 assembled genomes from bacterial pathogens (<https://ftp.ncbi.nlm.nih.gov/pathogen/Results/> accessed 05-08-2021). 7,934 (0.87%) of the single isolates contained at least one of the nine *mcr* variants.

We grouped the isolates either by sampling year, location, or host, and report the log10 counts of samples available for each category and the isolates carrying *mcr* genes. Furthermore, we compared the relative count of each variant to the relative abundance levels in the metagenomes.

Metagenomic assembly of *mcr* samples

We assembled metagenomes where at least one of the *mcr* genes had a minimum coverage of 95% by trimmed reads, according to KMA. The trimmed reads were assembled with MetaSPAdes 3.14.0⁴⁰ with at least 1.2 terabytes of memory, 40 threads per node and a max runtime of one week. 145 metagenomes were not assembled out of 1014, as they did not complete within the chosen timeframe of a week. Contigs carrying the nine different *mcr*-genes variants were identified with blastn 11.0⁴¹ with a percentage identity ≥ 95 .

Flank analysis of metagenomic assemblies

The metagenomic contigs carrying *mcr* genes were used in the flank analysis. Flanks were created by masking the *mcr* gene in the contig and cutting out up- and down-stream regions of increasing sizes between 1,000 bp and 30,000 bp by intervals of 1,000 bp with bedtools⁴². The presence of plasmids in the flanks was identified with PlasmidFinder 2.1⁴³ and mobile elements with MobileElementFinder 1.0.3⁴⁴. Distance between the flanks was calculated as the Szymkiewicz-Simpson dissimilarity with KMA²⁷. Hierarchical clustering on the flank distances was done with Wards method⁴⁵ to create a dendrogram plotted with ggtree 2.0.4⁴⁶ and ggplot 3.3.3⁴⁷. This approach is similar to the workflow of the tool Flanker by Matlock et al. (2021)⁴⁸, except that we cluster with KMA.

Data and code availability

Source data for generating abundance figures, statistical and flank analysis can be found in the supplementary files and the supporting code is available upon request. Supplementary Data (SD) 1 contains ENA run accessions for the downloaded samples and which contains *mcr* genes, where

189 SD2 contains metadata and read fragment counts for *mcr*-positive samples and SD3 is a FASTA file
190 with contigs carrying *mcr* genes.

191 Results

192 Data set

193 Out of the 214,095 metagenomes, we found that 2.09% (4,465) of them contained read fragments
194 aligning to at least one of the nine *mcr* gene variants. The average number of reads per *mcr*
195 positive sample was 27 million reads, and on average, 0.003% of the reads were aligned to *mcr*
196 genes. Among the variants in the *mcr* family, *mcr-1* and *mcr-9* were the most frequent, with
197 25.91% and 57.47% of the *mcr*-aligned reads mapping to these variants, and disseminated across
198 respectively 10 and 13 sampling years, 21 and 56 countries, and 23 and 61 hosts. The rarest
199 variants were *mcr-2*, *mcr-6*, and *mcr-8* with read frequencies of 0.03%, 0.01%, and 0.08%,
200 respectively, and their metagenomic origins were more restricted (Table 1a). Overall, different log-
201 ratio abundance levels seemed to be different across the sampling years in different countries and
202 hosts (Figure S1).

203

204 Level of *mcr* variants over time

205 The *mcr* positive metagenomic samples were collected between 2003 and 2019, with the
206 exception of 2005, in which no *mcr* fragments were detected (Figure 1). Only two metagenomes
207 sampled in 2003 contained *mcr* fragments and a single *mcr* positive metagenome was from 2004.
208 Onwards, the percentage of positive samples fluctuated with the lowest value of 0.5% in 2008 and
209 highest 6.4% in 2019 (Figure S1). All the variants were frequently found in samples from 2016-
210 2017, except *mcr-6* that was only found in 2012 (Table 1b).

211

212 We found that the log ratio abundance fluctuated for the nine variants in each sampling year
213 (Figure 1). The oldest positive metagenomes were sampled in 2003 and 2004 and contained only
214 *mcr-3* and *mcr-5*. From 2006, the other variants began to emerge. *mcr-1* was detected first in
215 2009 at a low ALR abundance, and increased levels were observed between 2011 and 2019.
216 Similarly, *mcr-9* could be detected in low amounts in metagenomes from 2007. In 2012 and 2013,
217 *mcr-9* was the most abundant gene, with 81% and 86% of the read fragments aligning to this gene.
218 In 2007, only 3% of the *mcr* read fragments aligned to *mcr-7*, but more and more fragments for
219 each year were assigned to the *mcr-7* gene and peaked in 2019 with 95% of the *mcr* fragments
220 aligned being to it.

221

222 Significant levels of different *mcr* genes were observed for sampling years 2011, 2013, 2014, 2015,
223 2016, and 2017 (p-value < 0.05, Figure 4a). Even though the variance of *mcr* levels within the
224 sampling years was high, several variants stood out as having higher or lower levels in specific
225 years compared to other years. In 2011, *mcr-3* had a higher abundance than expected and
226 continued to be high in 2013-2014, together with *mcr-1* and *mcr-5*. *mcr-9* were lower in those
227 years. In 2016, the metagenomic picture changed as *mcr-3* and *mcr-5* had decreased levels, while
228 *mcr-1*, *mcr-4*, *mcr-7*, and *mcr-9* were increased.

229

230 Geographical distribution of *mcr* gene variants

231 The nine variants were spread across 95 different sampling locations (Figure S1), although samples
232 from different world regions were often different in which variant, they were positive for (Table

1c). A higher abundance of *mcr* gene variants were observed in the Americas, Asia, and Europe and decreased abundances in Africa. The highest total log-ratio abundances could be found in metagenomes from Australia, Lake Huron (USA), and Cambodia, and lowest levels in Kiribati, Greece, and the Caribbean Sea (Figure 2).

The individual variants were not equally distributed worldwide; instead, it seemed like specific variants were restricted to specific regions (Figure 2). For example, the variant *mcr-1* was less widely spread worldwide (Americas, Asia, and Europe) than *mcr-9* (Africa, Americas, Asia, Europe, Oceania, Atlantic Ocean, and the Pacific Ocean). No metagenomic locations contained all types of variants. In the Australian metagenomes, *mcr-9* was the most dominant gene, whereas *mcr-4* had high abundance levels in Lake Huron (USA) and *mcr-1* and *mcr-9* in Cambodian metagenomes. The only location of *mcr-6* was France.

15 of the 95 sampling locations had significant abundances of at least one of the *mcr* variants (p-value < 0.05), however the variance in the samples from most of the locations were high and did not have a large effect size compared to other locations (Figure 4c). Metagenomic locations that showed consistency within the group and were found to be different from the rest of the locations had lower levels of single variants: *mcr-1* in Bulgaria, *mcr-3* in Iceland, *mcr-5* in Malaysia, and *mcr-9* in Cambodia.

Host- and reservoir-specific *mcr* abundances

We found *mcr* genes present in 125 different sampling hosts and reservoirs, but with the various variants having different log-ratio abundances (Figure S2) and the two most frequent types differing for each variant (Table 1d). All six metagenomes from *Pomacea canaliculate* (golden apple snail) and the 11 *Danio rerio* (zebrafish) metagenomes contained *mcr* fragments. For two of the largest sampling groups, we found 897 out of 1,803 (49.75%) wastewater metagenomes and 13,831 out of 102,211 (1.35%) human derived samples *mcr* positive.

Out of the 125 hosts, only 20 of them showed significant levels of *mcr* gene fragments. These all had higher levels of colistin resistance genes (Figure 3). The dispersion within most of the 20 hosts was high, and their log-ratio levels were not significantly different from the other hosts, except a few (Figure 4e). The zebrafish samples had lower levels of *mcr-7* than expected, whereas golden apple snail metagenomes had higher levels. Panda metagenomes had elevated levels of *mcr-9* but had slightly lower amounts of *mcr-3*, *mcr-4*, and *mcr-5*. Metagenomes from pigs (*Sus scrofa* and pig gut) had increased levels of both *mcr-1* and *mcr-9*. Human metagenomes did not have large effect sizes but contained slightly less *mcr-1* and *mcr-9* than expected.

Diversity of *mcr* positive metagenomes

By performing compositional PCA analysis, we can visualize the variance in biplots showing which type of samples make the level of a resistance gene significant (Figure 4b,d,f). The biplots highlight a clear separation of metagenomes that contain *mcr-1* or *mcr-9*, and that these samples also differ a lot from each other. None of the samples from the different years are similar, which means that high levels of one of the variants cannot be explained simply due to a specific collection year (Figure 4b). Instead, we can see that several Panda metagenomes come from China in 2016, which most likely contribute to the higher levels of *mcr-1* in 2016 (Figure 4b,d). Likewise, human

metagenomes clearly show a geographical separation mainly driven by *mcr-1* being abundant in China and *mcr-9* in USA and Australia (Figure 4b,d, Figure S3), which could explain that even if these metagenomes contain significant levels of *mcr* genes, we could not observe large effect sizes. Excluding these two most abundant genes suggest however, that the differences are mainly driven by source and not by year or geographical location (Figure S4).

Distribution of *mcr* variants in pathogenic bacterial genomes

As several studies have performed retrospective screening of pathogenic bacterial isolates, we decided to compare the metagenomic *mcr* abundances to the prevalence in pathogenic single isolates. Out of 912,469 isolates screened by the NCBI Pathogen Detection Pipeline, only 7,934 (0.87%) were shown to carry at least one of the *mcr* genes. The majority of the *mcr* positive isolates contained either *mcr-1* (51.08%) or *mcr-9* (40.38%), while *mcr-6* and *mcr-7* were not detected at all (Table S1).

The congruence of relative counts in isolates and relative abundance levels in metagenomes varied depending on which allele and what kind of sample grouping it was. Grouped by the sampling location, the *mcr-1* gene appeared to be more widespread according to the isolates, whereas *mcr-3* had a larger global distribution based on the metagenomes. Similarly, for Human samples *mcr-1* had a higher prevalence in isolates, whereas metagenomes show a higher abundance of *mcr-9* (Figure S5).

Genomic background of *mcr* genes

The dissemination of colistin resistance genes between different reservoirs and countries in different sampling years was further investigated by creating assemblies of metagenomic samples with 95% coverage of at least one variant. We assembled 869 metagenomes, where we found 1,939 different contigs carrying *mcr* genes (range: 1-20 *mcr* contigs per metagenome). The most frequent gene present on these contigs was *mcr-9*, followed by *mcr-3* and *mcr-5* (Figure S6a). To identify structural patterns between different metagenomic origins, we analyzed the genetic signatures in regions up- and downstream of an *mcr* gene (the flanks) with a minimum size of 1,000 bp and a maximum of 21,000 bp to include most of the elements found in the flanks (Figure S6b). As most contigs were shorter than 1,000 (Figure S6a), only 138 contigs passed the size criteria. All 20 contigs containing plasmid replicons in their flanks carried *mcr-1* genes, whereas the 63 flanks with MGEs were on contigs with different *mcr* variants (Figure S7).

Six distinct clusters became apparent upon calculating the distance between the flanks surrounding the *mcr* genes (Figure 5). We find that the presence of specific MGEs seemingly correlate with a *mcr* variant is located on the contig, as ISAp1 occurs only on *mcr-1* contigs and IS903 on *mcr-9* contigs. Five of the six clusters are all flanks around the same variant, with two being *mcr-1* clusters, while the sixth contains flanks surrounding four different *mcr* gene variants.

In the first *mcr-1* cluster, an IncX4 plasmid replicon was present either upstream or downstream in most members. These contigs were found in metagenomes sampled in 2016 and 2017 from a diverse background, indicating that IncX4 plasmids are involved in multiple mobilization events in different settings. The second *mcr-1* cluster differs from the first in that we see an absence of IncX4 and IncI2 occurs in a few contigs instead. The cluster can instead be best characterized by the

323 presence of the insertion element IS*ApI1* in half of the flanks, that mainly originate from gut
324 metagenomes from Cambodia in 2014, while those without the insertion element are from Chinese
325 samples from 2015-2017.

326
327 Four of the *mcr-9* contigs contain IS102, IS26, or IS*Kpn43*, while the remaining carry IS903. Since all
328 these contigs contain at least one insertion sequence, it suggests that they were highly mobilized
329 between different metagenomes from human and environmental sources between 2008 and 2018.

330
331 The mixed cluster shows a surprising clustering of flanks on contigs of uneven lengths, with different
332 MGEs and replicons present, carrying four different *mcr* gene variants: *mcr-1*, *mcr-4*, *mcr-8*, and
333 *mcr-9*. This could indicate that these different alleles have been mobilized into the same genomic
334 region, despite the varying sample types.

335 Discussion

336 The growing collection of metagenomic sequencing data available in public repositories have the
337 potential to provide a much more detailed picture of the emergence, evolution and spread of
338 ARGs. We downloaded and analyzed 214,095 host derived and environmental metagenomes to
339 characterize the global distribution of nine mobilized colistin resistance genes that have been
340 identified since 2015. Among the downloaded samples, we found that 4,465 metagenomes (2%)
341 contained *mcr* reads distributed differently across the sampling period (2003-2019), geographical
342 and host origins. We found that all the nine different gene variants were present in metagenomes
343 sampled years before their discovery (Figure 1), confirming the notion of the resistance genes
344 circulating in the environment long before being reported^{18,19,49}. This confirms the value of
345 publicly sharing of raw NGS data allowing for later re-analyses.

346
347 To date, *mcr-1* is the most studied *mcr* gene variant, and the dissemination has been described in
348 detail. Multiple studies agree that even though *mcr-1* has been detected in a few isolates from the
349 1980s¹⁸, it has been appearing with increasing frequencies in samples between 2011 and 2017 and
350 decreasing in later years^{8,18,19,50}. We see a similar trend in the *mcr-1* positive frequencies, but the
351 levels seem to increase from 2008 with the highest levels in 2015; the year of *mcr-1* being
352 reported⁴ (Figure 1).

353
354 In just a few years after discovering *mcr-1*, multiple other *mcr* variants were reported in different
355 world regions, with *mcr-1* and *mcr-9* being the most disseminated genes⁶. Despite *mcr-9* being the
356 newest member¹⁷, we observe that it was the most abundant gene variant in publicly available
357 metagenomes, with *mcr-1* being the second most abundant. The two variants are not equally
358 distributed across sampling sites, as *mcr-1* appears to be more geographically restricted to Europe
359 and Asia, whereas *mcr-9* has reached a wider area (Figure 2). In human metagenomes, *mcr-1* and
360 *mcr-9* dominate, whereas other hosts and environmental origins display a considerable variation
361 in *mcr* variants. Despite earlier reports of *mcr-1* presence in both animal⁵¹ and environmental⁵²
362 origins, we see only very few environmental origins of the gene (Figure S2). Only a few hosts have
363 significantly different levels of *mcr* gene abundances than expected, where *mcr-1* and *mcr-9* tend
364 to be higher in pigs and pandas and *mcr-3*, *mcr-4*, *mcr-5*, and *mcr-7* lower in other hosts. *mcr-2*,
365 *mcr-6*, and *mcr-8* only appear in very few metagenomes, with *mcr-6* being the rarest variant. Since
366 the first report of *mcr-6*¹⁴, it has only been detected in very few places around the world and all in

367 2014 and 2015^{14,53}, but we can here report the presence of *mcr-6* in low amounts in a
368 metagenome from France sampled in 2012 (Figure 1, Figure 2). Overall, there appears to be a
369 connection between the abundance of a variant and the sampling source and location, but due to
370 the sparse nature of our dataset, we have not been able to determine the relative contribution of
371 these factors to the observed *mcr* levels.

372
373 The comparison of the relative abundance in metagenomes to the frequency in single isolates of
374 the different *mcr* alleles highlights the benefits of using metagenomic screening in addition to
375 single isolate analysis. Our metagenomic analysis showed that some of the variants were more
376 widely distributed than one would have known only based on the isolates analyzed in the NCBI
377 Pathogen Detection Pipeline, while other variants were more prevalent in single isolates. As both
378 data resources highly depend on what topic members of the scientific community focus on, there
379 is a sampling bias both in the results of screening single isolates and metagenomes. While the
380 compositional methods can handle some of the bias introduced by the different number of reads
381 available, bias resulting from the different nature of sequencing projects cannot be all removed.

382
383 Interestingly, we observed that the *mcr* contigs from the assembled metagenomes were well
384 conserved across reservoirs and locations except for *mcr-1* contigs. This suggests that most of the
385 variants have only been mobilized once and then spread globally and between reservoirs. In
386 contrast, *mcr-1* is known to be present in a variety of genomic backgrounds⁸, which we also
387 observe as the flanking regions of our *mcr-1* contigs grouped together in three distinct clusters
388 (Figure 5). *mcr-1* is commonly found on IncI2, IncHI2, IncX4 plasmids with ISApI1^{8,52}, although we
389 only observed ISApI1 on two IncI2 plasmids, a possible loss of ISApI1 near IncX4 replicons and that
390 no IncHI2 were present on *mcr-1* contigs. The absence of ISApI1 in one of the *mcr-1* clusters could
391 indicate a loss of mobility either due to their difference in sampling years or a shift in hosts. IS26
392 has been observed downstream of *mcr-9*⁵⁴, which we only observed once, and instead, we see
393 that IS903 occurs on both sides of *mcr-9* in the examined contigs. The metagenomic origin of *mcr-*
394 *9* contigs is highly diverse, suggesting that the presence of multiple different insertion sequences
395 has been a contributing factor in their mobilization between 2008 and 2018.

396
397 As we collected data by downloading publicly available metagenomic samples, we present a
398 dataset with uneven coverage of sampling locations and sources, which heavily influences our
399 ability to provide an in-depth understanding of the mobilization, emergence and spread of the *mcr*
400 genes. Regardless of this, we have shown the potential of using raw sequencing reads generated
401 by other researchers to improve our knowledge. It is, however, important that all such generated
402 data are shared publicly to allow for future exploration and improved understanding of the global
403 microbial biology⁵⁵. Since the start of this project, another *mcr* variant was discovered named *mcr-*
404 *10*⁵⁶, but we decided not to include the gene due to the massive computation task of mapping 442
405 Tb of raw sequencing reads. Nevertheless, it is indeed interesting to figure out when *mcr-10* first
406 appeared and characterize its dissemination as well, which we hope to do for this and for other
407 ARGs in the future.

a.		<i>mcr-1</i>	<i>mcr-2</i>	<i>mcr-3</i>	<i>mcr-4</i>	<i>mcr-5</i>	<i>mcr-6</i>	<i>mcr-7</i>	<i>mcr-8</i>	<i>mcr-9</i>
	Read frequency (%)	25.91	0.03	10.33	0.98	1.86	0.01	3.32	0.08	57.47
	No. years	10	6	14	13	13	1	13	11	13
	No. countries	21	6	59	42	43	1	27	14	56
	No. hosts	22	6	49	54	40	1	43	8	60

b.	Year (%)	<i>mcr-1</i>	<i>mcr-2</i>	<i>mcr-3</i>	<i>mcr-4</i>	<i>mcr-5</i>	<i>mcr-6</i>	<i>mcr-7</i>	<i>mcr-8</i>	<i>mcr-9</i>
	2010		16.67							
	2012					9.03	100.00		20.00	
	2013				17.04					
	2014									23.89
	2015	37.66								
	2016			58.77	20.89	61.11		30.68	16.00	22.11
	2017	22.86	16.67	14.36				17.40		

c.	Country (%)	<i>mcr-1</i>	<i>mcr-2</i>	<i>mcr-3</i>	<i>mcr-4</i>	<i>mcr-5</i>	<i>mcr-6</i>	<i>mcr-7</i>	<i>mcr-8</i>	<i>mcr-9</i>
	Angola		16.67							
	Cambodia	6.36								
	China	68.96						21.9		
	Denmark			40.11	12.16	36.24				
	France						100.0			
	Kenya								11.10	
	Netherlands		16.67							
	USA			13.43	39.38	18.78		23.63	11.11	45.18

d.	Host/reservoir (%)	<i>mcr-1</i>	<i>mcr-2</i>	<i>mcr-3</i>	<i>mcr-4</i>	<i>mcr-5</i>	<i>mcr-6</i>	<i>mcr-7</i>	<i>mcr-8</i>	<i>mcr-9</i>
	Homo sapiens	32.51	33.33		23.71				19.35	56.46
	Panda	22.59								
	Activated sludge metagenome					5.93				
	Freshwater metagenome			10.16				18.24		
	Marine metagenome		22.22						48.39	
	Microbial mat metagenome						100.0			
	Wastewater metagenome			65.77	26.20	60.81		25.88		13.44

Table 1: Read mapping of each *mcr* variant across different sample types. **a.** Read frequencies and count of unique sampling origins i.e., number of (No.) years, countries, and hosts and reservoirs. **b-d.** The top two sampling years (b.), countries (c.) or host and reservoir (d.) for the given variant was abundant in are shown in percentage of *mcr*-mapped reads.

Figure 1: **Discovery and the change of mcr genes over time.** Top: Timeline showing when each gene was first reported in the literature. Middle: Change in additive log-ratio (ALR) abundance values over time for each gene are shown, as well as the number of samples with or without a mcr hit from each year as bars. Bottom: The frequency of each gene as compared to the total mcr amount. Data were normalized with gene lengths to generate the charts.

Figure 2: **Global levels of mcr genes.** The large map shows the total abundance levels of all mcr genes, whereas the nine smaller plots show the individual gene levels in the world. A circle represents a collection of samples from water containing mcr genes. ALR: Additive log-ratio. White color indicates an absence of results, not that a specific location does not have any mcr genes. The circle markers indicate that the sample from a location comes from lakes or oceans.

Figure 3: Distribution of mcr genes in selected sampling hosts. Hosts were selected based on the host showing significant CLR values according to the ALDEx2 analysis (Figure 4). Top: Bar plot showing both the number of samples without and with an mcr hit and the overall mcr level for each host measured by ALR values. Bottom: The abundance of individual mcr genes relative to total mcr levels. Data were normalized with gene lengths before plotting. To see the distribution of mcr genes for all sampling hosts available, refer to Figure S2.

Figure 4: Analysis of significant *mcr* levels in sampling years, countries, and hosts. Left column: Visualizations of within-group dispersion of CLR values of individual *mcr* genes as compared to the between group difference in CLR values for **a.** sampling year, **c.** location and **e.** host. Right column: Compositional biplots of the first two principal components (PC) capturing 64% of the variation in the dataset, where samples are colored according to significant **b.** hosts, **d.** countries and **f.** hosts. Grey filled markers are samples that were non-significant.

Figure 5: The k-mer distance tree for flanks of mcr contigs with flank sizes between a 1000 bp and 21000 bp is drawn in the left panel with the metagenomic origins (year, country, host) added as colored tiles in the middle panel. The genomic background in the right is a schematic illustration of the size of each flank region in grey centered on the middle of the mcr gene and plasmid replicons and mobile genetic elements (MGEs) colored in the flanks.

References

1. *Antimicrobial Resistance: Global Report on Surveillance*. World Health Organization (2014).
2. Davies, J. Origins and evolution of antibiotic resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).
3. Falagas, M. E. & Kasiakou, S. K. Colistin: The revival of polymyxins for the management of multidrug-resistant gram-negative bacterial infections. *Clinical Infectious Diseases* **40**, 1333–1341 (2005).
4. Liu, Y. Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *Lancet Infect. Dis.* **16**, 161–168 (2015).
5. Gharaibeh, M. H. & Shatnawi, S. Q. An overview of colistin resistance, mobilized colistin resistance genes dissemination, global responses, and the alternatives to colistin: A review. *Veterinary World* **12**, 1735–1746 (2019).
6. Ling, Z. *et al.* Epidemiology of mobile colistin resistance genes mcr-1 to mcr-9. *J. Antimicrob. Chemother.* 1–9 (2020). doi:10.1093/jac/dkaa205
7. Cuadrat, R. R. C., Sorokina, M., Andrade, B. G., Goris, T. & Dávila, A. M. R. Global ocean resistome revealed: Exploring antibiotic resistance gene abundance and distribution in TARA Oceans samples. *Gigascience* **9**, 1–12 (2020).
8. Wang, R. *et al.* The global distribution and spread of the mobilized colistin resistance gene mcr-1. *Nat. Commun.* **9**, 1–9 (2018).
9. Luo, Q., Wang, Y. & Xiao, Y. Prevalence and transmission of mobilized colistin resistance (mcr) gene in bacteria common to animals and humans. *Biosaf. Heal.* **2**, 71–78 (2020).
10. Xavier, B. B., Lammens, C., Ruhai, R., Butaye, P. & Goossens, H. Identification of a novel plasmid-mediated colistin- resistance gene , mcr-2 , in Escherichia coli , Belgium , June. *Euro Surveill.* 6–11 (2016).
11. Yin, W. *et al.* Novel Plasmid-Mediated Colistin Resistance Gene mcr-3 in Escherichia coli. *MBio* **8**:e00543–1, (2017).
12. Carattoli, A. *et al.* Novel plasmid-mediated colistin resistance mcr-4 gene in Salmonella and Escherichia coli, Italy 2013, Spain and Belgium, 2015 to 2016. *Eurosurveillance* **22**, 1–5 (2017).
13. Borowiak, M. *et al.* Identification of a novel transposon-associated phosphoethanolamine transferase gene, mcr-5, conferring colistin resistance in d-tartrate fermenting Salmonella enterica subsp. enterica serovar Paratyphi B. *J. Antimicrob. Chemother.* **72**, 3317–3324 (2017).
14. AbuOun, M. *et al.* mcr-1 and mcr-2 variant genes identified in Moraxella species isolated from pigs in Great Britain from 2014 to 2015. *J. Antimicrob. Chemother.* **72**, 2745–2749 (2017).
15. Yang, Y. Q., Li, Y. X., Lei, C. W., Zhang, A. Y. & Wang, H. N. Novel plasmid-mediated colistin resistance gene mcr-7.1 in Klebsiella pneumoniae. *J. Antimicrob. Chemother.* **73**, 1791–1795 (2018).
16. Wang, X. *et al.* Emergence of a novel mobile colistin resistance gene, mcr-8, in NDM-producing Klebsiella pneumoniae article. *Emerg. Microbes Infect.* **7**, (2018).
17. Carroll, L. M. *et al.* Identification of novel mobilized colistin resistance gene mcr-9 in a multidrug-resistant, colistin-susceptible salmonella enterica serotype typhimurium isolate. *MBio* **10**, (2019).

- 477 18. Shen, Z., Wang, Y., Shen, Y., Shen, J. & Wu, C. Early emergence of mcr-1 in Escherichia coli
478 from food-producing animals. *Lancet Infect. Dis.* **16**, 293 (2016).
- 479 19. Skov, R. L. & Monnet, D. L. Plasmid-mediated colistin resistance (mcr-1 gene): Three months
480 later, the story unfolds. *Eurosurveillance* **21**, (2016).
- 481 20. Hendriksen, R. S. *et al.* Global monitoring of antimicrobial resistance based on
482 metagenomics analyses of urban sewage. *Nat. Commun.* **10**, (2019).
- 483 21. Brinch, C. *et al.* Long-Term Temporal Stability of the Resistome in Sewage from Copenhagen.
484 (2020).
- 485 22. Forsberg, K. J. *et al.* The shared antibiotic resistome of soil bacteria and human pathogens.
486 *Science (80-.).* **337**, 1107–1111 (2012).
- 487 23. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* **39**, 44–47 (2011).
- 488 24. Zankari, E. *et al.* Identification of acquired antimicrobial resistance genes. *J. Antimicrob.*
489 *Chemother.* **67**, 2640–2644 (2012).
- 490 25. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing
491 and web-based tools. *Nucleic Acids Res.* **41**, 590–596 (2013).
- 492 26. Bushnell, B. BBMap. (2014).
- 493 27. Clausen, P. T. L. C., Aarestrup, F. M. & Lund, O. Rapid and precise alignment of raw reads
494 against redundant databases with KMA. *BMC Bioinformatics* **19**, 1–8 (2018).
- 495 28. Aitchison, J. The Statistical Analysis of Compositional Data. *J. R. Stat. Soc. Ser. B* **44**, 139–160
496 (1982).
- 497 29. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G. & Gloor, G. B. ANOVA-Like Differential
498 Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLoS One* **8**, (2013).
- 499 30. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
- 500 31. Waskom, M. *et al.* mwaskom/seaborn: v0.11.0. (2020). doi:10.5281/zenodo.592845
- 501 32. Gillies, S. & Others, A. Shapely: manipulation and analysis of geometric objects. (2007).
- 502 33. Office, M. Cartopy: a cartographic python library with a Matplotlib interface. (2021).
- 503 34. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets:
504 Characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by
505 compositional data analysis. *Microbiome* **2**, 1–13 (2014).
- 506 35. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A practical and Powerful
507 Approach to Multiple Testing. *Journal of the Royal Statistical Society* **57**, 289–300 (1995).
- 508 36. Gloor, G. B., Macklaim, J. M. & Fernandes, A. D. Displaying Variation in Large Datasets:
509 Plotting a Visual Summary of Effect Sizes. *J. Comput. Graph. Stat.* **25**, 971–979 (2016).
- 510 37. Calle, M. L. Statistical Analysis of Metagenomics Data. *Genomics Inform.* **17**, e6 (2019).
- 511 38. The NCBI Pathogen Detection Project. *Bethesda (MD): National Library of Medicine (US),*
512 *National Center for Biotechnology Information.* (2016). Available at:
513 <https://www.ncbi.nlm.nih.gov/pathogens/>. (Accessed: 5th August 2021)
- 514 39. Feldgarden, M. *et al.* AMRFinderPlus and the Reference Gene Catalog facilitate examination
515 of the genomic links among antimicrobial resistance, stress response, and virulence. *Sci.*
516 *Rep.* **11**, (2021).
- 517 40. Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. MetaSPAdes: A new versatile
518 metagenomic assembler. *Genome Res.* **27**, 824–834 (2017).
- 519 41. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
520 (2009).
- 521 42. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic

features. *Bioinformatics* **26**, 841–842 (2010).

43. Carattoli, A. *et al.* In Silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).

44. Johansson, M. H. K. *et al.* Detection of mobile genetic elements associated with antibiotic resistance in *Salmonella enterica* using a newly developed web tool: MobileElementFinder. *J. Antimicrob. Chemother.* **76**, 101–109 (2021).

45. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* **58**, 236–244 (1963).

46. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

47. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

48. Matlock, W., Lipworth, S., Constantinides, B. & Peto, T. E. A. Flanker : a tool for comparative genomics of gene flanking regions. (2021).

49. Dalmolin, T. V, De Lima-Morales, D. & Barth, A. L. *Plasmid-mediated Colistin Resistance: What Do We Know? Journal of Infectiology Mini Review* **1**, (2018).

50. Jiang, Y. *et al.* Clinical relevance and plasmid dynamics of mcr-1-positive *Escherichia coli* in China: a multicentre case-control and molecular epidemiological study. *The Lancet Microbe* **1**, e24–e33 (2020).

51. Liu, Y. Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: A microbiological and molecular biological study. *Lancet Infect. Dis.* **16**, 161–168 (2016).

52. Anyanwu, M. U., Jaja, I. F. & Nwobi, O. C. Occurrence and characteristics of mobile colistin resistance (Mcr) gene-containing isolates from the environment: A review. *International Journal of Environmental Research and Public Health* **17**, 1028 (2020).

53. Khanawapee, A. *et al.* Distribution and Molecular Characterization of *Escherichia coli* Harboring mcr Genes Isolated from Slaughtered Pigs in Thailand . *Microb. Drug Resist.* (2020). doi:10.1089/mdr.2020.0242

54. Tyso, G. H. *et al.* The mcr-9 gene of salmonella and escherichia coli is not associated with colistin resistance in the United States. *Antimicrob. Agents Chemother.* **64**, (2020).

55. Aarestrup, F. M. & Woolhouse, M. E. J. Using sewage for surveillance of antimicrobial resistance. *Science* **367**, 630–632 (2020).

56. Wang, C. *et al.* Identification of novel mobile colistin resistance gene mcr-10. *Emerg. Microbes Infect.* **9**, 508–516 (2020).

Contributions

F.M.A. and T.N.P. designed the project. H.M.M. performed data acquisition, assembly, and analysis. J.S. and T.N.P. provided guidance with sequence downloading and mapping, P.M. and C.B. with compositional analysis and P.M. with flank analysis. H.M.M. wrote the first draft. All authors have read, commented, and accepted the final version.

Acknowledgements

This work was supported by the European Union's Horizon H2020 grant VEO (874735) and the Novo Nordisk Foundation (grant NNF16OC0021856: Global Surveillance of Antimicrobial Resistance).

Figure S1: Sample origins of metagenomes with mcr genes and the number of metagenomes without in parentheses.

Figure S2: Distribution of *mcr* genes in all sampling hosts.

Figure S3: Biplot of human samples from PCA on the full dataset colored by **a.** country and **b.** sampling year.

Figure S4: Biplots produced by doing compositional PCA on read counts of mcr-2 to mcr-8, where samples are colored by **a.** sampling year, **b.** sampling location and **c.** sampling host.

	<i>mcr-1</i>	<i>mcr-2</i>	<i>mcr-3</i>	<i>mcr-4</i>	<i>mcr-5</i>	<i>mcr-6</i>	<i>mcr-7</i>	<i>mcr-8</i>	<i>mcr-9</i>
Isolate frequency (%)	51.08	0.25	6.83	0.74	1.26	0.00	0.00	1.18	40.38

Table S1: Prevalence of *mcr* gene variants in single isolates screened by the NCBI Pathogen Detection pipeline.

Figure S5: Pairwise comparison of *mcr* gene prevalence in single isolates and metagenomes for common sample origins, where the prevalence is higher than zero for both measurements. The relative count is the percentage of *mcr* positive isolates carrying a variant and the relative abundance is the frequency of metagenomic *mcr* reads aligned to a variant. The closer to the grey line a marker is, the more agreement there is between the two measurements.

588

589

590

591

Figure S6: a. Histogram of contig lengths and which mcr gene a contig carries. b. Count of mobile elements or plasmids occurring in flanks of different sizes.

592

Figure S7: Count of the occurrence of mobile genetic elements (left) and plasmids (right) in *mcr* carrying contigs with flank sizes between 1000 and 21,000 bp. The number does not consider if two elements are on the same contig.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementarydata.tar.gz](#)