

Reviewing the phylogenetic relationships and nomenclature of common antimicrobial resistance genes

This manuscript ([permalink](#)) was automatically generated from jackiepurdue/phylogenetic-amr-survey-manuscript@119127d on May 30, 2020.

Authors

- **Jackie Purdue**
ID [0000-0003-3131-0688](#) ·  [jackiepurdue](#) ·  [jackiepurdue](#)
Faculty of Computer Science, Dalhousie University
- **Jocelyn McDonald**
Faculty of Computer Science, Dalhousie University
- **Dayna Mikkelsen**
Faculty of Computer Science, Dalhousie University
- **Finlay Maguire**
ID [0000-0002-1203-9514](#) ·  [fmaguire](#) ·  [finlaym](#)
Faculty of Computer Science, Dalhousie University · Funded by Donald Hill Family Fellowship in Computer Science
- **Robert G. Beiko**
Faculty of Computer Science, Dalhousie University

Abstract

Antimicrobial resistance (AMR) is a global threat to human health. Efforts to prevent the spread of AMR rely on surveillance of possible AMR determinants. Using phylogenetics to highlight potential sources of AMR could guide researchers in choosing organisms for phenotypic resistance testing. A systematic way of classifying AMR gene variants is important in comparing these phylogenetic relationships. In this study, the phylogenetic neighborhoods of several named AMR genes are examined for diversity, spread, and consistency in classification. Canonical sequence data from the Comprehensive Antibiotic Resistance Database (CARD) for mobile colistin resistance (MCR), New Delhi beta-lactamase (NDM), Klebsiella pneumoniae carbapenemase (KPC), and OXA beta-lactamase were used to query CARD prevalence data, NCBI sequence data, and draft quality metagenome assembled genomes (MAGs) from various uncultivated bacterial and archeal sources (UBAs). A phylogeny of this expanded set of candidate resistance genes was produced with a limited amount of new diversity, most notably with genes which were potentially associated with MCR. In this phylogenetic context, some of the names of the genes were found to be inconsistent or inconsequential.

Introduction

Antimicrobials, substances which can kill or inhibit the growth of microbes [1], are key effectors in the ecology of microorganisms. The evolution of resistance to these substances, known as antimicrobial resistance (AMR), is an important and ancient adaptive process [2]. In the last century we have adopted and used antimicrobials to great effect in clinical medicine and agriculture. However, the use and abuse of antimicrobials has led to increasing levels of observed AMR [3]. This poses a growing global health risk by undermining our ability to treat infectious diseases and perform surgeries [4]. Infections by resistant microbes have greater mortality and morbidity [4] with the European Centre for Disease Prevention and Control estimating that 25,000 people per year die due to AMR [5].

AMR can be due to intrinsic resistance within the organism or the acquisition of resistance via lateral gene transfer [2]. These mechanisms allow for the constant change of the resistome, making determining the scope of AMR within both pathogenic and non-pathogenic bacteria difficult [3]. To mitigate the spread of AMR the World Health Organization (WHO) created an action plan which emphasizes the need to strengthen our understanding and surveillance of AMR. An important aspect of this research is to examine the spread of AMR between environmental and clinical samples (<https://www.who.int/antimicrobial-resistance/global-action-plan/en/>). One way to try to understand the evolution and spread of AMR is to perform large-scale evolutionary analyses with samples from lots of different environments/contexts.

Some of the most critical pathogens are the multi-drug resistant "ESKAPE" pathogens: *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Enterobacter* spp. [6]. These are the leading cause of hospital-acquired infections and are largely only treatable with carbapenem or polymyxin antibiotics [5]. Carbapenems are broad spectrum beta-lactam antibiotics typically used to treat life threatening, high risk, and multi-drug resistant bacterial infections. beta-lactamase producing bacteria have difficulty degrading carbapenems when combined with beta-lactamase inhibitors making them effective treatments against pathogens with multi-drug resistance [5]. However, there has been a recent emergence of resistance to these antimicrobials, presenting a new threat to public health. One way microbes can develop carbapenem resistance is through the acquisition of carbapenamase genes. [5]. Three main carbapenamase genes of concern are *Klebsiella pneumoniae* carbapenamase (KPC), New Delhi metallo-beta-lactamase (NDM) and OXA beta-lactamases [5].

Carbapenamases are classified as one of four classes (A-D) based on their molecular structure [7]. Class A, C, and D both use a serine residue to hydrolyze beta-lactam antibiotics [8]. KPC genes are a part of class A. Many variants of KPC are spreading and being discovered, however carbapenamases within this class are more rare than the other three classes [5]. Class D contains the OXA family of beta-lactamases. Unlike KPC and NDM, this family is characterized by phenotype rather than genotype. This family is characterized by specific hydrolases that can hydrolyze oxacillin and cloxacillin, with the enzyme being poorly inhibited by clavulanic acid (<https://card.mcmaster.ca/ontology/36026>). This results in a low amount of sequence homology within the family. A subset of this family, OXA-48 possesses carbapenem hydrolyzing activity [9]. The OXA-48 subfamily consists of five variants: OXA-48, OXA-162, OXA-163, OXA-181, OXA-204, and OXA-232 [10]. Class B beta-lactamases are metallo-enzymes that use a zinc active site to hydrolyze beta-lactam antibiotics. The NDM gene is classified as a class B beta-lactamase [5]. When pathogens become resistant to carbapenems, the last line of defense is the polymyxin antibiotic colistin. Polymyxins act via disrupting membrane permeability [11]. Colistin is limited to a last line defense against carbapenem resistant infections due to its neurotoxic and nephrotoxic effects [11]. Despite its already limited use, mobilized colistin resistance (MCR) have been discovered [12]. Due to the emergence of resistance to many last-line antibiotics, it is important to characterize the total diversity and spread of these key resistance determinants. However, to perform these types of analyses we need two resources: a clear consolidated definition of what constitutes an AMR-related gene and a large amount of sequencing microbial genomes.

Naming and defining AMR genes is difficult due to the legacy of research pre-DNA sequencing technology and the large number of different researchers and stakeholders involved in AMR. AMR determinants are typically classified into AMR families, some of which are based on phenotypic properties, and some based on sequence variation. The nomenclature is usually an acronym representing the mechanism of resistance, along with a numerical value to distinguish variants. Each sequence, determined to be novel by some arbitrary criteria, is assigned a new number. For example, the MCR phosphoethanolamine transferase gene family is a family of genes which code for a hinderance of the ability of colistin to bind to the cell membrane. The MCR-3 gene has a sequence identity which has been deemed to be significantly different than that of the other MCR genes in the family. MCR-3.1, is has been deemed to be a variant of the MCR-3 gene, which has a slight variation in gene sequence from other MCR-3 genes. In looking at another AMR gene family, NDM beta-lactamase, a different set of criteria are used to name variants. NDM betalactamases are classified as such by their ability to confer resistance to carbapenems and some other antibiotics. In this case, NDM-1, and NDM-2 are only different by a single amino acid substitution. This system for classification has the potential to be misleading when conducting AMR research. As in the examples, AMR families could appear to have a large amount of diversity, when in reality, sequences are closely related, and only a small number are actually distinct. Additionally, sequences which are not homologues, could potentially be classified in the same family, simply based on their function. An example of this is the OXA betalactamase family, which is a phenotypically characterized family, and there are examples of determinants which have a low amount of sequence homology. OXA-20 has less than 20% amino acid sequence identity when compared with several other OXA family betalactamases [13]. This is a problem when attempting to characterize the AMR determinants by sequence similarity.

High-quality manually curated AMR databases such as the Comprehensive Antibiotic Resistance Database (CARD) [14] provide a unified resource for the definition, nomenclature and classification of AMR genes. CARD organizes this information CARD through the antibiotic resistance ontology (ARO), a controlled vocabulary with defined relationships. There are two data-sets within CARD, canonical and prevalence. Canonical is a conservative set of AMR genes and mutations that have been experimentally verified as being associated with resistance in a peer-reviewed publication. This is therefore slightly biased towards the organisms that have been most heavily studied. On the other hand, prevalence contains AMR sequences that have been detected by searching a broader set of WHO priority organism genomes from repositories such as National Center for Biotechnology Information (NCBI) for sequences similar to canonical sequences. This database is meant to represent the diversity of all current sequence data available to the public. This *in silico* search is performed using CARD's BLAST-based Resistance Gene Identifier (RGI) tool and greatly increases the sequence diversity in CARD [14].

However, this is still limited to a subset of all genomes currently available in central repositories like NCBI. If we want to thoroughly understand the evolution and spread of AMR genes we need to analyses as many genomes as possible. Unfortunately, the genomes in databases are largely sequenced from microbes that can be easily . As only a subset of microbial diversity can be cultured, this means many of our existing genomes aren't necessarily

representative of the environment from which they were sampled. Recently, techniques have been developed that allow the recovery of genomes from metagenomic data e.g. [15]. As metagenomic sequencing, the direct sequencing of all DNA in a sample, doesn't require culturing these metagenome-assembled genomes (MAGs) represent a huge source of novel microbial diversity. Parks et. al. [17], generated new 7,903 bacterial and archaeal (Uncultured Bacteria and Archaea; UBA) MAGs representing >30% increase in the sampled phylogenetic diversity of the bacteria and archaea.

Identifying and phylogenetically analyzing key carbapenemase and colistin resistance genes in this dataset, CARD, CARD-prevalence and NCBI genomes, could greatly improve AMR surveillance of these genes. We would characterize previously unseen diversity in genomes not yet analyzed for AMR, and provide insights into the diversity of AMR across non-clinical samples. This could inform our understanding of the transmission of these mobile critical AMR genes and help refine the nomenclature of this ever-expanding dataset. Therefore, in this work we present a comprehensive phylogenetic survey and analysis of KPC, NDM, OXA-48, and MCR across all currently sequenced genomes.

Methods

Data

To sample diverse sources for AMR, various sequence databases were sampled for phylogenetic analyses. These sources include CARD canonical and prevalence sequences for the AMR genes under study, and associated homologs from NCBI non-redundant data and metagenomic data. The 32 canonical MCR sequence variants (table 1), 14 canonical NDM sequence variants (table 3), 6 canonical OXA-48-like sequence variants (table 4), and 18 canonical KPC sequence variants (table 2) as labeled in the CARD database, were obtained as a reference. The AMR prevalence data was queried from CARD Prevalence 3.0.5. The NCBI non-redundant data was queried from all non-redundant GenBank, PDB, SwissProt, PDB, PIR, and PRF on May 17, 2019. The metagenomic data came from a data-set of 7903 draft quality MAGs which were recovered from the Sequence Read Archive by Parks [18]. These genomes were chosen specifically because they were likely to be from lineages which were under-sampled, environmental and non-human gastrointestinal samples being the main focus.

In the case of OXA beta-lactamase, only OXA-48-like genes were used for analysis. The OXA family is characterized by phenotype rather than genotype, and results in a low amount of sequence homology within the family. The phenotype of OXA-48 results from carbapenem hydrolyzing activity [9]. This subfamily of OXA contains homologous sequences suitable for this study. Incorporating other subfamilies of OXA proved to be too cumbersome.

Querying the data

The CARD canonical sequences from each family were used to perform a multiple query BLASTP (version 2.5.0 [19]) against the prevalence BLAST database with a e-value threshold and query coverages shown in Table 1. Many of the sequences are nearly identical, thus they were further processed by clustering with CD-HIT version 4.8.1 at a minimum sequence identities as per table {[??]}.

The reference CARD canonical sequences were also used to perform a multiple query BLASTP against the NCBI non-redundant database with a e-value threshold, and query coverage, also indicated in Table 1. There are many highly sampled taxa and genes in the non-redundant database. To balance the distribution, and reduce the size of the non-redundant sequence set, CD-HIT version 4.8.1 was also used to cluster the data as per table {[??]}.

For the metagenomic data, RGI version 5.0 with CARD database version 3.02 [3] was run on the contigs of the 7903 MAGS with the inclusion of loose, perfect, and strict hits. Sequences possibly containing AMR gene prediction data for each gene family was produced by filtering RGI output based on the its association with the search strings for each determinant in Table 1. The filtered data were translated to a blast database. The CARD canonical sequences were used to perform a multiple query BLASTP against this UBA blast database with a e-value threshold and query coverage in Table 1. The E-values were chosen such that a tractable or sufficient number of sequences resulted from their respective database query. The minimum BLAST query coverages were chosen as to reduce noise in the final alignment, thus reducing the possibility of attempting to align non-homologous sequences.

For all sequence variants obtained from each data source, redundant results for the prevalence, NCBI, and UBA queries were filtered from these BLASTP results by retrieving only the longest sequence for each uniquely labeled result.

AMR Determinant	Minimum BLAST E-value			Minimum BLAST Query Coverage			CD-HIT Sequence Identity Threshold		
	CARD	NCBI	Parks	CARD	NCBI	Parks	CARD	NCBI	Parks
MCR	1e-160	1e-160	1e-160	98	98	98	100	100	100
KPC	1e-100	1e-40	1e-10	99	99	60	99	70	100
NDM	1e-160	1e-160	1e-10	98	98	60	100	100	100
OXA	1e-100	1e-100	1e-10	90	99	95	N/A	N/A	N/A

Table 1: Parameters used with BLAST and CD-HIT to obtain sequences for the phylogenies of MCR phosphoethanolamine transferase, and KPC, NDM, and OXA beta-lactamases.
For each determinant, the minimum BLAST E-values, minimum query coverages, and CD-HIT Sequence Identity threshold for clustering sequences are displayed for each database.
CARD represents the RGI-CARD prevalence sequence data, NCBI represents the NCBI non-redundant sequence data, and Parks represents the MAG data source.

Sequence Alignment

Two alignments were created for each AMR determinant under study. The first alignment was created to compare the phylogenetic relationship of only the putative sequences. The second alignment was created for an overall comparison of sequences.

The putative AMR alignment was made up of the sequences from the CARD prevalence data were concatenated in one multi-FASTA format file with the canonical sequences and an outgroup chosen for each AMR family as per table 5. In the second, more diverse alignment, The filtered sequences from NCBI non-redundant data, CARD prevalence data, UBA data were all concatenated to one multi-FASTA format file with the canonical sequences with the same outgroup sequences (table 5).

This set of concatenated amino acid sequences were aligned with MAFFT-LINSI version 7.40 and trimmed by trimal version 1.4.rev22 using the automated1 option.

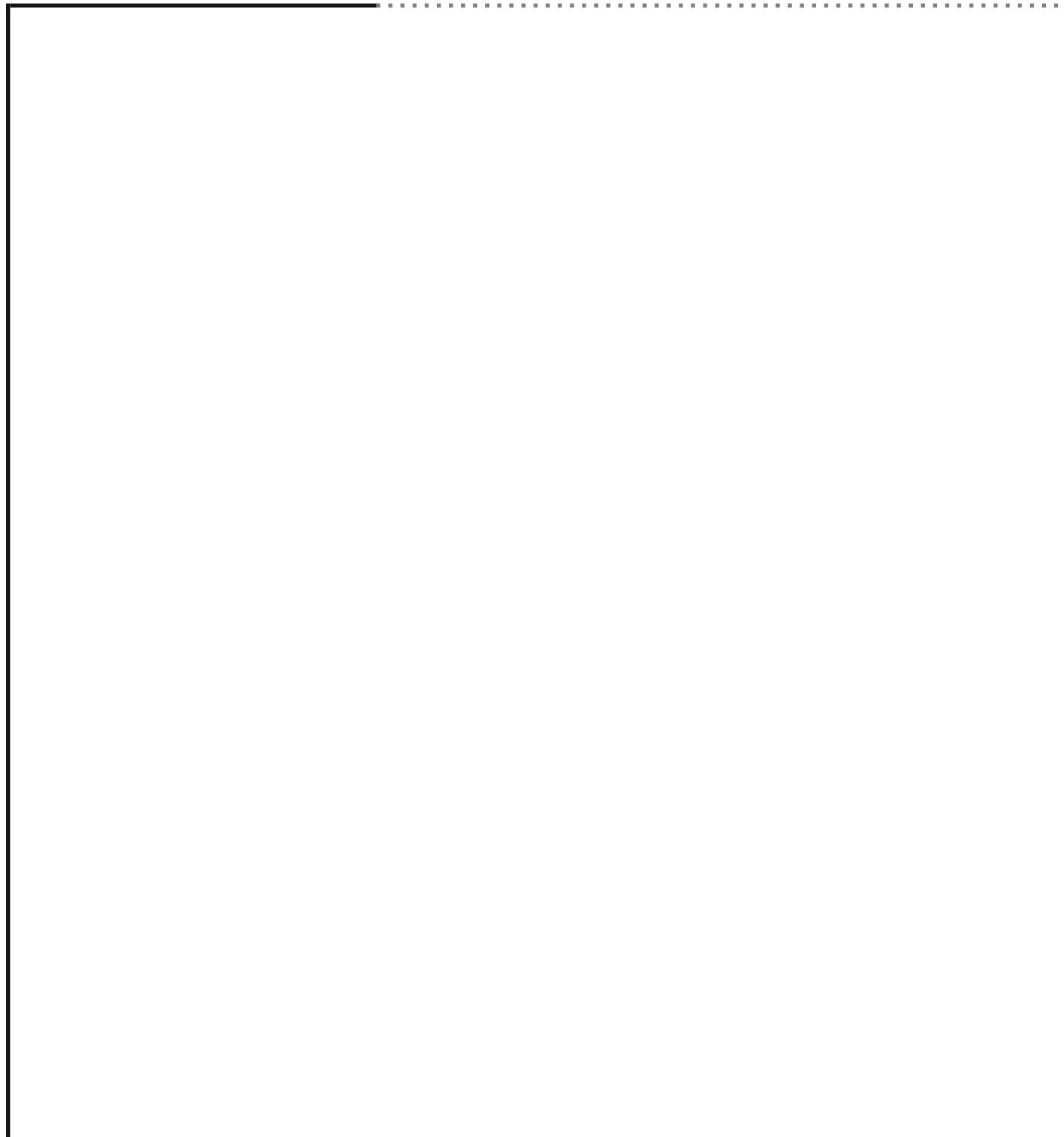
Creation of phylogenies

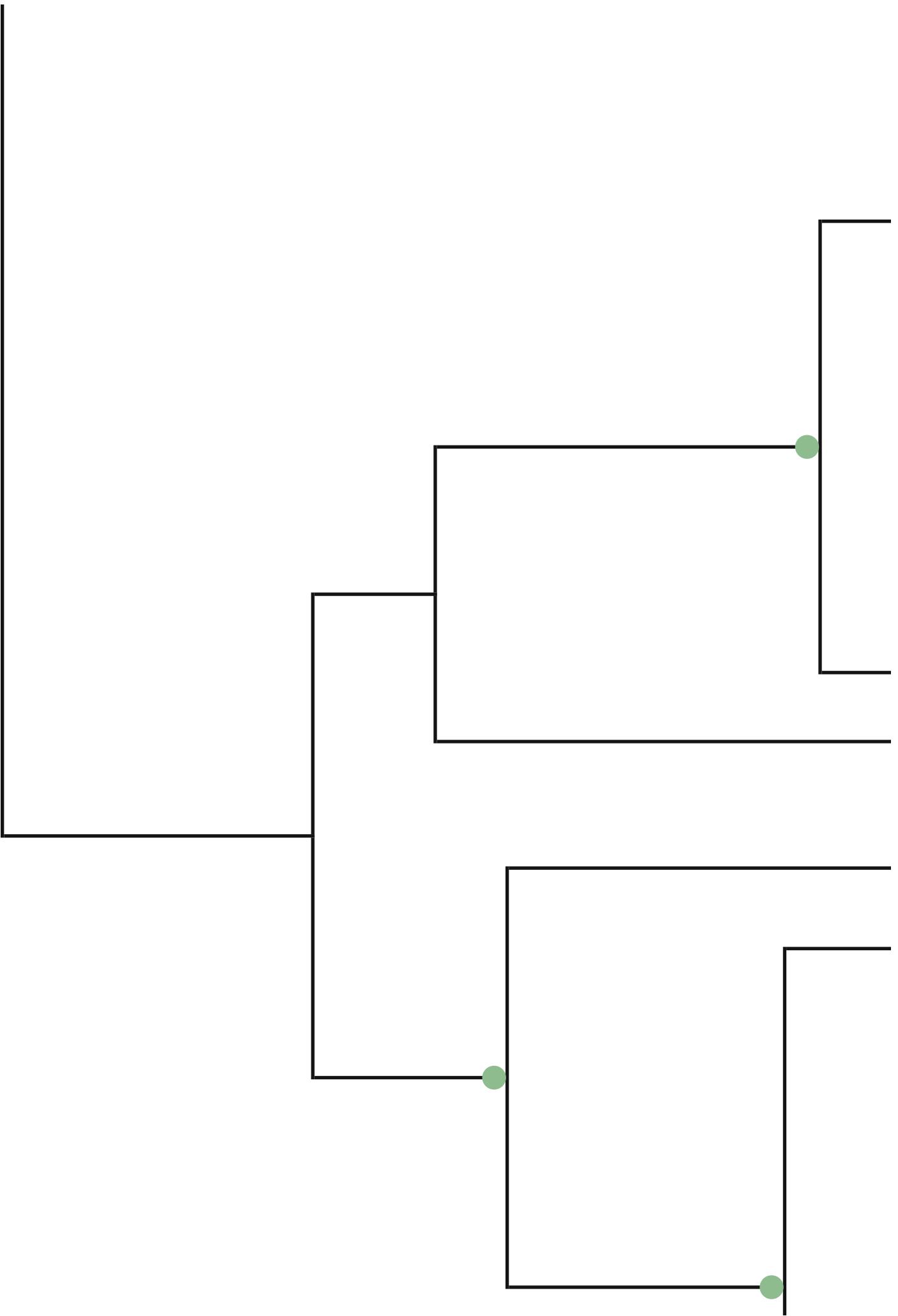
For each alignment under each AMR family under study, IQ-TREE multicore version 1.6.9 [20](#) was used to build a bootstrapped tree with -bb 1000 with the G+I+G4 model of substitution. Tree visualizations were created with ETE Toolkit version 3. The trees were annotated with taxonomic information for each rank by using ETE's NCBI taxonomy module to translate the names result set from the various databases to taxonomic identifiers. The taxonomic identifier was used to query NCBI's taxonomy database for the taxonomy information. In cases where there were multiple hits for genes, the most recent common ancestor was used to query the taxonomic information, therefore some leaves in the trees were annotated with only partial taxonomic information.

Results

Phylogenetic analysis of MCR sequences

A phylogeny with only putative MCR gene family sequences was created, showing the relationships without additional diversity. An out-group, 32 canonical sequences, and 54 prevalence sequences were selected for this phylogeny. The resulting tree in Figure {[??]} shows MCR family members 1 through 9 forming distinct sub-clades in the tree. MCR-1, MCR-2, and MCR-6 form a closely related clade. This clade is also closely related to the ICR-Mc clade. ICR-Mc [21](#) is another phosphoethanolamine transferase which confers colistin resistance. MCR-3, and MCR-7, also form a closely related clade.





0.05

The relationships were further condensed with a representative for each MCR family member (figure 1). The overall diversity gradient becomes more apparent in this tree. Regions of the tree vary in their potential to accommodate additional diversity. Regions containing more evolutionary distance are accommodating to additional diversity, while closely related regions would not.

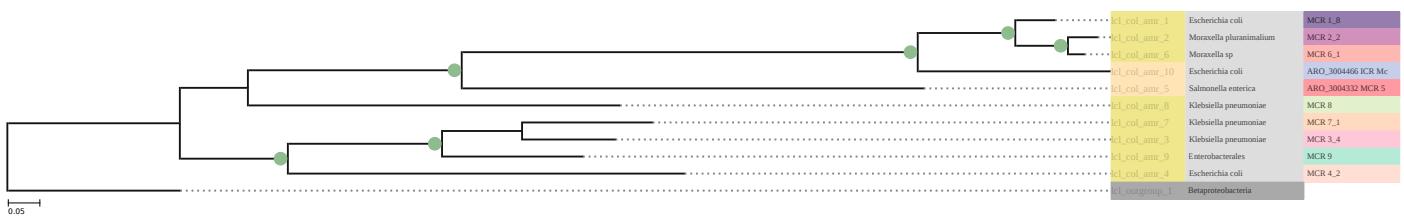


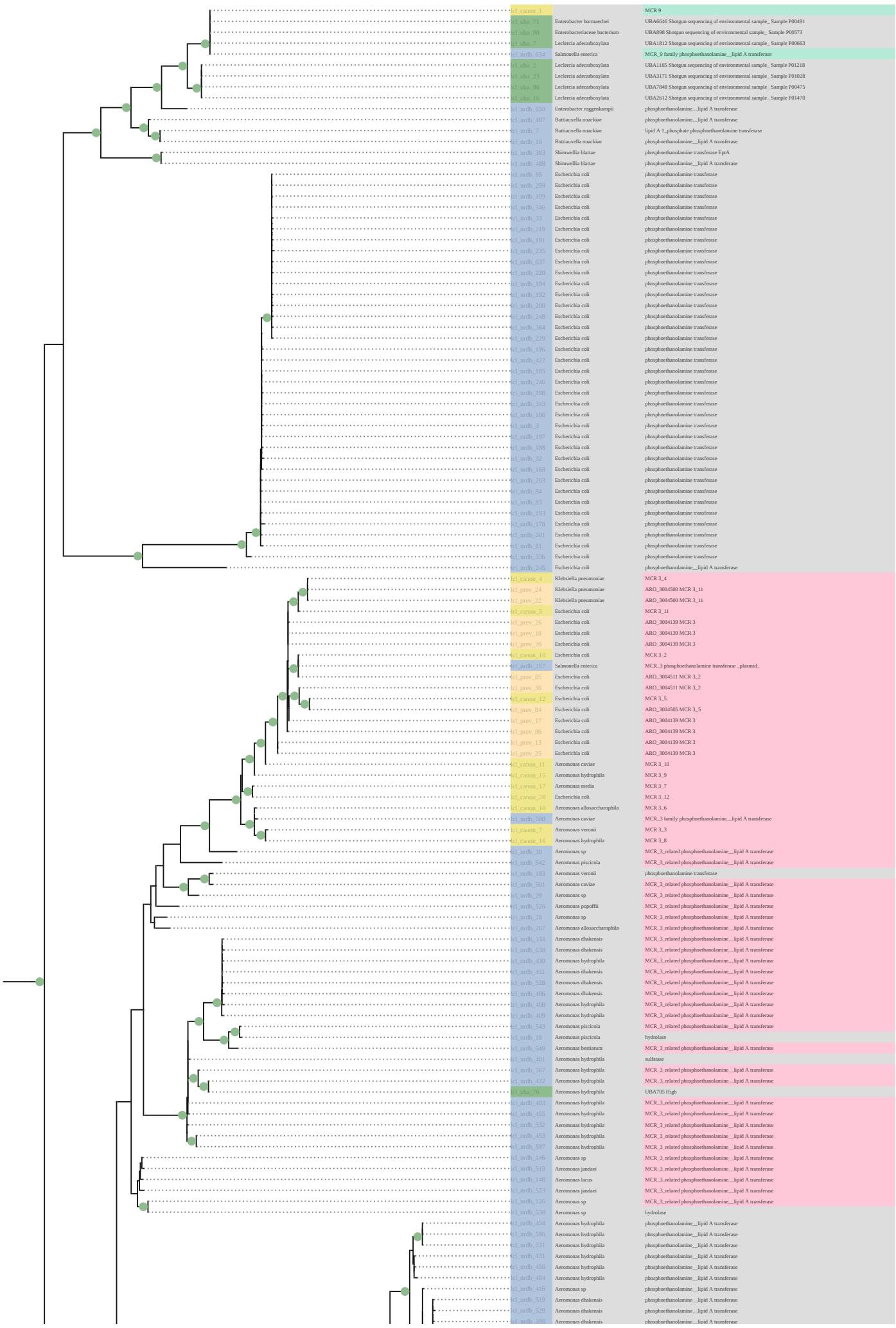
Figure 1: Phylogenetic relationship of 9 representative MCR family sequences, and an outgroup from Betaproteobacteria. (labels prefixed with 'lcl_canon' in yellow are canonical, labels prefixed with 'lcl_prev' in tan are prevalence, and an outgroup from Betaproteobacteria is prefixed with 'lcl_outgroup' in grey)

The phylogenies represented by Figures 2, 3, and 7 were created with additional diversity by querying the NCBI non-redundant database, and the MAGs from the Parks data 18. The filtering criteria in Table 1 were applied to these data sets.

After filtering, 409 NCBI sequence hits, labeled as phosphoethanolamine lipid A transferase, and 104 NCBI sequence hits, labeled as MCR family, were selected for addition to the phylogeny. The 7903 draft quality MAGs were queried for AMR genes with RGI. RGI produced 1457246 loose hits, 7171 strict hits, and 310 perfect hits. After applying filters, only 63 MAG sourced genes were retained for the phylogeny. Due to the large size, the phylogeny displayed in Figure 7 was subdivided into Figures 2, and 3. The subdivisions were chosen based on a pruning at the point of the most recent common ancestor of sets of canonical MCR labeled leaves. The set including MCR-3, MCR-7, and MCR-9, and the set including MCR-1, MCR-2, MCR-5, MCR-6, and ICR-Mc, contained several clades of diversity between some of the canonical and prevalence leaves. Other choices for leaf sets contained added diversity that was distant from canonical sequences, and were not chosen as a focused figures.

Between the clade containing MCR-3, and the most recent common ancestor of MCR-3 and MCR-7 clades (figure 2), there is a clade of sequences from NCBI which have been reported as MCR-3. Present within this clade is a single result from the Parks dataset 18, UBA705. This leaf is labelled as *Aeromonas hydrophilia*, and is associated with rice paddy field-soil microbial fuel cells. This is present within the clade alone with several other canonical and non-redundant Aeromonadalacea. *Aeromonas hydrophilia* is a species which has been found to have an MCR-3 gene.

Present between the clade containing MCR 7, and the most recent common ancestor of MCR 3 and MCR 7 clades, is a clade of NCBI sourced phosphoethanolamine lipid A transferase. This clade consists of the genus of gram negative bacteria, *Aeromonas*, 22 which is sometimes involved in human infection. This clade also includes an *Aeromonas veronii* hit from epidermal mucus of *Anguilla anguilla* in the Parks data. Between MCR 9 and the most recent common ancestor of MCR 3 and 9, a clade of *Aeromonas* associated phosphoethanolamine lipid A transferases appear. The MCR 9 containing clade contains 8 UBAs and several NCBI non-redundant hits not reported as MCR family genes. 5 Loose hits for MCR in a *Leclercia adecarboxylata* branch within this clade. These *Leclercia* branch below the common ancestor of MCR-9 and MCR-3 which are well supported. The *Leclercia* UBAs were all sampled from New York City MTA subway samples Metagenome. *Acinetobacter* is another opportunistic pathogen 23 which is becoming resistant to many antimicrobials.



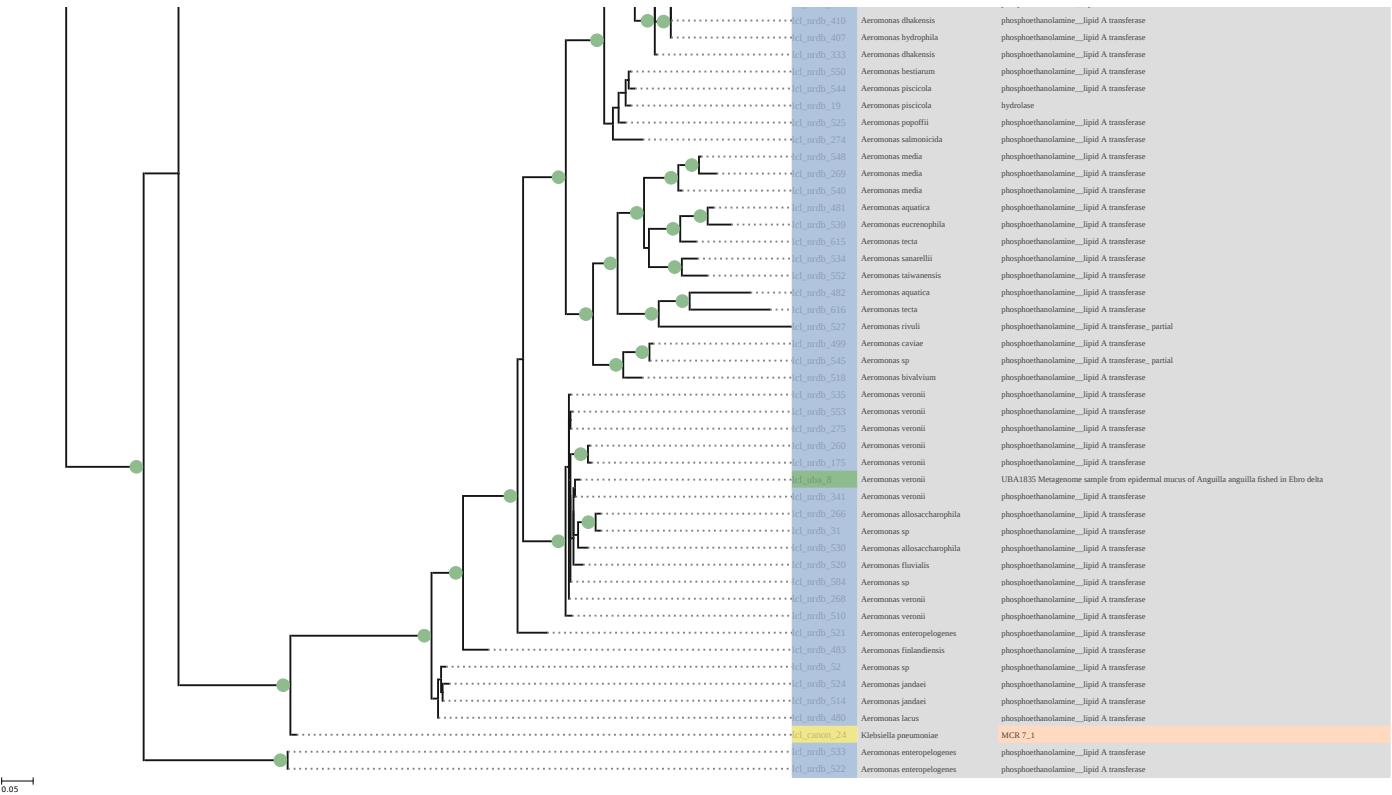
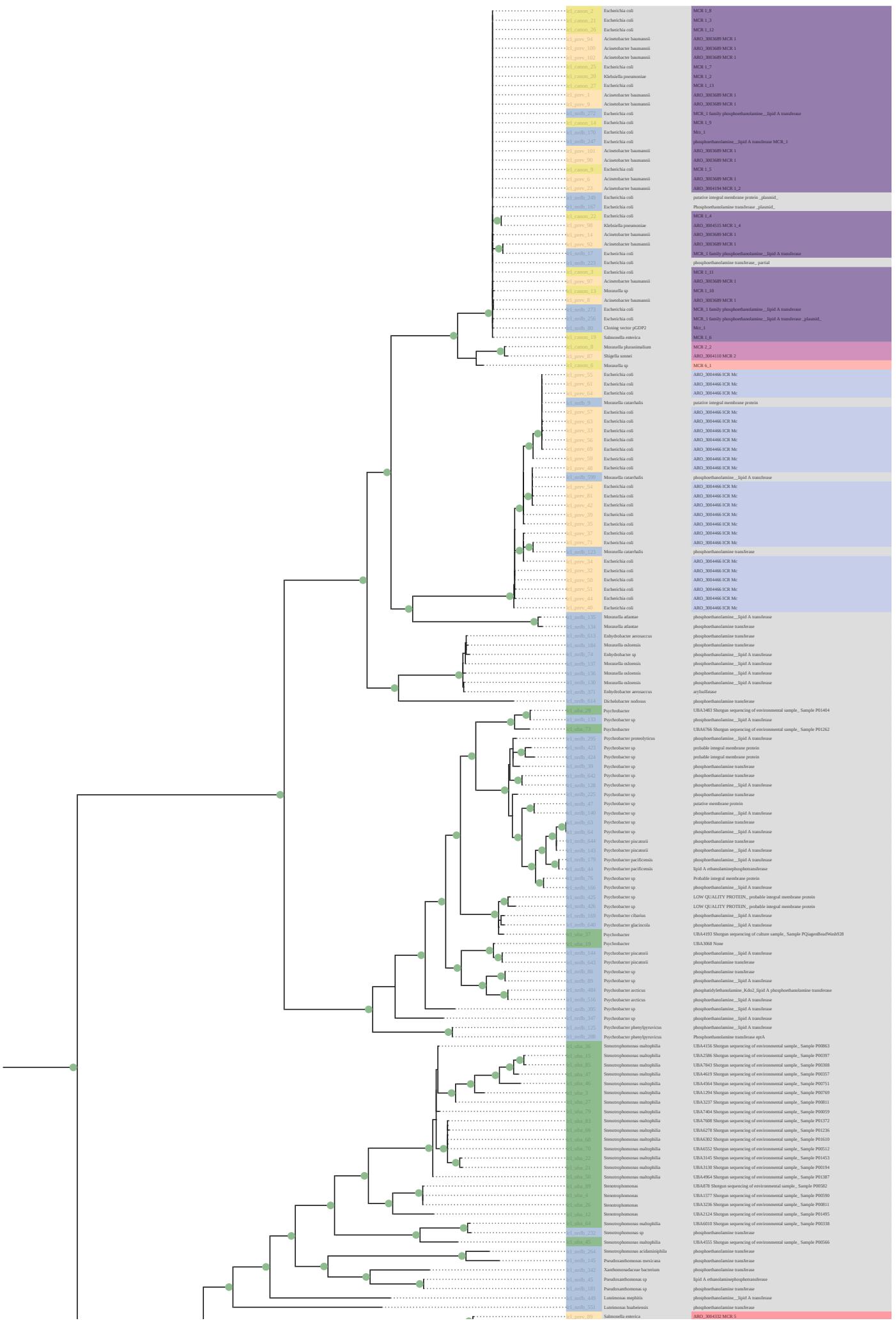


Figure 2: Clade containing putative MCR3 and MCR 9 clades pruned from Phylogenetic relationship of 32 canonical (labels prefixed with 'cl_canon'_ in yellow), and 54 prevalence (labels prefixed with 'cl_prev' in tan) MCR family sequences, 595 NCBI non-redundant sequences (labels prefixed with 'cl_nrdb' in blue), and 91 UBA sourced sequences (labels prefixed with 'cl_uba' in green). Each MCR variant is coloured based on its primary numerical value. If the sequence is not reported to be MCR family it is coloured in grey.

Between MCR 5 and ICR-Mc (figure {#fig:mcr-5-icr}), there appear clades of diversity in the genus *Psychrobacter*. Branching between these two clades, as a descendant to this ancestor, is a clade of *Psychrobacter* species bacteria. This clade includes several hits from the non-redundant database, and two hits from the UBA data. According to the Parks 18 data, the identity of these samples are, UBA3068, A *Psychrobacter* sp., sampled from Oil polluted marine microbial communities and, UBA4193, a *Psychrobacter* sp., sampled from the New York City subway. The quality information for these sequences, shown in Table ??, shows that UBA3068 is near complete, while UBA4193 is only partial. It is encouraging to see that even with the quality difference, these two sequences branch in the expected clade. *Psychrobacter* 24 is a Genus is widespread and includes many cold adapted bacteria, it is an opportunistic pathogen, and has been found to sometimes be a cause of infections in humans, animals, and fish. Many new species of *Psychrobacter* have been discovered in cold climates 25. Some of the species have been shown to be resistant to colistin, like *Psychrobacter vallis* ps. nov. and *Psychrobacter aquaticus* ps. nov. 26 and is sister to [TODO: display new tree such that *Acinetobacter* is shown] Another clade between these two variants contains phosphoethanolamine transferases from the genus *Stenotrophomonas*, including *Stenotrophomonas maltophilia* and *Stenotrophomonas acidaminiphila* which is a multidrug-resistant opportunistic pathogen 27. Several UBA hits for *Stenotrophomonas maltophilia* show up as shotgun sequencing of environmental samples. [TODO: quality information]



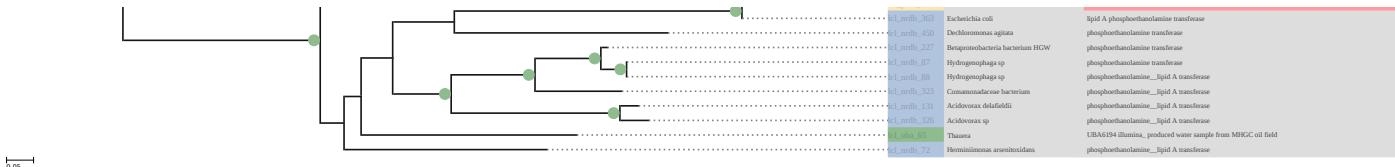


Figure 3: Clade containing putative MCR 5 and ICR-Mc clades pruned from Phylogenetic relationship of 32 canonical labels (prefixed with 'lcl_canon'_ in yellow), and 54 prevalence (labels prefixed with 'lcl_prev' in tan) MCR family sequences, 595 NCBI non-redundant sequences (labels prefixed with 'lcl_nrdb' in blue), and 91 UBA sourced sequences (labels prefixed with 'lcl_uba' in green). Each MCR variant is coloured based on its primary numerical value. If the sequence is not reported to be MCR family it is coloured in grey.

[TODO: visualization with plasmid distribution]

Phylogenetic analysis of KPC

Using the 18 canonical sequences from Table 2 as query sequences, 25 prevalence sequences, 376 non-redundant sequences, and 6 UBA sequences were recovered from the various databases. The result was combined in the phylogeny displayed in figure 11. This phylogeny shows that many of the resulting sequences are distant from the canonical sequences. The main clade in figure 4 shows that many of the amr determinants are named as distinct sequences, while not having much phylogenetic distance between them. The most closely related clade is a clade containing IMI and SME-type beta lactamases. The remaining sequences are in a clade containing a large amount of class A beta-lactamases from the non-redundant data. All of the UBA sequences were present in this portion of the tree.

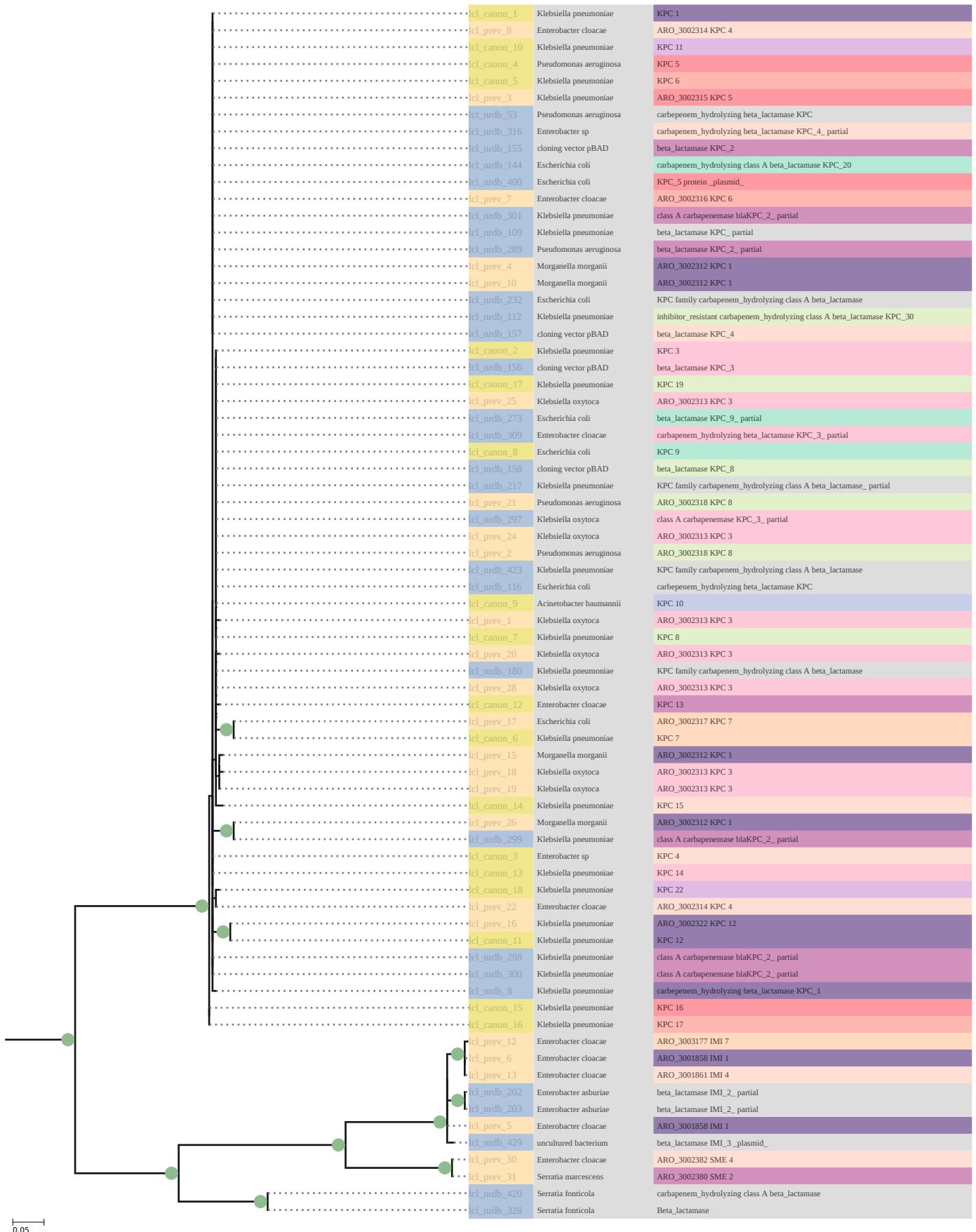


Figure 4: A clade pruned from the phylogeny in figure {#fig:kpc-tree} which represents the most closely related genes to the KPC family genes. 18 canonical labels (prefixed with 'cl_canon'_ in yellow), and 25 prevalence (labels prefixed with 'cl_prev' in tan), and 28 NCBI non-redundant sequences (labels prefixed with 'cl_nrdb' in blue). Each KPC variant is coloured based on its primary numerical value. If the sequence is not reported to be KPC family it is coloured in grey.

Phylogenetic analysis of NDM

14 canonical sequences were used to query the data for NDM beta-lactamase. The genes retained for the phylogeny were 14 canonical sequences, 8 prevalence sequences, 12 non-redundant sequences, 1 UBA sequence, and an out-group. This resulted in the phylogenetic relationship in Figure 10.

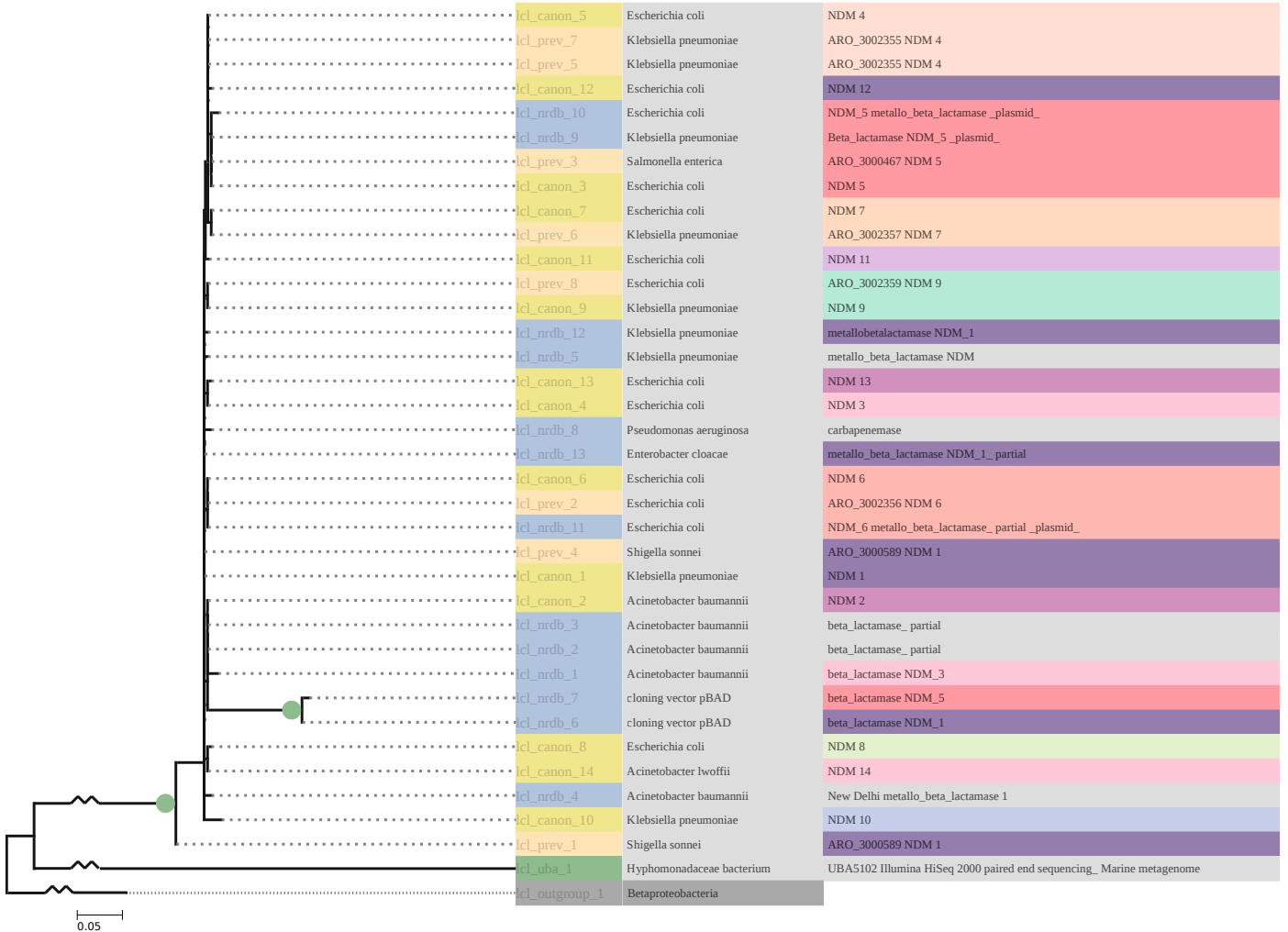


Figure 5: Phylogenetic relationship of 14 canonical (labels prefixed with 'icl_canon' in yellow), 8 prevalence (labels prefixed with 'icl_prev' in tan), 12 NCBI non-redundant database sequences (labeled as 'icl_nrdb' in blue), 1 UBA sequence (labeled as 'icl_uba' in green), and an outgroup from Betaproteobacteria (labeled as 'icl_outgroup' in grey). Each NDM variant is coloured based on its primary numerical value.

The 12 sequences resulting from the NCBI data included 3 genes reported to be NDM-1, NDM-3, and NDM-5, while the other 4 genes were reported to be general beta lactamase/carbapenemase hits. As in the KPC phylogeny (figure 4), there is very little phylogenetic difference between the named variants of NDM. The only phylogenetically distinct genes in this clade were the prevalence hit for NDM-1, and the ncbi hits labeled as NDM-1 and NDM-5. The naming of these sequences seem not to be congruent with the phylogenetic relationships present in the tree. Some of the NDM-1 and NDM-5 genes are more closely related to the other NDM variants than themselves. One BLAST result from the Parks data branches far from the clade containing the canonical sequences. This indicates that under the coverage queried, there are no reasonably detectable NDM homologues in the UBA data. The alignment of this UBA under this relaxed query coverage of 60% (see Table 1) is questionable. Reducing query coverage further could produce more sequences from the Parks data, but the resulting alignment would produce a meaningless phylogeny.

Phylogenetic analysis of OXA-48

In querying for the OXA-48 portion of the analysis, multiple BLAST results for UBA sequences were found, but the hits were too phylogenetically dissimilar to draw a conjecture about their relationship to the OXA family. 6 canonical sequences, and the 14 prevalence sequences, and 20 non-redundant sequences are combined in the phylogeny in Figure 8

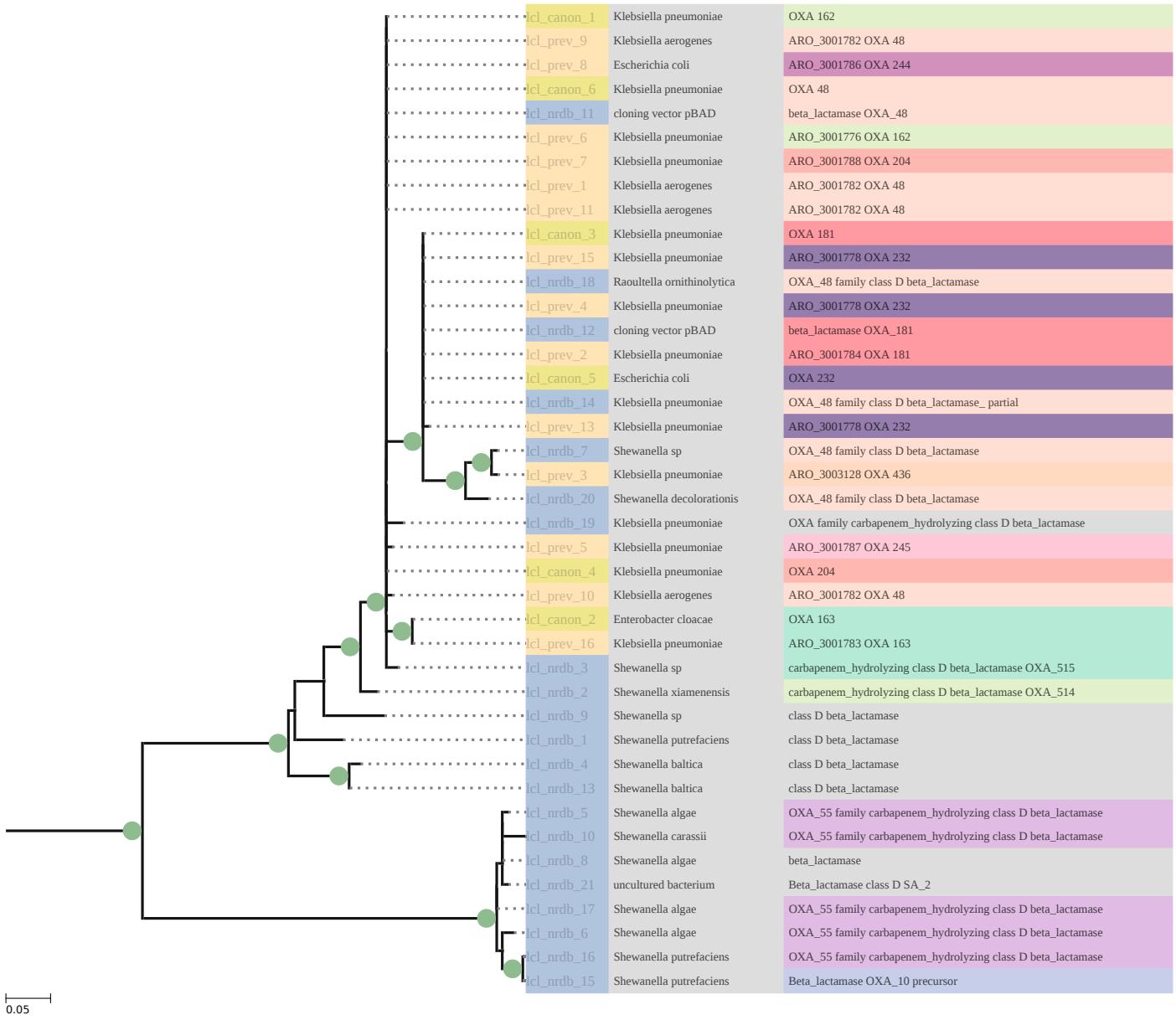


Figure 6: The main phylogenetic relationship pruned from the tree inferred from OXA-48-like sequences from figure (#fig:oxa-tree). 6 canonical (labels prefixed with 'cl_canon'_ in yellow), and 14 prevalence (labels prefixed with 'cl_prev' in tan), 20 NCBI non-redundant sequences (labels prefixed with 'cl_nrd' in blue). Each OXA-48 variant is coloured based on its primary numerical value. If the sequence is not reported to be OXA-48 family it is coloured in grey.

The OXA-48 prevalence hits added further diversity to the reference OXA-48-like sequences. OXA-436 28 was found, clustered with no other gene, in the prevalence data, and OXA-514 and OXA-515 were found in the non-redundant data. Much like in NDM, this diversity seems only to be diversity in name. The various named variants are too similar to produce significant phylogenetic resolution.

Discussion

The aim of this study was to use the phylogenetic relationships of AMR genes to expand on current AMR surveillance efforts of CARD. Loose RGI matches in the MAGs for MCR, NDM, KPC, and OXA-48 were examined, but after applying conservative filtering criteria, only the MCR survey produced matches which would be worthy of further consideration for resistance testing. This contradistinction between the MCR analysis and the beta lactamase analyses could be explained by plasmid distributions. Plasmids could have been poorly recovered in the assembly phase of the SRA experiments. These experiments were performed with high-throughput sequencing, many using Illumina HiSeq 2000 and 2500, which produce short reads. Plasmids from short reads are known to be difficult to assemble due to sequence characteristics. For example repeat sequences, which are common on plasmids, are often shared with other genomic elements from many genomes, and produces many contigs of ambiguous origin [29]. There are tools which exist, like PlasmidSPAdes, Recycler, cBar and PlasmidFinder (TODO:doi), which are designed specifically for the task of assembling plasmids, but with metagenomic data, many of these tools have weaknesses which are prohibitive, especially for plasmids above 50 kbp in length[29]. In the SRA experiments, this specialized assembly is not used, and many of the plasmid fragments would have been binned poorly. If the MCR homologs were mostly present on chromosomes or other extrachromosomal molecules, and the beta lactamase genes were present on plasmids, one would expect a lack of beta lactamase hits (TODO: get this data in the results). If the criteria for filtering the downstream data are also considered, then there are a multitude of reasons that these plasmid fragments would not appear in the final gene alignments for NDM, KPC, and OXA-48. This data's genomes were recovered using Metabat [15] (TODO: Should this be mentioned when describing the dataset?), which clustered sequences into genome bins. After binning, filters were applied such that only contigs meeting a certain minimum size, and coverage were used. In the phylogenetic portion of the analysis, conservative measures were taken to produce trees with high support values. These constraints included strict ranges for query coverage. It would be unlikely that genes on plasmids were included in the analysis after this entire process took place. Even though the MCR tree included 91 MAG sourced sequences, only a portion of these were of any interest. Many of the sequences fell too far from any canonical MCR clade. This shows that to provide a real effort with metagenomic data and surveillance involving phylogenetic relationships, it would be desirable to sequence with longer reads, have better assembly for plasmids, and to have more data. In treating this data in a conservative way, we sacrifice many low quality, questionable phylogenetic relationships.

for few, potentially meaningful, high quality relationships. When dealing with second generation sequencing methods and metagenomics, it appears that high quality data is rare. One should be cautious when drawing conclusions about relationships derived from this sort of data.

Another way in which the analyses of MCR and the beta lactamases differ is in the spread and diversity of the relationships. In addition to MAG sources, which are completely lacking in the beta lactamase trees, the MCR tree contains several sequences from NCBI's non-redundant database which land close to canonical clades. There are several clades made up of canonical and prevalence sequence genes separated by swathes of diversity from both NCBI and the MAGs. This suggests that there could be more diversity to be found within the MCR class. In contrast, the added diversity in phylogenies from OXA-48, KPC is distant from the canonical and prevalence sequences. The matches which are close, are nearly identical to the canonical and prevalence sequences. The beta-lactamase phylogenies are low in gene diversity, and thus form very tight, closely related clades. This displays that this study is very sensitive to the diversity of the gene being studied, whether it be from bias in sampling, or high amount of genetic conservation in the gene pool.

In assessing the value of the relationships present in these more tightly related phylogenies, it is surprising that the literature has produced so many named variants. In the tree for NDM, we have named variants numbered from NDM 1 to NDM 14, yet most of the tree would be more appropriately displayed with multifurcating nodes. There are fifteen named variants, but there are only three weakly distinct clades, and little bootstrap support (TODO: look at SNPs put this in results?). The most divergent NDM sequence is a prevalence sequence which is labelled as NDM-1, yet is more distant from other homologs named NDM-1 than other NDM-1 genes. The criteria for naming these genes seem to be completely arbitrary. Similarly, KPC is very tightly related and some of the same arguments could be made for classification of these sequences. In the tree for OXA-48 like genes, we see more diversity, with branches of more support. The tree forms few tightly related clades, and allows some diversity from NCBI to be added to the tree. The OXA group, as a whole, is very diverse, but only because they are classified phenotypically. If the OXA were more thoroughly separated into actual homologous groups, perhaps we would see more diversity added to the group, and perhaps even from metagenomic data. These groups of OXA could behave very differently and some could have data recovery similar to that of MCR, but we are missing out on this because of the phenotypic naming scheme over a rigorous scheme which takes into account the genetic similarity. In MCR, the naming situation is less arbitrary. The closely related variants are classified under the same primary number. Several versions of MCR-1 exist, for example, and all form a clade.

TODO: Should we talk about things like Leclercia? e.g: Most of the clades of new organisms from UBAs are not surprising. For the most part, taxa branched where expected, there doesn't seem to be much unexplained here. Many of these families added to the tree are associated with human disease/animal disease already, and one could expect to find colistin resistance there, however one family, Leclercia, branches in a curious place, and has not been shown to carry the MCR gene like this in any study. If anything, these are at least an indicator that our conservative treatment of the metagenomic data was a good idea

Conclusion

- The data for three of the four phylogenetic analyses were not ideal for expanding on the known diversity of AMR homologs. These beta lactamase phylogenies did, however show the importance of unambiguous naming schemes, and care when it comes to presenting data honestly.
- The study also showed that it is important to take the type of data into consideration, and metagenomic sequencing must come a long way before these studies can produce quality information.
- MCR is a good candidate for this type of study because it has the right balance of homology and diversity in the group, and is an example of a better naming scheme.

Supplemental

Table 1: Names of the canonical MCR phosphoethanolamine transferase variants from CARD database

MCR-9	MCR-1.8	MCR-1.11	MCR-3.4	MCR-3.11	MCR-6.1	MCR-3.3	MCR-2.2	MCR-1.5	MCR-3.6	MCR-3.10	MCR-3.5	MCR-1.10	MCR-1.9	MCR-3.9	MCR-3.8	MCR-3.7	MCR-3.2	MCR-1.6
MCR-1.2	MCR-1.3	MCR-1.4	MCR-8	MCR-7.1	MCR-1.7	MCR-1.12	MCR-1.13	MCR-3.12	MCR-4.2	MCR-4.3	MCR-4.4	MCR-4.5	MCR-5.2					

Table 2: Names of the canonical beta-lactamase variants from CARD database

KPC-1	KPC-3	KPC-4	KPC-5	KPC-6	KPC-7	KPC-8	KPC-9	KPC-10	KPC-11	KPC-12	KPC-13	KPC-14	KPC-15	KPC-16	KPC-17	KPC-19	KPC-22	KPC-24
-------	-------	-------	-------	-------	-------	-------	-------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------

Table 3: Names of the canonical NDM beta-lactamase variants from CARD database

NDM-1	NDM-2	NDM-5	NDM-3	NDM-4	NDM-6	NDM-7	NDM-8	NDM-9	NDM-10	NDM-11	NDM-12	NDM-13	NDM-14	NDM-17
-------	-------	-------	-------	-------	-------	-------	-------	-------	--------	--------	--------	--------	--------	--------

Table 4: Names of the canonical OXA beta-lactamase variants from CARD database

OXA-162	OXA-163	OXA-181	OXA-204	OXA-232	OXA-48
---------	---------	---------	---------	---------	--------

TODO: Quality/CHECKM information for UBAs

Table 5: Outgroups chosen in building the phylogenies of the 4 AMR families.

Outgroup to AMR Family	Gene	Species	Class
MCR	phosphoethanolamine-lipid A transferase	Oxalobacter formigenes	Betaproteobacteria
KPC	class A beta-lactamase	Mesorhizobium loti	Alpha Proteobacteria
NDM	VIM-4 metallo-beta-lactamase	Alcaligenes faecalis	Betaproteobacteria
OXA-48	OXA-10 family class D beta-lactamase OXA-454	Delftia acidovorans	Betaproteobacteria

MCR Alignment

[TODO: visualization with alignment/entropy] 

KPC Alignment

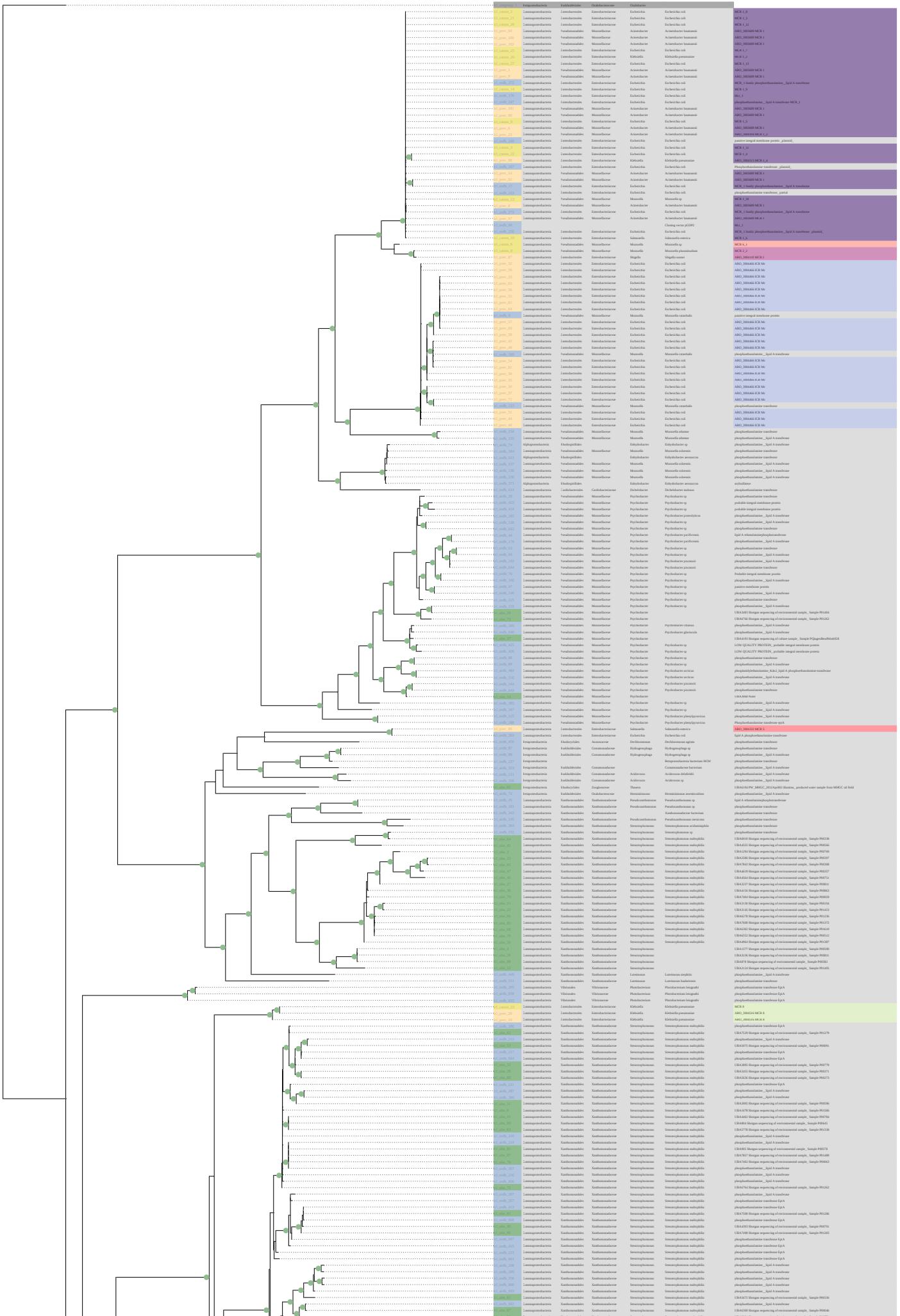
[TODO: visualization with alignment/entropy] 

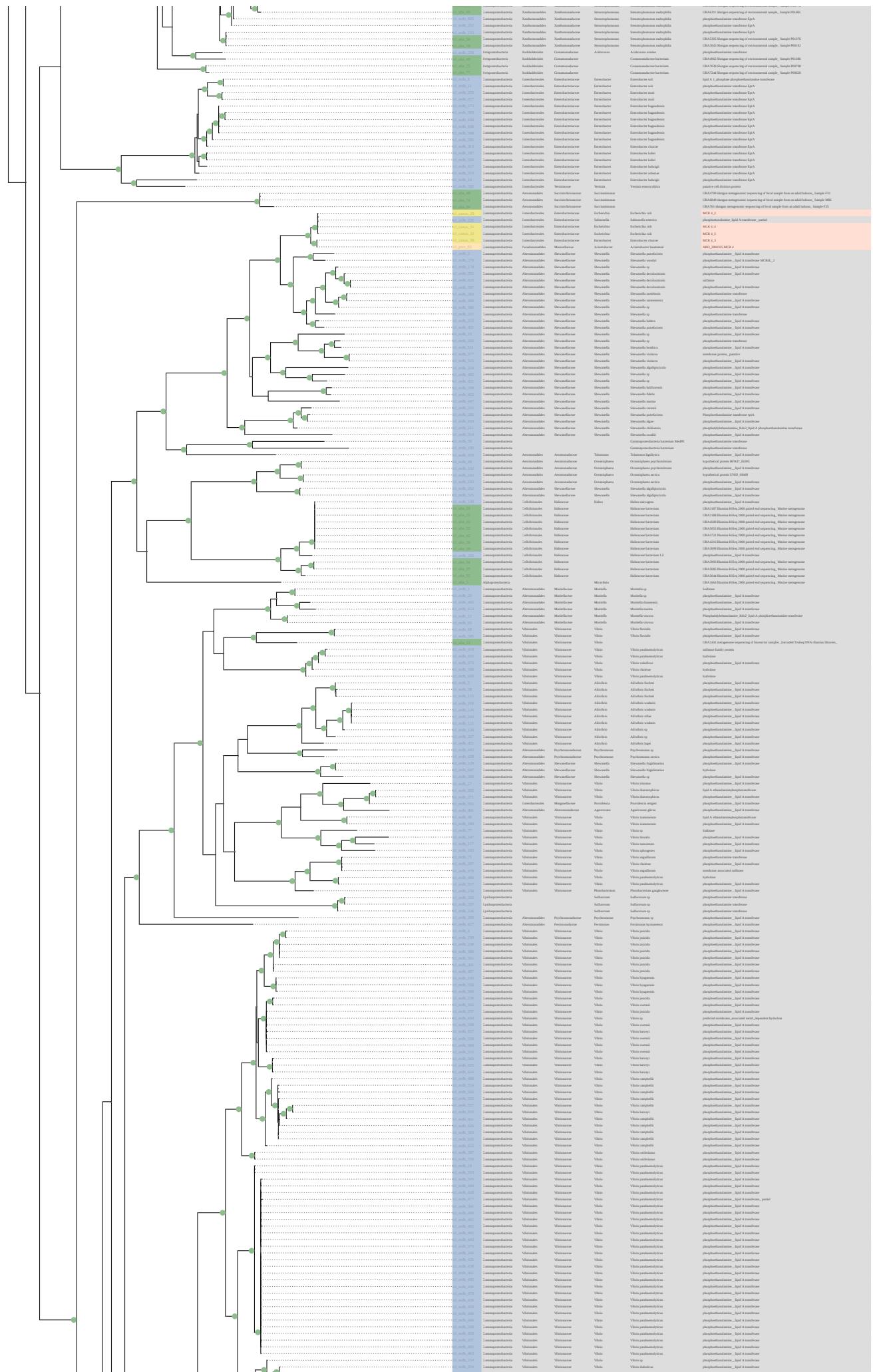
NDM Alignment

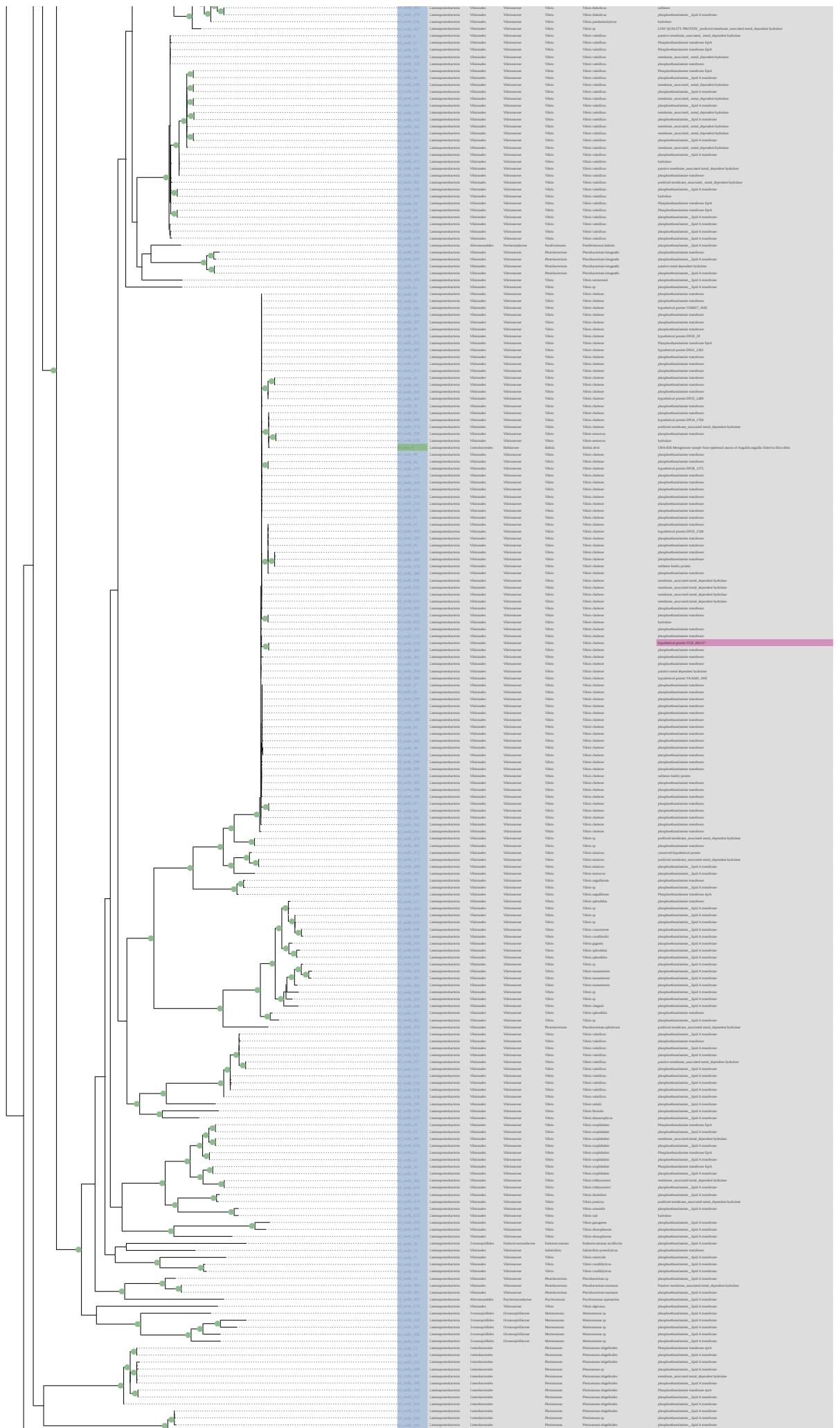
[TODO: visualization with alignment/entropy] 

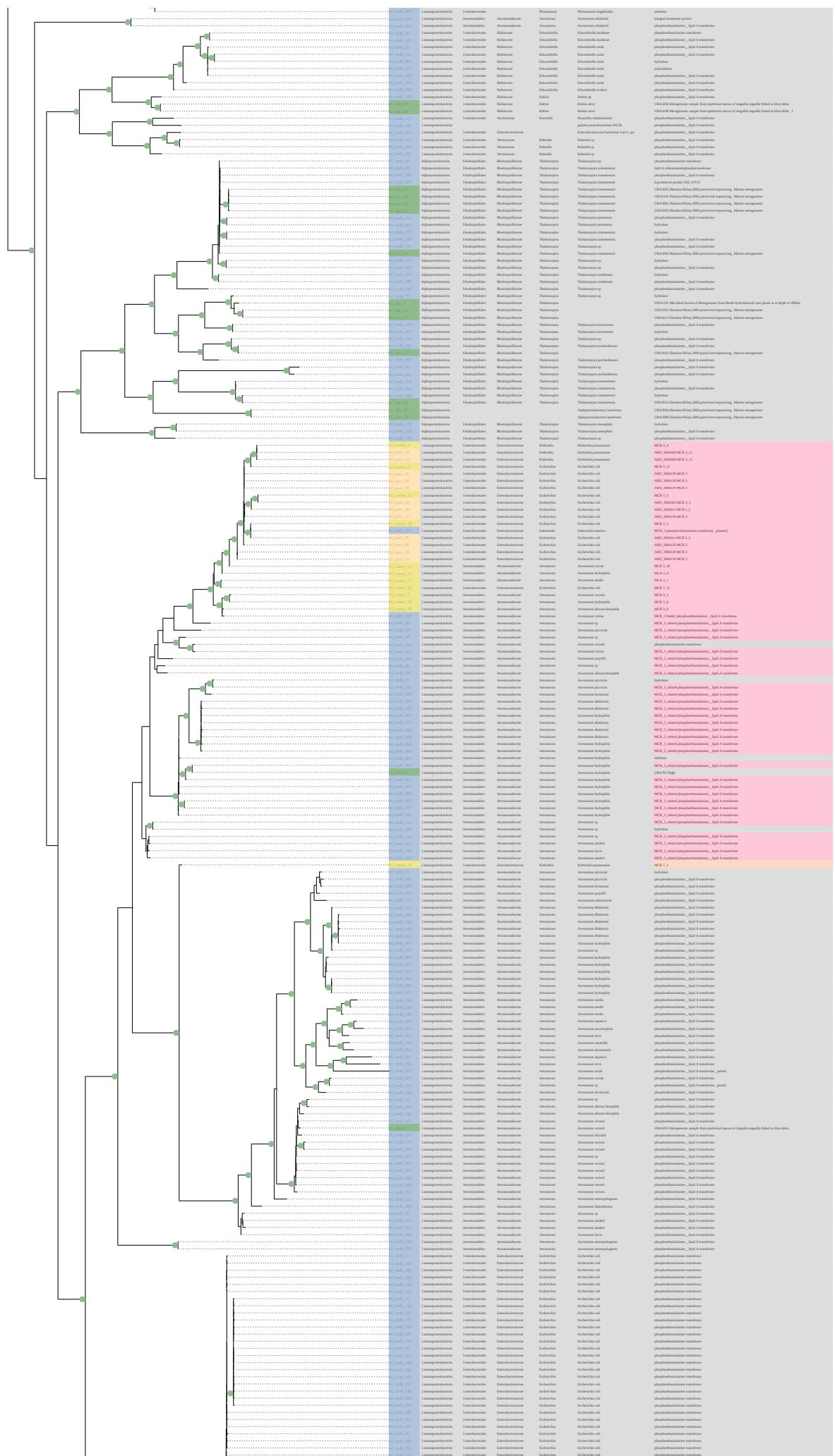
OXA-48 Alignment

[TODO: visualization with alignment/entropy] 









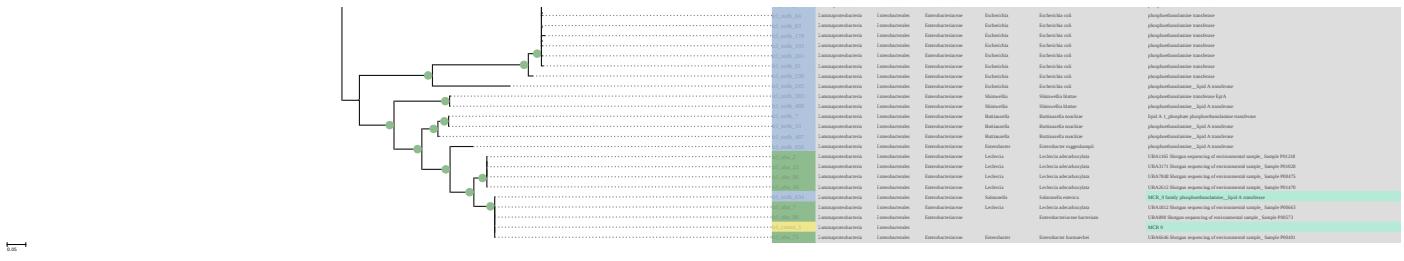


Figure 7: Phylogenetic relationship of 32 canonical (labels prefixed with lcl_canon_ in yellow), and 54 prevalence (labels prefixed with lcl_prev_ in tan) MCR family sequences, 595 NCBI non-redundant sequences (labels prefixed with lcl_prev_ in blue), and 91 UBA sourced sequences (labels prefixed with lcl_prev_ in green). Each MCR variant is coloured based on its primary numerical value. If the sequence is not reported to be MCR family it is coloured in grey.

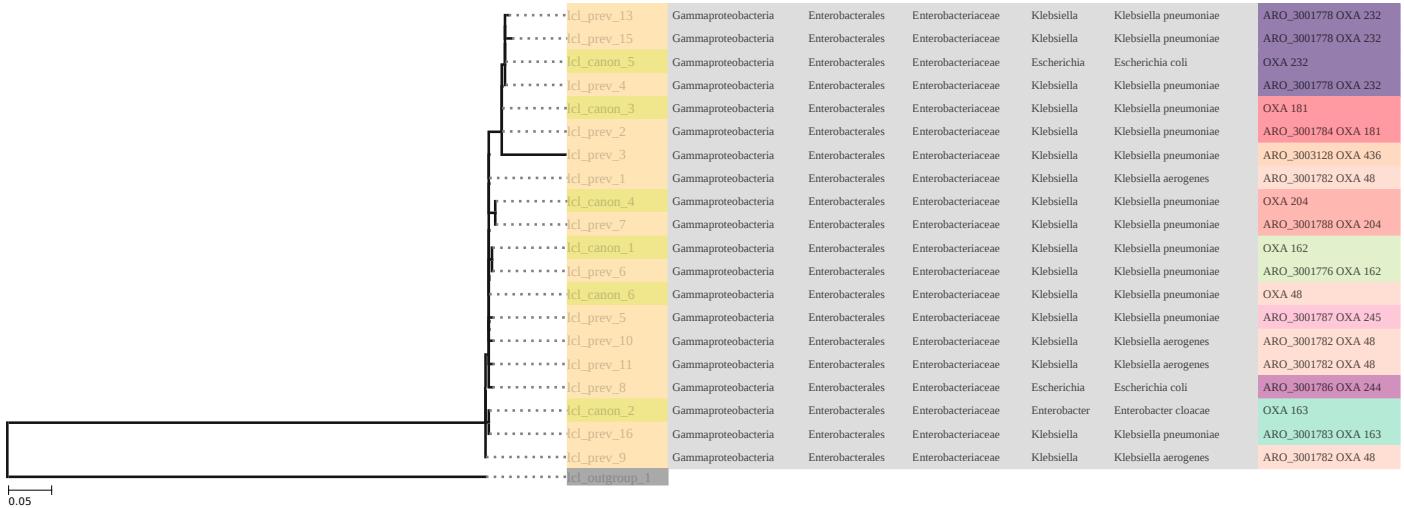
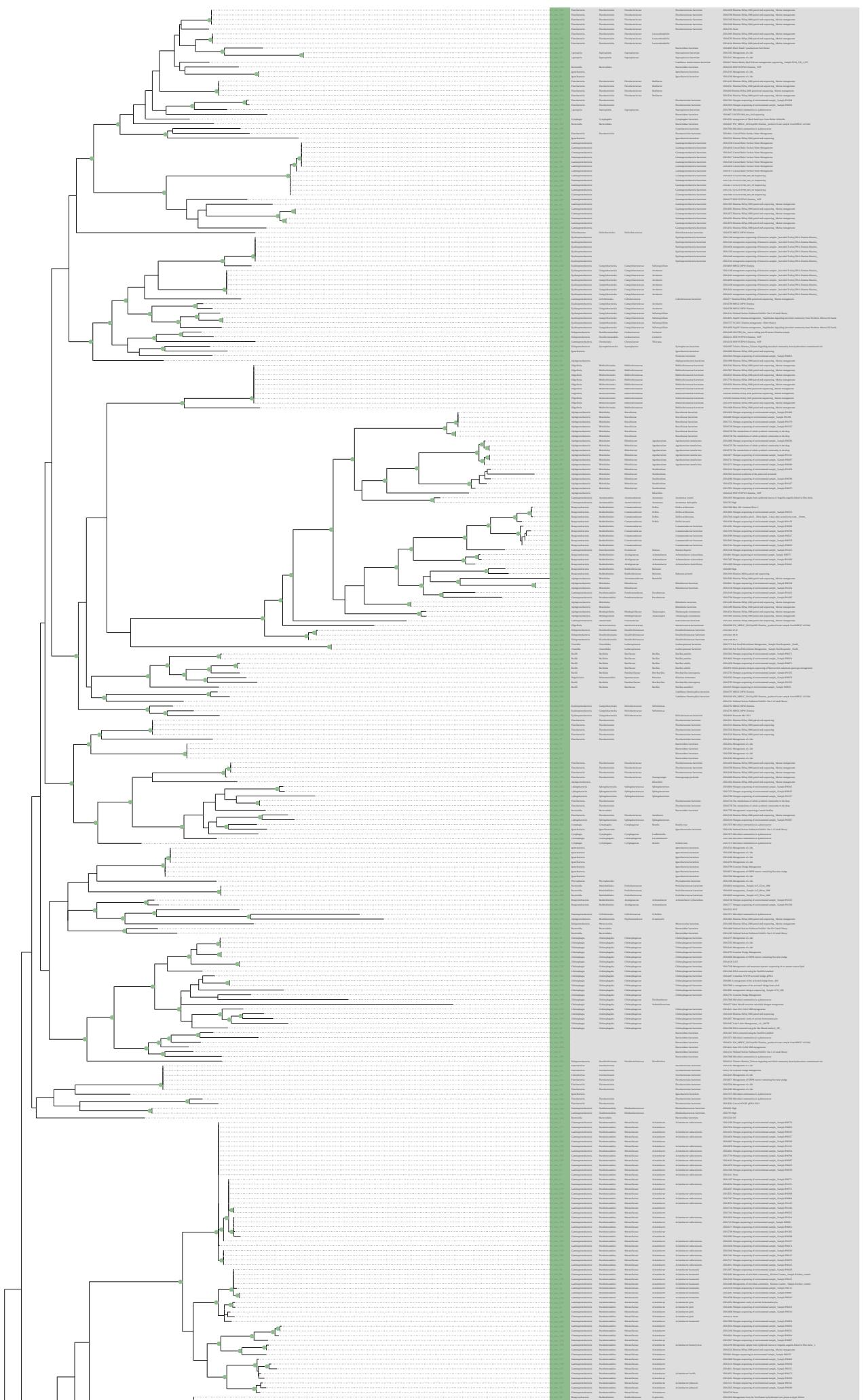


Figure 8: Phylogenetic relationship (lcl_canon) OXA-48 family sequences along with 14 prevalence (lcl_prev) sequences



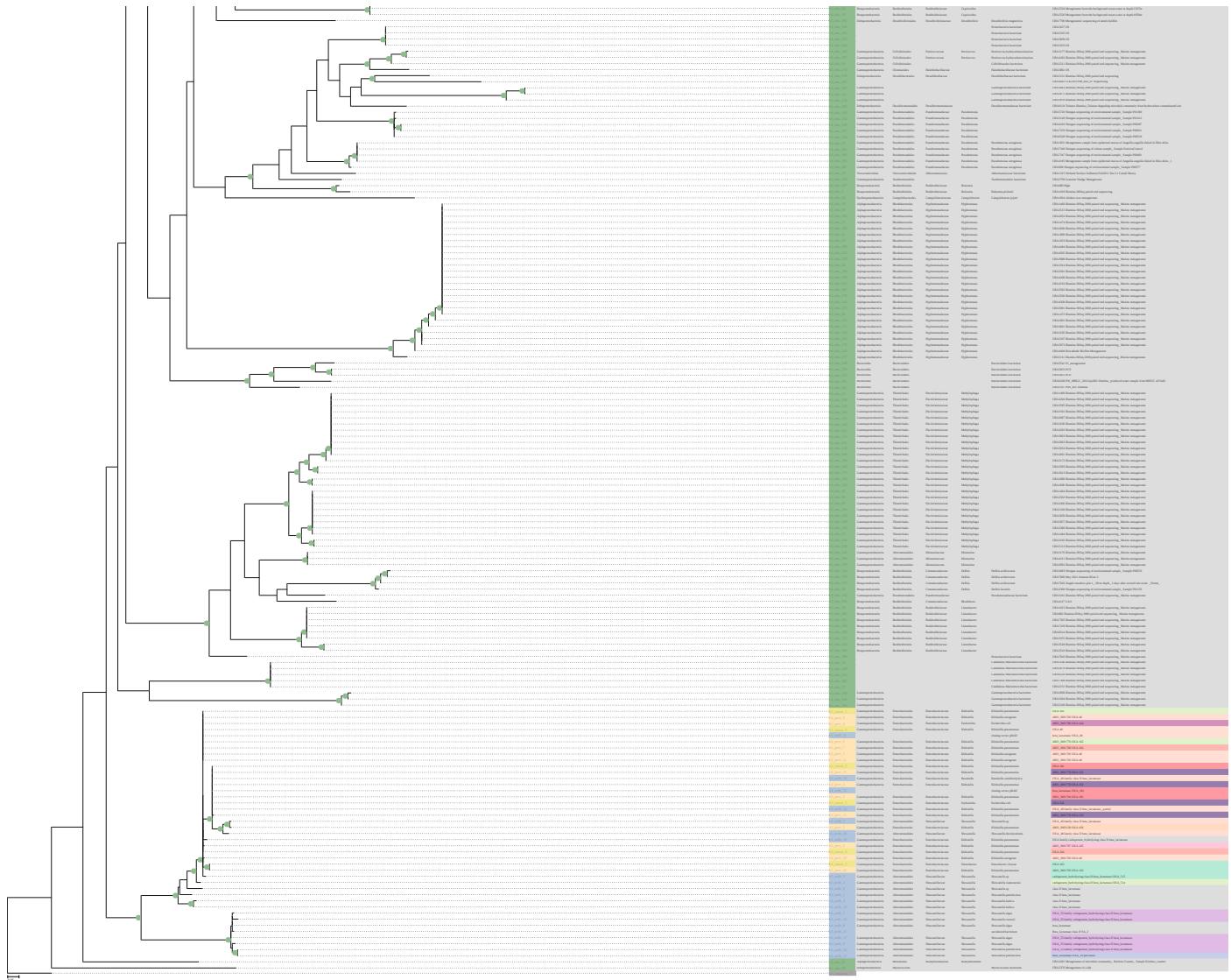


Figure 9: Phylogenetic relationship (lcl_canon) OXA-48 family sequences along with 14 prevalence (lcl_prev) sequences.

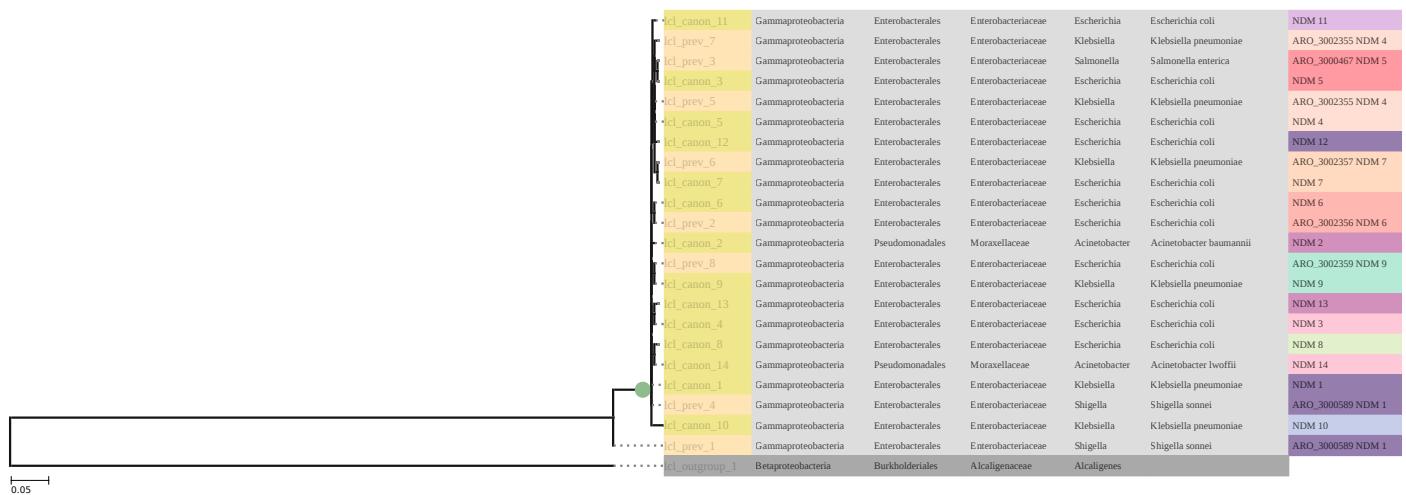


Figure 10: Phylogenetic relationship (lcl_canon) NDM family sequences along with 14 prevalence (lcl_prev) sequences.

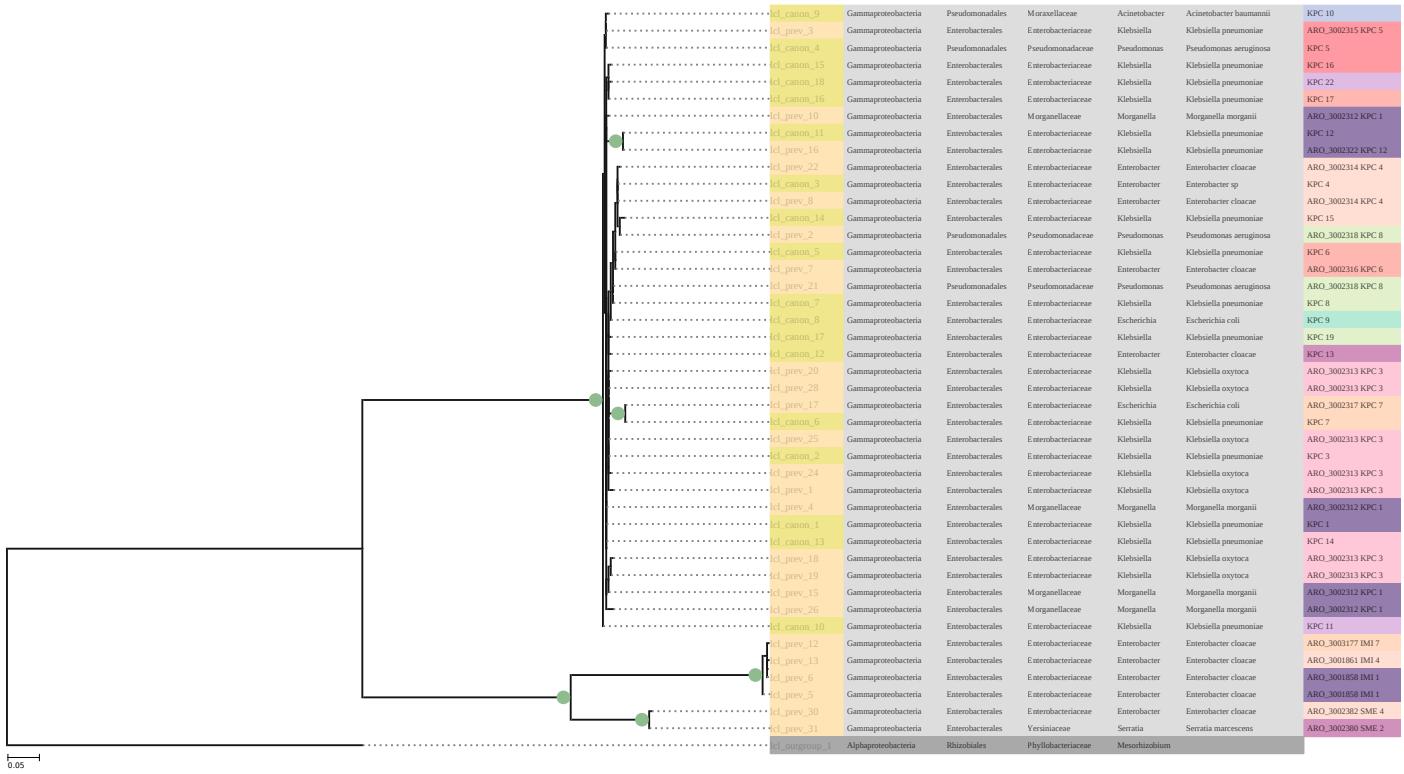
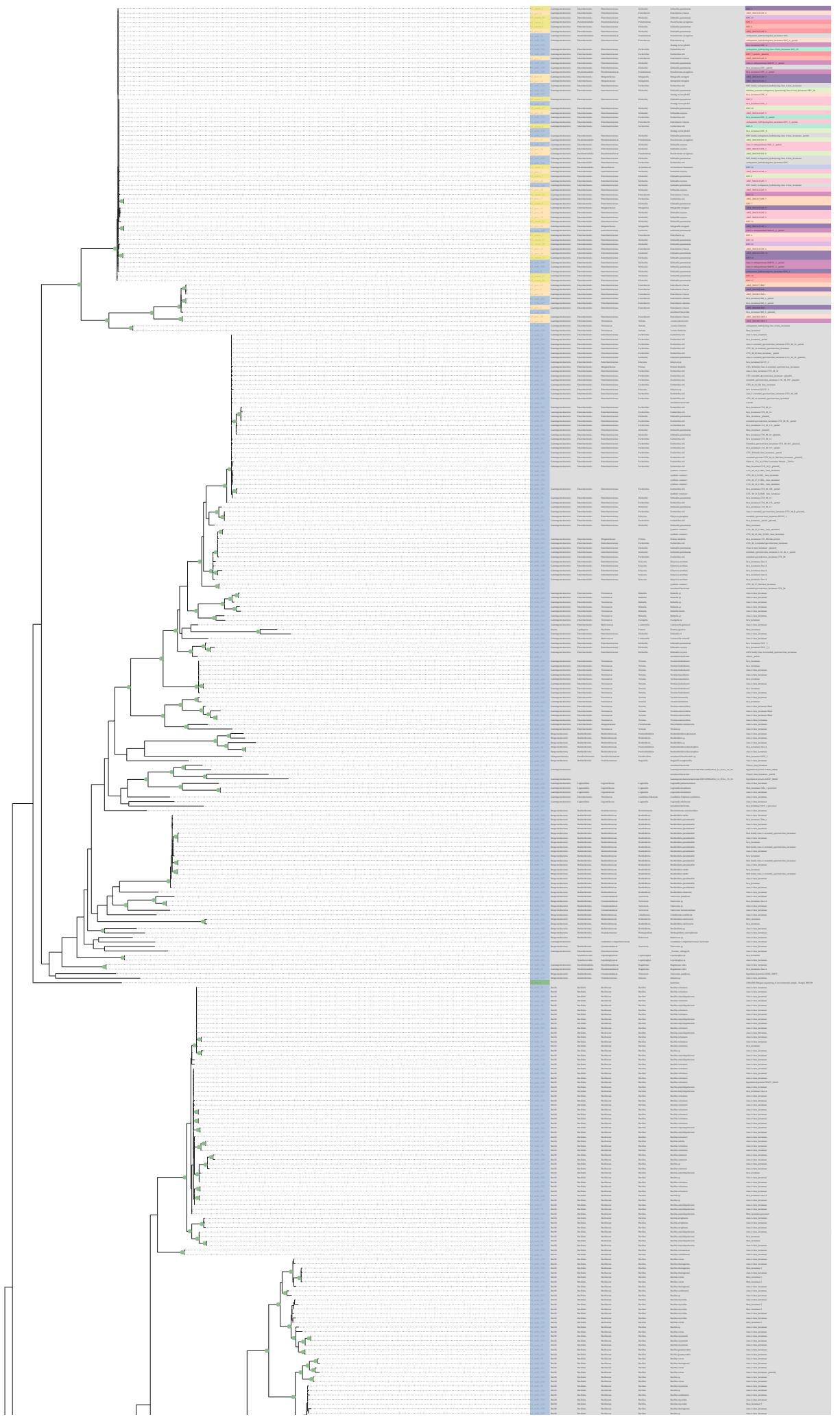


Figure 11: Phylogenetic relationship (lcl_canon) KPC family sequences along with 14 prevalence (lcl_prev) sequences.



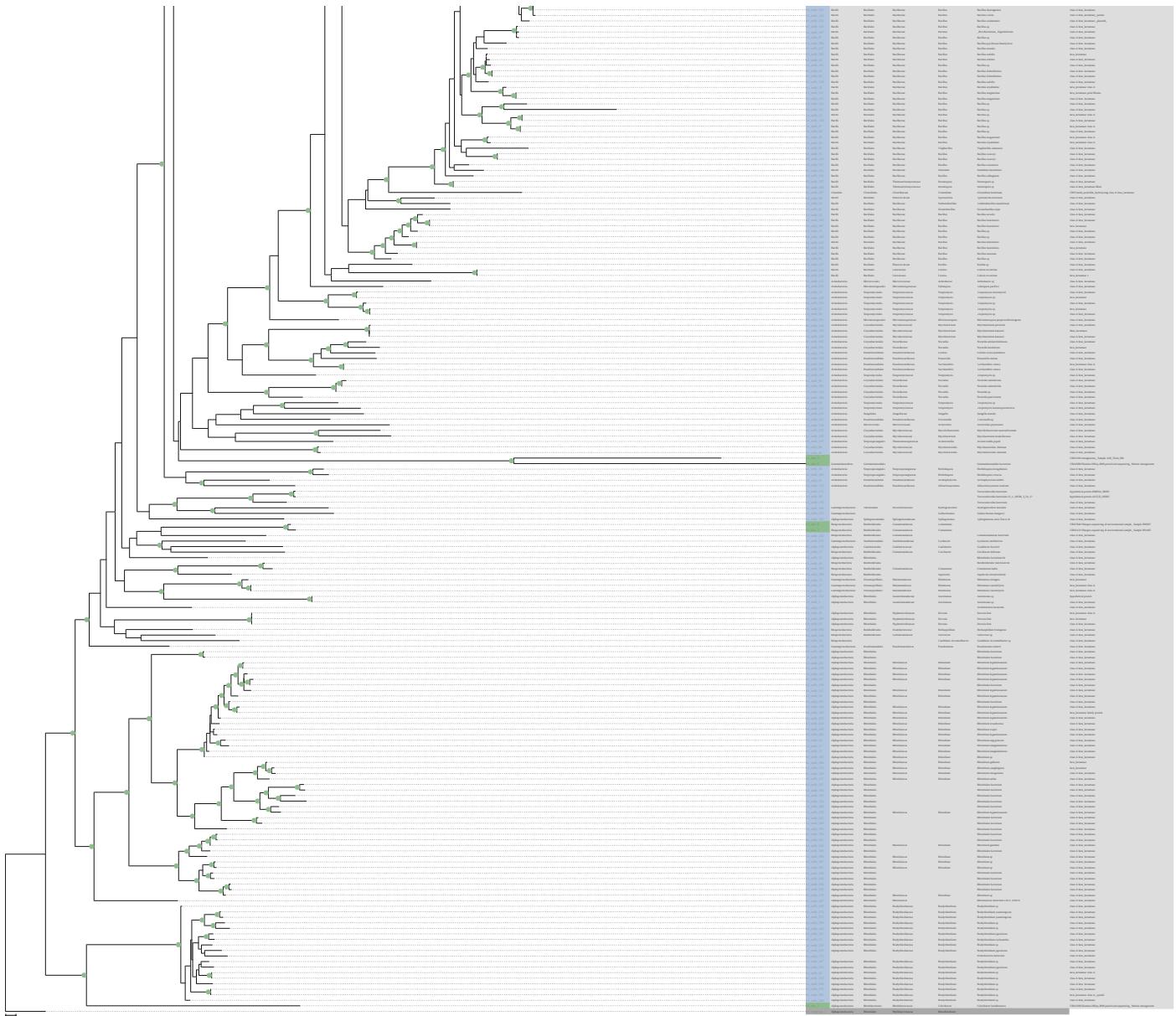


Figure 12: Phylogenetic relationship (lcl_canon) KPC family sequences along with 14 prevalence (lcl_prev) sequences.

References

- 1. Antibiotic Resistance**☆
B. Périchon, P. Courvalin, C. W. Stratton
Reference Module in Biomedical Sciences (2015) <https://doi.org/dbhg>
DOI: [10.1016/b978-0-12-801238-3.02385-0](https://doi.org/10.1016/b978-0-12-801238-3.02385-0)
- 2. Antibiotic resistance is ancient**
Vanessa M. D'Costa, Christine E. King, Lindsay Kalan, Mariya Morar, Wilson W. L. Sung, Carsten Schwarz, Duane Froese, Grant Zazula, Fabrice Calmels, Regis Debruyne, ... Gerard D. Wright
Nature (2011-08-31) <https://doi.org/b3wbvx>
DOI: [10.1038/nature10388](https://doi.org/10.1038/nature10388) · PMID: [21881561](#)
- 3. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database**
Baofeng Jia, Amogelang R. Raphenya, Brian Alcock, Nicholas Waglechner, Peiyao Guo, Kara K. Tsang, Briony A. Lago, Biren M. Dave, Sheldon Pereira, Arjun N. Sharma, ... Andrew G. McArthur
Nucleic Acids Research (2016-10-26) <https://doi.org/f9wbjs>
DOI: [10.1093/nar/gkw1004](https://doi.org/10.1093/nar/gkw1004) · PMID: [27789705](#) · PMCID: [PMC5210516](#)
- 4. Horizontal transfer of antibiotic resistance genes in clinical environments**
Nicole A. Lermniaux, Andrew D. S. Cameron
Canadian Journal of Microbiology (2019-01) <https://doi.org/gfnrkj>
DOI: [10.1139/cjm-2018-0275](https://doi.org/10.1139/cjm-2018-0275) · PMID: [30248271](#)
- 5. Carbapenem Resistance: A Review**
Francis Codjoe, Eric Donkor
Medical Sciences (2017-12-21) <https://doi.org/c9zd>
DOI: [10.3390/medsci6010001](https://doi.org/10.3390/medsci6010001) · PMID: [29267233](#) · PMCID: [PMC5872158](#)
- 6. Mechanisms of Antimicrobial Resistance in ESKAPE Pathogens**
Sirijan Santajit, Nitaya Indrawattana
BioMed Research International (2016) <https://doi.org/f9h4q3>
DOI: [10.1155/2016/2475067](https://doi.org/10.1155/2016/2475067) · PMID: [27274985](#) · PMCID: [PMC4871955](#)
- 7. The structure of β-lactamases**
Philosophical Transactions of the Royal Society of London. B, Biological Sciences (1980-05-16) <https://doi.org/cdckks>
DOI: [10.1098/rstb.1980.0049](https://doi.org/10.1098/rstb.1980.0049) · PMID: [6109327](#)
- 8. Updated Functional Classification of β-Lactamases**
Karen Bush, George A. Jacoby
Antimicrobial Agents and Chemotherapy (2009-12-07) <https://doi.org/bp6dp2>
DOI: [10.1128/aac.01009-09](https://doi.org/10.1128/aac.01009-09) · PMID: [19995920](#) · PMCID: [PMC2825993](#)
- 9. Class D OXA-48 Carbapenemase in Multidrug-Resistant Enterobacteria, Senegal**
Olivier Moquet, Coralie Bouchiat, Alfred Kinana, Abdoulaye Seck, Omar Arouna, Raymond Bercion, Sébastien Breurec, Benoit Garin
Emerging Infectious Diseases (2011-01) <https://doi.org/cn235v>
DOI: [10.3201/eid1701.100244](https://doi.org/10.3201/eid1701.100244) · PMID: [21192883](#) · PMCID: [PMC3204621](#)
- 10. Rapid Identification of OXA-48 and OXA-163 Subfamilies in Carbapenem-Resistant Gram-Negative Bacilli with a Novel Immunochemical Lateral Flow Assay**
Fernando Pasteran, Laurence Denorme, Isabelle Ote, Sonia Gomez, Denise De Belder, Youri Glupczynski, Pierre Bogaerts, Barbara Ghiglione, Pablo Power, Pascal Mertens, Alejandra Corso
Journal of Clinical Microbiology (2016-08-17) <https://doi.org/dbhs>
DOI: [10.1128/jcm.01175-16](https://doi.org/10.1128/jcm.01175-16) · PMID: [27535687](#) · PMCID: [PMC5078564](#)
- 11. Treatment Options for Carbapenem-Resistant Enterobacteriaceae Infections**
H. J. Morrill, J. M. Pogue, K. S. Kaye, K. L. LaPlante
Open Forum Infectious Diseases (2015-05-05) <https://doi.org/dbht>
DOI: [10.1093/ofid/ofv050](https://doi.org/10.1093/ofid/ofv050) · PMID: [26125030](#) · PMCID: [PMC4462593](#)
- 12. Towards Understanding MCR-like Colistin Resistance**
Jian Sun, Huimin Zhang, Ya-Hong Liu, Youjun Feng
Trends in Microbiology (2018-09) <https://doi.org/gdqcfc>
DOI: [10.1016/j.tim.2018.02.006](https://doi.org/10.1016/j.tim.2018.02.006) · PMID: [29525421](#)
- 13. Molecular Characterization of OXA-20, a Novel Class D β-Lactamase, and Its Integron from *Pseudomonas aeruginosa***
Thierry Naas, Wladimir Sougakoff, Anne Casetta, Patrice Nordmann
Antimicrobial Agents and Chemotherapy (1998-08-01) <https://doi.org/dwsr>
DOI: [10.1128/aac.42.8.2074](https://doi.org/10.1128/aac.42.8.2074)
- 14. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database**
Brian P Alcock, Amogelang R Raphenya, Tammy TY Lau, Kara K Tsang, Mégane Bouchard, Arman Edalatmand, William Huynh, Anna-Lisa V Nguyen, Annie A Cheng, Sihan Liu, ... Andrew G McArthur
Nucleic Acids Research (2019-10-29) <https://doi.org/ggckg6>
DOI: [10.1093/nar/gkz935](https://doi.org/10.1093/nar/gkz935) · PMID: [31665441](#) · PMCID: [PMC7145624](#)
- 15. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities**
Dongwan D. Kang, Jeff Froula, Rob Egan, Zhong Wang

16. Recovery of genomes from metagenomes via a derePLICATION, aggregation and scoring strategy

Christian M. K. Sieber, Alexander J. Probst, Alison Sharrar, Brian C. Thomas, Matthias Hess, Susannah G. Tringe, Jillian F. Banfield

Nature Microbiology (2018-05-28) <https://doi.org/gfwwg>

DOI: [10.1038/s41564-018-0171-1](https://doi.org/10.1038/s41564-018-0171-1) · PMID: [29807988](https://pubmed.ncbi.nlm.nih.gov/29807988/) · PMCID: [PMC6786971](https://pubmed.ncbi.nlm.nih.gov/PMC6786971/)

17. Author Correction: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, Gene W. Tyson

Nature Microbiology (2017-12-12) <https://doi.org/c8rq>

DOI: [10.1038/s41564-017-0083-5](https://doi.org/10.1038/s41564-017-0083-5) · PMID: [29234139](https://pubmed.ncbi.nlm.nih.gov/29234139/)

18. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz, Gene W. Tyson

Nature Microbiology (2017-09-11) <https://doi.org/cczd>

DOI: [10.1038/s41564-017-0012-7](https://doi.org/10.1038/s41564-017-0012-7) · PMID: [28894102](https://pubmed.ncbi.nlm.nih.gov/28894102/)

19. Basic local alignment search tool

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, David J. Lipman

Journal of Molecular Biology (1990-10) <https://doi.org/cnsjsz>

DOI: [10.1016/s0022-2836\(05\)80360-2](https://doi.org/10.1016/s0022-2836(05)80360-2)

20. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies

Lam-Tung Nguyen, Heiko A. Schmidt, Arndt von Haeseler, Bui Quang Minh

Molecular Biology and Evolution (2014-11-03) <https://doi.org/f3srt>

DOI: [10.1093/molbev/msu300](https://doi.org/10.1093/molbev/msu300) · PMID: [25371430](https://pubmed.ncbi.nlm.nih.gov/25371430/) · PMCID: [PMC4271533](https://pubmed.ncbi.nlm.nih.gov/PMC4271533/)

21. Substrate Recognition by a Colistin Resistance Enzyme from *Moraxella catarrhalis*

Peter J. Stogios, Georgina Cox, Haley L. Zubyk, Elena Evdokimova, Zdzislaw Wawrzak, Gerard D. Wright, Alexei Savchenko

ACS Chemical Biology (2018-04-09) <https://doi.org/gdnrmn>

DOI: [10.1021/acscchembio.8b00116](https://doi.org/10.1021/acscchembio.8b00116) · PMID: [29631403](https://pubmed.ncbi.nlm.nih.gov/29631403/) · PMCID: [PMC6197822](https://pubmed.ncbi.nlm.nih.gov/PMC6197822/)

22. The Genus *Aeromonas*: Taxonomy, Pathogenicity, and Infection

J. Michael Janda, Sharon L. Abbott

Clinical Microbiology Reviews (2010-01) <https://doi.org/cbnfww>

DOI: [10.1128/cmr.00039-09](https://doi.org/10.1128/cmr.00039-09) · PMID: [20065325](https://pubmed.ncbi.nlm.nih.gov/20065325/) · PMCID: [PMC2806660](https://pubmed.ncbi.nlm.nih.gov/PMC2806660/)

23. Natural antimicrobial susceptibility patterns and biochemical profiles of *Leclercia adecarboxylata* strains

I Stock, S Burak, B Wiedemann

Clinical Microbiology and Infection (2004-08) <https://doi.org/b2zr9k>

DOI: [10.1111/j.1469-0998.2004.00892.x](https://doi.org/10.1111/j.1469-0998.2004.00892.x) · PMID: [15301675](https://pubmed.ncbi.nlm.nih.gov/15301675/)

24. The Genus *Psychrobacter*

Elliot Juni

The Prokaryotes (1992) <https://doi.org/dbpr>

DOI: [10.1007/978-1-4757-2191-1_12](https://doi.org/10.1007/978-1-4757-2191-1_12)

25. Novel *Psychrobacter* Species from Antarctic Ornithogenic Soils

J. P. BOWMAN, J. CAVANAGH, J. J. AUSTIN, K. SANDERSON

International Journal of Systematic Bacteriology (1996-10-01) <https://doi.org/fd9rb9>

DOI: [10.1099/00207713-46-4-841](https://doi.org/10.1099/00207713-46-4-841) · PMID: [8863407](https://pubmed.ncbi.nlm.nih.gov/8863407/)

26. *Psychrobacter vallis* sp. nov. and *Psychrobacter aquaticus* sp. nov., from Antarctica

Sisinty Shivaji, Gundlapalli S. N. Reddy, Korbole Suresh, Pratima Gupta, Suresh Chintalapati, Peter Schumann, Erko Stackebrandt, Genki I. Matsumoto

International Journal of Systematic and Evolutionary Microbiology (2005-03-01) <https://doi.org/cdq9w4>

DOI: [10.1099/ijs.0.03030-0](https://doi.org/10.1099/ijs.0.03030-0) · PMID: [15774658](https://pubmed.ncbi.nlm.nih.gov/15774658/)

27. *Stenotrophomonas maltophilia*: an Emerging Global Opportunistic Pathogen

J. S. Brooke

Clinical Microbiology Reviews (2012-01-01) <https://doi.org/fxx2tt>

DOI: [10.1128/cmr.00019-11](https://doi.org/10.1128/cmr.00019-11) · PMID: [22232370](https://pubmed.ncbi.nlm.nih.gov/22232370/) · PMCID: [PMC3255966](https://pubmed.ncbi.nlm.nih.gov/PMC3255966/)

28. Dissemination and Characteristics of a Novel Plasmid-Encoded Carbapenem-Hydrolyzing Class D β -Lactamase, OXA-436, Found in Isolates from Four Patients at Six Different Hospitals in Denmark

Ørjan Samuelsen, Frank Hansen, Bettina Aasnæs, Henrik Hasman, Bjarte Aarmo Lund, Hanna-Kirsti S. Leiros, Berit Lilje, Jessin Janice, Lotte Jakobsen, Pia Littauer, ... Anette M. Hammerum

Antimicrobial Agents and Chemotherapy (2017-10-23) <https://doi.org/db7x>

DOI: [10.1128/aac.01260-17](https://doi.org/10.1128/aac.01260-17) · PMID: [29061750](https://pubmed.ncbi.nlm.nih.gov/29061750/) · PMCID: [PMC5740310](https://pubmed.ncbi.nlm.nih.gov/PMC5740310/)

29. On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data

Sergio Arredondo-Alonso, Rob J. Willems, Willem van Schaik, Anita C. Schürch

Microbial Genomics (2017-10-01) <https://doi.org/gf6b63>

DOI: [10.1099/mgen.0.000128](https://doi.org/10.1099/mgen.0.000128) · PMID: [29177087](https://pubmed.ncbi.nlm.nih.gov/29177087/) · PMCID: [PMC5695206](https://pubmed.ncbi.nlm.nih.gov/PMC5695206/)