The challenges of defining the human nasopharyngeal resistome

Lucy O’Connor1,* and Robert Heyderman1

The nasopharynx is an important microbial reservoir for the emergence and spread of antibiotic-resistant organisms. The nasopharyngeal resistome is an extensive, adaptable reservoir of antibiotic-resistance genes (ARGs) within this niche. Metagenomic sequencing decodes the genetic material of all organisms within a sample using next-generation technologies, permitting unbiased discovery of novel ARGs and associated mobile genetic elements (MGEs). The challenges of sequencing a low-biomass bacterial sample have limited exploration of the nasopharyngeal resistome. Here, we explore the current understanding of the nasopharyngeal resistome, particularly the role of MGEs in propagating antimicrobial resistance (AMR), explore the advantages and limitations of metagenomic sequencing technologies and bioinformatic pipelines for nasopharyngeal resistome analysis, and highlight the key outstanding questions for future research.

Introduction

AMR has become a worrying threat to global public health, in part due to widespread use and misuse of antibiotics. Resistance is now detected to all antibiotics, including those that form the last line of defence against multidrug-resistant infections [1]. It is predicted that deaths due to AMR will rise to 10 million per annum by the year 2050 [2].

The resistome (see Glossary) is an extensive and adaptable reservoir of ARGs that exist within a microbial niche [3]. The human resistome is largely derived from harmless colonising bacteria; bacterial pathogens account for just a small fraction of ARGs within these communities [4]. The resistome is dominated by low-risk ARGs that are dormant, expressed at low levels, or are less able to transfer between bacteria, including precursor genes, which have the potential to generate resistance under antibiotic selective pressure, and cryptic resistance genes, which are embedded in bacterial chromosomes [3]. However, some colonising bacteria harbour ARGs colocalised with MGEs that can transfer to bacterial pathogens, causing antibiotic-resistant, life-threatening, invasive disease [5,6].

The nasopharynx is an important microbial reservoir for the emergence and spread of AMR organisms that cause pneumonia, meningitis, and sepsis, particularly in children aged under 5 years [7–9]. Key pathogens include Staphylococcus aureus, Streptococcus pyogenes, Streptococcus pneumoniae, and Haemophilus influenzae; the latter two have sufficient associated mortality to warrant vaccination programs, which in turn have promoted epidemiological shifts in carriage [8]. While there are multiple studies demonstrating the impact of antibiotics on AMR amongst individual bacterial species [9–11], analysis of the overall nasopharyngeal resistome remains challenging. There has been comparatively more investigation of the lower respiratory tract resistome, a separate ecological niche with similar sequencing challenges. Here, the search for targeted antibiotic strategies based upon individual resistome analysis has been driven by the

Highlights

The nasopharynx is an important, understudied, reservoir for emergence and spread of antibiotic-resistant bacteria that cause invasive disease, including pneumonia, meningitis, and sepsis, a leading cause of mortality in children under 5 years of age worldwide.

Metagenomic sequencing techniques provide a culture-unbiased, high-resolution tool for characterising nasopharyngeal resistome composition, with the ability to capture microbial taxonomy, function, and resistance from a single specimen.

Challenges in nasopharyngeal resistome characterisation include low sample biomass, high burden of human host DNA, the need to functionally validate novel ARGs, varying bioinformatic approaches, and limited computational resources.
abundant AMR in patients with chronic respiratory disease (e.g., cystic fibrosis), who have a high respiratory bacterial burden and prolonged antibiotic exposure [12].

Despite advances in culture-based and rapid-identification techniques, sometimes referred to as ‘culturomics’ [13,14], there remain uncultivable bacteria in the mucosal microbial community [15]. Next-generation metagenomic sequencing decodes the genetic material of all organisms within a sample [4], allowing more unbiased characterisation of bacterial diversity, ARGs, and MGEs [1,4,16,17]. Technological advances have improved affordability and throughput capacity [1], such that studies exploring the use of metagenomic sequencing for surveillance of emerging AMR in agricultural environments are now possible [18].

However, there are several challenges to effectively using metagenomics to study the nasopharyngeal resistome and its contribution to invasive infections. Nasopharyngeal swabs typically yield a low bacterial biomass with sequencing outputs that are dominated by human host DNA, thus reducing sensitivity of bacterial DNA detection [19–21]. Relating detected ARGs to phenotypic resistance is problematic, due to a lack of information about the genomic context of these resistance determinants [22].

In this review, we explore current understanding of the nasopharyngeal resistome, particularly the role of MGEs in propagating AMR; we examine the advantages and limitations of current metagenomic sequencing technologies and bioinformatic pipelines for nasopharyngeal resistome analysis and highlight outstanding questions for future research.

The resistome

Origins of the resistome

Genes with AMR potential are ubiquitous in the natural environment, driven by the evolutionary processes that have enabled bacteria to survive in proximity to other antibiotic-producing microbes within the same microbial niche [3,23]. Therefore environmental bacteria provide an inexhaustible genomic reservoir for the development and dissemination of AMR to human pathogens [4,6]. Consequently, environmental resistome surveillance studies highlight potential AMR evolution, but not all ARGs pose a risk to public health [6,24]. Nonetheless, resistome analysis provides a broad understanding of the complex evolutionary processes in AMR, identifying the genetic origin and mechanisms of dissemination. Resistome analysis may also allow us to alter our use of antibiotics to exploit synergistic combinations, enhance the bactericidal effect of existing antibiotics, and reduce selection pressure [25,26]. For example, the synergy between mecillinam and cefotaxime against extended-spectrum β-lactamase CTX-M-15-mediated resistance [27], where mecillinam promotes a resistance mutation that confers increased susceptibility to cefotaxime, suppressing further resistance evolution.

Structure of the resistome

The resistome consists of intrinsic and acquired (extrinsic) ARGs (Figure 1). Intrinsic ARGs are ancient and predate therapeutic use of antibiotics [28]. These are typically involved in basic bacterial physiological processes and generate resistance to an entire antibiotic class [6,29]. They are often universal within a bacterial species and largely independent of antibiotic selective pressure [29]. For example, intrinsic genes encode the outer membrane of Gram-negative bacteria, which is impermeable to glycopeptide antibiotics, rendering them resistant to glycopeptides, whereas Gram-positive bacteria do not have an outer membrane and are therefore generally susceptible [30].

Many intrinsic ARGs are inactive evolutionary precursors with inducible resistance function under antibiotic selective pressure, representing a large reservoir of AMR [3]. These cryptic resistance

Glossary

Binning: the clustering of contigs that originate from the same species.
Contig: a set of overlapping DNA segments that together contain a contiguous genomic region.
Fitness cost: reduced competitive ability of an organism to replicate and survive.
Horizontal gene transfer (HGT): the nonhereditary transmission of genetic material between bacterial organisms by transformation, transduction, or conjugation.
Integron: a genetic element that can capture and disseminate genes using a site-specific recombination mechanism.
K-mer: a nucleotide sequence of a certain length, for example, a dinucleotide is a k-mer where k = 2.
Metagenomic sequencing: a process of decoding the genetic material of all organisms within a sample.
Microbiome: the collective genetic material of the microbial community or commensal, symbiotic and pathogenic organisms within one ecological area.
Resistome: reservoir of antibiotic-resistance genes within a microbial niche.
genes are often embedded in bacterial chromosomes, expressed at a low level or not at all, and therefore may not result in AMR [1], for example, multidrug efflux pump gene *hmrM*, which is chromosomally encoded in *H. influenzae* [31].

Extrinsic ARGs can be acquired by *de novo* mutation or horizontal gene transfer (HGT). Bacteria have a remarkable ability to mobilise genetic material within genomes by transposition, and between genomes via conjugation, interorganism contact with DNA exchange; or transformation, the direct uptake of naked DNA, on plasmids or as linear DNA released from dead cells; or phage-mediated transduction [3].

**MGEs in AMR development**

ARGs in MGEs, the leading drivers of AMR evolution [32], pose a greater risk to human health than chromosomally embedded ARGs [5,6]. MGEs include transposons; entire resistance gene cassettes sandwiched between integrons; and integrative and conjugative elements (ICEs), which often contain virulence factors (e.g., antitoxin systems) promoting their vertical inheritance within the bacteria [4,33].

Although MGE-mediated genetic exchange is more common between less taxonomically related bacteria [4], a gene with AMR potential that is acquired through HGT is removed from its original regulatory and metabolic networks. This decontextualisation alters gene expression, and therefore function, leading to constitutive gene expression or upregulation, which can generate AMR [24]. Thus, genes with a ‘housekeeping’ role in their original bacterial setting (e.g., cell wall structural proteins) can become potent AMR determinants.

β-Lactam antibiotics are particularly vulnerable to MGE-mediated AMR. A study of the nasal resistome in infants with cystic fibrosis identified MGEs for over 30% of β-lactamase genes [34]. Long-term macrolide therapy results in an increase in transmissible plasmid-encoded *ermB* amongst viridans streptococci [35], a group of nasopharyngeal commensals that are known to be a source of MGEs for more pathogenic pneumococcal species [36].

---

**Figure 1. The role of antibiotic selective pressure on the evolution of the resistome.** The purple italicised text highlights the role of antibiotic selective pressure on each part of the resistome, as indicated by the purple dashed arrows. The yellow arrows illustrate the route of horizontal gene transfer (HGT) between environmental and human microbiomes. Abbreviations: ARGs, antibiotic-resistance genes; MGE, mobile genetic element.
**Antibiotic selective pressure**

When ARGs acquired by HGT provide a survival advantage, antibiotic use promotes resistome expansion, ensuring their persistence within the resistome [4,37,38]. Even subtherapeutic antibiotic concentrations, such as those found in environmental sources due to agricultural use, can modulate gene transcription, ensuring ARG persistence within a genome [3]. However, in the absence of antibiotic selective pressure, HGT events can be lost by genetic drift [37].

**The nasopharyngeal resistome**

There are limited studies of the nasopharyngeal resistome (Table 1), characterised by relatively small sample sizes and single sampling time points, preventing assessment of resistome evolution. Differing study populations, DNA extraction methods, and sequencing techniques make it challenging to compare their results.

Unfortunately, most of these studies do not analyse the nasopharyngeal microbiome, focussing instead on resistance determinants or streptococcal species, particularly *S. pneumoniae* [10,11,22,34]. *S. pneumoniae* readily acquires ARGs by HGT on MGEs, and from environmental DNA by transformation [39], particularly in the context of multiple serotype carriage, which enhances the opportunities for HGT [22]. Lack of microbiome analysis precludes identifying whether the ARGs identified are present in pathogenic bacteria, which is key to assessing their risk to human health. Furthermore, most studies published to date do not present an unbiased sample of the resistome; they use targeted techniques for sequencing macrolide resistance genes [10,11,40], or enrichment for streptococci prior to DNA extraction [22,41]. This creates the incorrect impression that the nasopharyngeal microbiome and resistome are dominated by *S. pneumoniae* [42]. Other pathogens in the nasopharynx exhibit AMR, including *H. influenzae* and *S. aureus* [43]. Importantly, microbiome diversity may facilitate genetic exchange of ARGs within a resistance reservoir [42].

Nonetheless, there are some interesting themes from these studies. Firstly, whilst the prevalence of ARG carriage in the nasopharynx is high, there are varying estimates of the prevalence of phenotypic resistance. Functional metagenomics identified a resistance phenotype to β-lactam antibiotics, tetracyclines, trimethoprim-sulfamethoxazole, and chloramphenicol in 53% of nasopharyngeal samples [34], whilst a study of pneumococcal carriage showed that just 52% of isolates carrying ARGs demonstrated phenotypic resistance to penicillins, and only 39% of macrolide-resistant isolates carried ARGs for this phenotype: *msrD*, *mefA*, or *ermB* [22]. This demonstrates the complexity of conferring phenotype based on ARG carriage. Secondly, whilst most ARGs are associated with streptococcal species, β-lactamase genes are predominantly found on non-streptococcal species, which may provide a source for HGT between species [41]. For example, functional metagenomics has identified β-lactamase genes *blaTEM-1* and *blaBRO-type* on *H. influenzae* and *Moraxella catarrhalis*, and *blaZ* on *S. aureus* and coagulase-negative staphylococci [34].

Thirdly, in keeping with other bacterial niches, colonisation of the nasopharynx in neonates includes ARG originating from environmental sources [40,44]. This is particularly relevant in hospitalised neonates, who can acquire nosocomial colonisers that contain ARGs on MGEs, with the potential for HGT to nasopharyngeal pathogens. For example, the *bro* β-lactamase gene is a common MGE in *M. catarrhalis*, with hypothesised origins in *Psychrobacter, Acinetobacter* and *Neisseria* species [45].

It is important to identify which ARGs produce a resistance phenotype as not all ARGs will be functional. Predictions of resistance phenotype based on sequence data alone are prone to false-positive bias [1]. For example, low concordance has been found between penicillin
Table 1. Summary of nasopharyngeal resistome studies

<table>
<thead>
<tr>
<th>Year of study</th>
<th>Sample type</th>
<th>DNA extraction and sequencing methods</th>
<th>Results/conclusions</th>
<th>Limitations</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>Nasopharyngeal swabs</td>
<td>Healthy versus diseased participants</td>
<td>- DNA extracted using enzymatic lysis.</td>
<td>- 48 ARGs, plus resistance-associated mutations were detected; but demonstrated poor correlation with phenotypic resistance.</td>
<td>- Enrichment for streptococci using Todd-Hewitt broth performed prior to DNA extraction.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy infants over first year of life</td>
<td>- Shotgun sequencing performed.</td>
<td>- ARG-ANNOT database used.</td>
<td>- No assessment of nasopharyngeal microbiome outside of pneumococcus.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy infants</td>
<td>- ARG-ANNOT database used.</td>
<td>- BLAST pairwise sequence alignment.</td>
<td>- Sample selection not representative of study population.</td>
</tr>
<tr>
<td>2019</td>
<td>Nasopharyngeal swabs</td>
<td>Infants with cystic fibrosis in first year of life</td>
<td>- DNA extracted using enzymatic lysis.</td>
<td>- ARG were detected in 64% of samples.</td>
<td>- <strong>Enrichment for streptococci using Todd-Hewitt broth performed prior to DNA extraction.</strong></td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Low-birthweight preterm neonates</td>
<td>- No host DNA depletion.</td>
<td>- Shotgun sequencing.</td>
<td>- Low stringency criteria of ≥50%.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy children from an intrapartum azithromycin randomised clinical trial</td>
<td>- Targeted sequencing of 16S rRNA gene and for ARGs on microbial DNA qPCR array for antibiotic resistance genes.</td>
<td>- Identified at least one resistant phenotype in 53% of samples.</td>
<td>- Enrichment for streptococci altered resistome composition.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy children aged 0–60 months</td>
<td>- Total DNA extracted for qPCR with a commercial kit, with adjustments for Gram-positive bacteria.</td>
<td>- Absence of ARGs in nasal resistome at birth, but ARGs detected after NICU admission and increased with prolonged hospitalisation.</td>
<td>- ARG-ANNOT database excludes ESBLs.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy infants</td>
<td>- No host DNA depletion.</td>
<td>- Targeted enrichment for macrocide resistance genes msrA, ermC with PCR and gel electrophoresis.</td>
<td>- Antibiotic treatment may not be major factor in early resistome development.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy children from an intrapartum azithromycin randomised clinical trial</td>
<td>- DNA extraction with enzymatic lysis.</td>
<td>- Intrapartum azithromycin increases carriage of macrolide-resistance genes msrA and ermC in the infant in the first month of life; this increase is not sustained at 1 year.</td>
<td>- Not a true metagenomic study; used two targeted sequencing techniques.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy children aged 0–60 months</td>
<td>- No host DNA depletion.</td>
<td>- Targeted enrichment for macrolide resistance genes msrA, ermC with PCR and gel electrophoresis.</td>
<td>- Small study size; 55 samples from 30 neonates over different time points.</td>
</tr>
<tr>
<td></td>
<td>Nasopharyngeal swabs</td>
<td>Healthy children from an intrapartum azithromycin randomised clinical trial</td>
<td>- DNA extraction using enzymatic lysis.</td>
<td>- Intrapartum azithromycin increases carriage of macrolide-resistance genes msrA and ermC</td>
<td>- No functional assessment of ARGs detected.</td>
</tr>
</tbody>
</table>

### Results/conclusions

- **Enrichment for streptococci using Todd-Hewitt broth performed prior to DNA extraction.**
- Low-stringency criteria of ≥50%.
- **Enrichment for streptococci altered resistome composition.**
- ARG-ANNOT database excludes ESBLs.
- Antibiotic treatment may not be major factor in early resistome development.
- Absence of ARGs in nasal resistome at birth, but ARGs detected after NICU admission and increased with prolonged hospitalisation.
- Most ARGs in the respiratory mucosa also detected in NICU environment, including β-lactam.
- Identification of antibiotic resistance genes in NICU environment, including β-lactam.
- Intrapartum azithromycin increases carriage of macrolide-resistance genes msrA and ermC in the infant in the first month of life; this increase is not sustained at 1 year.
- Increase in ermC and msrA prevalence at day 28 was associated with increased macrolide resistance in study population.
- Not a true metagenomic study; used two targeted sequencing techniques.
- Small study size; 55 samples from 30 neonates over different time points.
- No assessment of nasopharyngeal microbiome, and whether this affected prevalence of macrolide ARGs.
- Not a true metagenomic study.
- Small study size.
- Limited time points.
- No assessment of nasopharyngeal microbiome.
- Not a true metagenomic study.
- Limited time points.
- No assessment of nasopharyngeal microbiome.
- Not a true metagenomic study.
- Limited time points.
- No assessment of nasopharyngeal microbiome.
- Not a true metagenomic study.
- Limited time points.
susceptibility and ARG presence among strains of S. pneumoniae [22]. The genomic context of an ARG affects its ability to function as a resistance determinant. Proximity to promoter regions will influence ARG expression, colocation with an MGE will enable it to mobilise between species (for example, streptococcal species within the nasopharyngeal microbiome) and host compatibility will determine whether it functions as a resistance determinant [6,46]. However, it is also important to recognise the fallibility of using a Gram-negative heterologous host, favoured in functional metagenomic studies for ease of culture and vector transformation, to examine phenotypic function in the resistome of a niche dominated by Gram-positive organisms [34,47,48].

Although there are many pressures that can influence the nasopharyngeal microbiome, and therefore its resistome – for example age, season and environment [8] – a key theme amongst these studies is the transient effect of antibiotic selective pressure on the nasopharyngeal resistome. Sputum samples originating from the lung show a core respiratory resistome, dominated by macrolide resistance, which remains stable despite disease state and antibiotic pressure [49,50]. In contrast, the nasopharyngeal resistome appears to change with antibiotic pressure. Several studies show an increase in macrolide-associated ARGs, particularly amongst streptococcal and staphylococcal species, following prolonged macrolide exposure [10,11,35]. This change is transient [51,52], suggesting that wild-type susceptible bacteria reclaim the nasopharyngeal niche in the absence of continued antibiotic selective pressure, potentially due to the fitness cost of resistance carriage [52]. Some ARGs (mefA/E) persist longer than others [11] and are predominantly associated with streptococcal carriage [10,53]. Other ARGs (msr(A) and ermC) are carried by MGEs, indicating a potential for HGT [54], particularly among staphylococcal and enterococcal species [55].

Metagenomic sequencing technologies

Targeted and shotgun metagenomics

Metagenomic sequencing for resistome analysis involves extraction and sequencing all DNA from a biological sample, processing the resulting reads in a bioinformatic pipeline, and annotating reads based on homology to catalogued sequences in ARG databases. There are two main approaches to metagenomic sequencing [4]: functional and sequence-based, as summarised in Figure 2. These approaches are often preceded by depletion of host DNA, the methods for which are detailed elsewhere [20,56]. Sequence-based approaches can be further categorised into targeted-enrichment and shotgun techniques, although only the latter is a true metagenomic method [19].

Targeted-enrichment techniques selectively capture and enrich genomic regions of interest from a sample prior to sequencing [57]. This increases the proportion of reads of interest, providing better sensitivity for the target region, but limits sequencing breadth and does not represent the complete sample metagenome [19,57]. Targeted techniques cannot be used for novel ARG discovery [23], and are limited by the expense of generating target-specific probes and the time required to amplify genomic targets [57].

Shotgun sequencing is rapid, relatively unbiased and as a truly metagenomic technique can be used for novel ARG discovery [58]. There are challenges to using sequence-based approaches on clinical samples to derive the nasopharyngeal resistome (Table 2). As mentioned, nasopharyngeal swabs yield a low bacterial biomass, thus contaminants and host DNA can conceal bacterial signals, preventing pathogen and ARG detection [19,46]. Clinical sample storage, biological variation in GC content (affecting DNA thermal stability) and extraction methods can all introduce bias before sequencing [46]. Shotgun sequencing techniques produce short, fragmented reads...
that lack genomic context [58]. Direct analysis of these short reads can lead to incorrect mapping based on sequence homology and false-positive identification of ARGs [59]. Short reads require assembly into longer contigs to improve contextualisation of ARGs within the genome [23].
There are also universal challenges to sequencing the resistome that apply to the nasopharyngeal niche, for example functional validation of novel ARGs [1].

**Long-read sequencing to provide genomic context for resistome analysis**

Long-read sequencing could prove game-changing by improving genomic context in producing reads with better contiguity and more complete gene profiling [60], overcoming the challenge of assembling homopolymers. This could be key to uncovering functions encoded within large operons or ARG clusters [61] and more accurately detecting mutations, permitting detection of more novel ARGs [62]. Several long-read sequencing platforms are under development, two of which are currently widely used: single-molecule real-time sequencing (Pacific Biosciences) [62] and nanopore technology (Oxford Nanopore Technologies) [63]. Nanopore technology enables real-time base-calling, drastically reducing the sequencing time for a biological sample. However, both technologies are limited by throughput, cost, accuracy, and the processing limitations of current bioinformatic pipelines.

**The utility of functional metagenomics**

Functional metagenomics involves expression of a DNA sequence in a heterologous host to screen for a functional phenotype (Figure 2), bridging the divide between culture-based and genomic techniques, enabling identification of functional AMR from novel ARGs [23]. Functional metagenomics identifies only those ARGs with the capacity for HGT that are functional in a heterologous host – genes that are capable of transfer to, and expression in, pathogens [47]. Over the past decade, this technique has identified novel rifamycin phosphorylases [64] and tetracycline-inactivating enzymes [65]. Creation of functional metagenomic libraries, a process detailed elsewhere in the literature [47,66], could enable future proactive antibiotic development strategies by screening novel compounds for AMR [1,23,67].

The major limitations of functional metagenomics for the nasopharyngeal resistome are the low bacterial biomass of nasopharyngeal samples and the lack of phylogenetically diverse heterologous hosts [47,48]. Functional metagenomics can also miss ARGs arising from target mutations; if just one copy of the ARG with a target mutation is present, it can produce a dominant susceptible phenotype [1]. Functional metagenomics can also simulate HGT in otherwise immobile genetic elements (e.g., ARGs not colocated with an MGE) [1]. The limited insert size of a functional metagenomic library hinders its ability to identify combinatorial AMR mechanisms and MGEs, which have long genetic sequences [4]. Finally, comparison of functional screening techniques between laboratories is impeded by use of different media, incubation methods, and antibiotic concentrations [4]. Ultimately, this method has limited throughput, and is best combined with sequence-based techniques for resistome analysis [23]. In time, metatranscriptomics could replace current functional approaches [4].

**Key factors in bioinformatic resistome analysis**

**The challenges of assembling metagenomic data**

Metagenomic assembly is very challenging in low-biomass samples as the human genome is a thousand times larger than a bacterial genome, so can completely obscure low numbers of bacterial reads [19]. Metagenomic assembly is also complicated by uneven sequence coverage of multiple different organisms, with long stretches of identical sequences in unrelated species, making it difficult to assign reads to species. Algorithms attempt to account for this by using an optimised de Bruijn graph in which reads are divided into \textit{k-mers} that form a network graph; assemblers reconstruct the genome sequence by finding the optimum (Euler’s) path through this graph [59,68].

The plethora of software packages available for \textit{de novo} assembly of metagenomic data is potentially overwhelming [59]. Despite attempts to benchmark these programs [69], there is no consensus on the optimal assembly package [59]. The current approach is to use multiple
assembly tools to determine the best fit for a dataset; this is time-consuming and increases the already intensive computational demands of metagenomic assembly [59]. Metagenomic assembly typically requires hundreds of gigabytes of Random Access Memory, distributed across multiple processor cores and nodes in a cluster [46]. This requires considerable investment in supercomputing resources or cloud-based computing services with the data storage, management, and processing capabilities to cope with large volumes of data [4].

There are also specific challenges to assembling short reads for resistome analysis. HGT has led to sequence homology in unrelated species, and MGEs often contain homopolymers, long stretches of repeated sequences which are difficult to assemble [32,70]. Low-quality assembly can result in composite genomes that incorporate contigs from distinct populations, leading to flawed conclusions about microbial ecology [71]. This can be partially overcome by high-quality assembly with longer contigs, fewer gaps in the genome, and low error rates [32].

### Sequence alignment and identity cut-offs

ARGs are identified from assembled contigs or genomes by comparison to reference databases using similarity based alignment search tools [59]. However, whilst closely related ARGs often contain identical regions, they will not be identical across their entire sequence length and may be split onto several shorter contigs, resulting in contigs containing partial sequences from resistance determinants [46].

### Table 2. The challenges and potential solutions in metagenomic sequencing of bacterial samples for resistome analysis

<table>
<thead>
<tr>
<th>Stage in sequencing process</th>
<th>Challenges</th>
<th>Ways to address these challenges</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>DNA extraction for resistome metagenomics</td>
<td>▪ Ensuring equal lysis of Gram-positive and Gram-negative organisms for even species coverage within the resistome. ▪ Different DNA extraction kits with varying reagents, wash steps and incubation periods for enzymatic lysis can introduce bias.</td>
<td>▪ Choice of extraction technique: mechanical lysis improves overall yield, but shear DNA producing shorter reads that lack genomic context. Enzymatic lysis is not as effective for Gram-positive organisms, but produces longer reads. ▪ DNA extraction kits should be carefully selected for the intended sequencing technique, e.g., to optimise high-molecular-weight DNA for long-read sequencing.</td>
<td>[46,70,82,83]</td>
</tr>
<tr>
<td>DNA extraction from low-biomass samples</td>
<td>▪ Overwhelmed by human host DNA, which conceals bacterial signals, preventing detection of clinical pathogens and ARGs. ▪ Obtaining sufficient DNA volumes for sequencing.</td>
<td>▪ Depletion of human host DNA prior to or after sequencing improves bacterial DNA resolution, but both reduce overall DNA yield, and the latter is computationally intensive. ▪ Ultra-deep sequencing, which can be prohibitively expensive. ▪ Use of targeted amplification sequencing techniques e.g., multiplex PCR for 16S rDNA.</td>
<td>[19,20,46,63,84]</td>
</tr>
<tr>
<td>Sequencing the bacterial resistome</td>
<td>▪ High variation between samples requires multiple replications, which can be prohibitively expensive. ▪ Uneven species coverage in diverse resistome.</td>
<td>▪ No solution: remains a difficult balance between sufficient sequencing depth and number of samples that can be replicated. ▪ ‘Binning’ of coabundant species: sequences are classified into separate ‘bins’ (e.g., species) based on DNA composition patterns of the genomes (e.g., tetramer frequencies). This reduces assembly complexity and targets regions of interest.</td>
<td>[46,76,85]</td>
</tr>
</tbody>
</table>
Table 3. A comparison of currently available ARG databases

<table>
<thead>
<tr>
<th>ARG databases</th>
<th>Updated/archived</th>
<th>Curated</th>
<th>Link</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generalised databases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antibiotic Resistance Gene Database (ARDB)</td>
<td>Archived</td>
<td>Yes</td>
<td><a href="http://ardb.cbcb.umd.edu/">http://ardb.cbcb.umd.edu/</a></td>
<td>• Uses BLAST to identify and annotate ARGs.</td>
<td>• The original ARG database, which was manually curated.</td>
<td>• Does not distinguish between putative ARGs based on homology and functionally confirmed ARGs.</td>
<td>[46,59,86]</td>
</tr>
<tr>
<td>ResFinder</td>
<td>Updated</td>
<td>Yes</td>
<td><a href="https://cge.cbs.dtu.dk/services/ResFinder/">https://cge.cbs.dtu.dk/services/ResFinder/</a></td>
<td>• Combines ARDB, Bush-Jacoby Beta-Lactamase list and other published ARG sequences.</td>
<td>• Allows user to adjust the identify and length coverage thresholds.</td>
<td>• Unable to classify chromosomal mutations causing antibiotic resistance or novel ARGs.</td>
<td>[46,68–90]</td>
</tr>
<tr>
<td>SARG (version 2)</td>
<td>Updated</td>
<td></td>
<td><a href="http://smile.hku.hk/SARGs">http://smile.hku.hk/SARGs</a></td>
<td>• Hierarchical structured database derived from ARDB, CARD and NCBI--NR database.</td>
<td>• Comprehensive database with lower identity matching for ARG annotation of metagenomic sequence data.</td>
<td>None identified.</td>
<td>[59,91]</td>
</tr>
<tr>
<td>Comprehensive Antibiotic Resistance Database (CARD)</td>
<td>Updated</td>
<td>Yes</td>
<td><a href="https://card.mcmaster.ca/">https://card.mcmaster.ca/</a></td>
<td>• Accompanied by a BLAST-based pairwise sequence alignment tool.</td>
<td>• The most comprehensive resource for ARG available.</td>
<td>• Poor at detection of point mutations in chromosomal target genes known to be associated with ARGs.</td>
<td>[4,46,76,89,92]</td>
</tr>
</tbody>
</table>
Table 3. (continued)

<table>
<thead>
<tr>
<th>ARG databases</th>
<th>Updated/archived</th>
<th>Curated</th>
<th>Link</th>
<th>Description</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Refs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ResfinderFG</td>
<td>Updated; last November 2016</td>
<td>Yes</td>
<td><a href="https://cge.cbs.dtu.dk/services/ResFinderFG/">https://cge.cbs.dtu.dk/services/ResFinderFG/</a></td>
<td>ARG variants identified in functional metagenomics studies. • Built by aggregating data from four functional metagenomics studies selected against 23 antibiotics.</td>
<td>• Functionally validated ARGs.</td>
<td>• Less comprehensive list of ARGs (only functionally validated), which may limit detection of novel resistance determinants.</td>
<td>[59,93]</td>
</tr>
<tr>
<td>Resqu</td>
<td>Yes</td>
<td>Yes</td>
<td><a href="https://www.1928diagnostics.com/resdb/">https://www.1928diagnostics.com/resdb/</a></td>
<td>Only contains ARGs with experimentally verified function and evidence of HGT between species.</td>
<td>• More stringent approach to identifying ARGs; the database only contains sequences of acquired ARGs present on MGEs. • DNA sequences included in the database are manually assessed for errors, nonredundant and consistent with established gene families.</td>
<td>• Consistently reports lowest numbers of ARGs. • Less comprehensive list of ARGs (only functionally validated), which may limit detection of novel resistance determinants.</td>
<td>[46,94]</td>
</tr>
<tr>
<td>Resfams</td>
<td>Updated; last in January 2015</td>
<td>Yes</td>
<td><a href="http://www.dantaslab.org/resfams/">http://www.dantaslab.org/resfams/</a></td>
<td>Database of HMM profiles built on resistance proteins compiled from the CARD, the LacED and Jacoby and Bush’s collection of curated β-lactamases, based on the gene ontology from the CARD database. Curated database for ARGs with confirmed antibiotic resistance function.</td>
<td>• HMMs identify novel ARGs with greater precision and accuracy than pairwise annotation. • These HMMs include a broader range of ARGs identified through culture-independent functional metagenomic selection. • HMM profiles are confirmed for functional resistance [1].</td>
<td>• Less comprehensive list of ARGs (only functionally validated), which may limit detection of novel resistance determinants.</td>
<td>[6,23,59,74]</td>
</tr>
<tr>
<td>Database Name</td>
<td>Updated Date</td>
<td>Available</td>
<td>Website</td>
<td>Description</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------------------------------------------------</td>
<td>-----------------------</td>
<td>-----------</td>
<td>----------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- Includes protein--coding antibiotic resistance sequences, regulatory elements, MGE and predicted proteins flanking ARGs.  
- Less comprehensive list of ARGs (only functionally validated), which may limit detection of novel resistance determinants.  |
| Mustard                                           | Updated; last in November 2018 | Yes       | http://mgps.eu/Mustard/                     | - Contains >6000 resistance determinants from 20 families.  
- Includes a curated set of ARG identified in functional metagenomics studies.  
- Functional validation of all ARGs included in the database.  
- Less comprehensive list of ARGs (only functionally validated), which may limit detection of novel resistance determinants. |
| Beta-lactamase Database (BLDB)                    | Updated; last June 2022 | Yes       | http://blpdb.eu/                            | - Compiles ARG sequence information, and biochemical and structural data for all currently known β-lactamases.  
- Limited to a defined subset of ARGs. |
- One of the first databases to compile known β-lactamases and assign nomenclature to new ones.  
- Focussed on compiling a list of a small subset of ARGs. |
- This database constructs protein families, using DWARF to integrate protein sequence and structure information.  
- Includes methods to predict TEM, SHV, and class B enzymes by merging information from both the NCBI databases and TEM mutation table.  
- Provides information on mutations, sequences and structures of TEM and SHV β-lactamases.  
- Limited to a defined subset of ARGs. |
- Uses proximity/synteny of MGEs with ARGs to form cassettes.  
- First database to simultaneously automate identification of ARGs and place them in broader genetic context.  
- Gene names differ from multiple other nomenclatures.  
- Limited to a defined subset of ARGs. |

Specialized databases – provide comprehensive information for specific gene families or species

<table>
<thead>
<tr>
<th>Database Name</th>
<th>Updated Date</th>
<th>Available</th>
<th>Website</th>
<th>Description</th>
</tr>
</thead>
</table>
| Beta-lactamase Database (BLDB)                    | Updated; last June 2022 | Yes       | http://blpdb.eu/                            | - Compiles ARG sequence information, and biochemical and structural data for all currently known β-lactamases.  
- Limited to a defined subset of ARGs. |
- One of the first databases to compile known β-lactamases and assign nomenclature to new ones.  
- Focussed on compiling a list of a small subset of ARGs. |
- This database constructs protein families, using DWARF to integrate protein sequence and structure information.  
- Includes methods to predict TEM, SHV, and class B enzymes by merging information from both the NCBI databases and TEM mutation table.  
- Provides information on mutations, sequences and structures of TEM and SHV β-lactamases.  
- Limited to a defined subset of ARGs. |
- Uses proximity/synteny of MGEs with ARGs to form cassettes.  
- First database to simultaneously automate identification of ARGs and place them in broader genetic context.  
- Gene names differ from multiple other nomenclatures.  
- Limited to a defined subset of ARGs. |
Basic Local Alignment Search Tool (BLAST) pairwise sequence alignment [72,73] is the most commonly used alignment approach [59] but is biased by the comparatively over-sampled human microbiome [74]. One alternative method is multiple sequence alignment using probabilistic prediction algorithms based on Hidden Markov Models (HMMs), enabling identification of a more diverse range of ARGs [23,74] with greater precision and accuracy than BLAST [75]. However, both BLAST and HMM statistics lack fine resolution, prompting difficult decisions regarding sequence identity cut-offs to prevent false-positive predictions [76]. Each technique can lead to differing and sometimes contradictory conclusions [74].

Cut-offs for alignment identity, score, and coverage breadth are important parameters to decide prior to ARG annotation, as they affect the number and type of ARGs detected [32]. Stringent sequence identity cut-offs avoid over-classification of ARGs, but the challenge is defining what is appropriate. Metagenomic reads will contain some sequencing errors, and not all genes with similar functions will have identical sequences. It is reasonable to allow some mismatch between read and reference sequences, but the degree of stringency in these mismatches affects the validity of results [46]. Accommodating mismatches when mapping short-read data also consumes computational resources [46]. Functional metagenomic studies help to validate identity cut-offs, but reads only represent part of a genetic sequence, so even if a read is 100% identical to an ARG it may not be functional [46]. The current solution for many resistome studies is a sequence identity cut off of 80–95%, a compromise between efficiency, sensitivity, and stringency of data analysis [46]. Some ARG databases have corresponding software packages for ARG annotation that provide gene-family specific alignment cut-offs [34].

ARG databases
There are multiple ARG reference databases (Table 3), compiled from studies of genetic determinants of AMR [59]. Databases that selectively include ARGs on MGEs identify resistance determinants most at risk of HGT, but miss novel and emerging ARGs [46]. Conversely, poorly curated databases that include chromosomal point mutations and putative ARGs, which lack functional validation, over-estimate the abundance and diversity of resistance determinants [6,46].

Some chromosomal point mutations in specific genes can produce AMR, for example, rpoB, but reads from wild-type genes will also map to these sequences due to their similarity. Many are ubiquitously occurring essential genes that dilute a database (e.g., PointFinder [77] and AMRFinder [78]), unless they are proven to confer functional resistance [32,46]. Therefore, a priority for the field is the functional validation of ARGs before database inclusion.

Another limitation of current ARG databases is the exclusion of environmental and nonclinical ARGs, limiting the ability to detect emerging resistance mechanisms in surveillance studies [74]. Many ARG databases also catalogue resistance determinants in a manner that ignores their taxonomic context, disregarding the host-specific functionality of some ARGs [76]; for example, ARGs for certain antibiotics in Gram-positive organisms may be irrelevant in Gram-negative organisms that are intrinsically resistant to those antibiotics, as they cannot penetrate their outer membrane. To address this, ARG databases will need to include species-specific risk estimations for functional resistance and HGT of ARGs [23].

Database curation
The challenge of ARG database curation impedes global data sharing and multinational surveillance of emerging AMR [63,76]. The best ARG databases are regularly updated with novel, functionally validated resistance determinants [1,23], but this requires manual curation and linking of complex metadata on resistance profiles and bacterial epidemiology [76]. A more automated method of database curation, which could operate through automatic incorporation of parameters from published data, is therefore urgently required if resistome analysis is to fulfil its potential.
Concluding remarks and future perspectives

AMR is placing an increasingly intolerable clinical and economic burden on global health systems, threatening a public health crisis on a scale that exceeds recent pandemics. Defining the nasopharyngeal resistome in human populations will improve AMR surveillance and enable the discovery of new ARGs and resistance mechanisms. It will also inform response strategies, including the repurposing of existing antibiotics into synergistic combinations, development of regimens that induce collateral sensitivity, and screening of novel compounds for resistance with functional metagenomic libraries.

The nasopharynx is an important, understudied, reservoir for emergence and spread of antibiotic-resistant bacteria that cause invasive disease. There are few studies of the nasopharyngeal resistome due to the challenges of characterising a resistome with a low sample biomass and high burden of human host DNA, alongside the general challenges of resistome characterisation, which include determining genetic context and phenotypic expression of ARGs, differing bioinformatic approaches and limited computational resources [63,79].

Advances in shotgun and long-read metagenomic sequencing techniques are providing tools to overcome these challenges, but require further development to improve reliability. ARG databases require optimisation to maximise their potential for AMR surveillance, including automated metadata curation and functional validation of all catalogued ARGs.

Current sequencing technologies could be used to characterise nasopharyngeal resistome evolution under specific selective pressures, for example antibiotic mass drug administration or change in pneumococcal vaccination regimen. However, for future nasopharyngeal resistome studies to fully inform preventative public health policy, for example the impact of mass drug administration and vaccination strategies on the nasopharyngeal resistome in children in high-disease-burden settings [80,81], there needs to be better linkage between ARGs and their bacterial host, and improved resolution of an individual’s resistome from a single sample (see Outstanding questions).

Acknowledgments

LO’C is supported by a British Infection Association (BIA) small project grant. R.S.H is supported by the National Institute for Health Research (NIHR) Global Health Research Unit on Mucosal Pathogens using UK aid from the UK Government. The views expressed in this publication are those of the author(s) and not necessarily those of the NIHR, BIA or the Department of Health and Social Care.

Declaration of interests

No interests are declared.

References

33. Nielsen, T.K. et al. (2022) Antibiotic resistance genes are differentially mobilized according to resistance mechanism. Science 373, eabg0072
36. Allemann, A. et al. (2022) Strepococcal pneumoniae sequences that frequently colonize the human nasopharynx are common recipients of penicillin-binding protein gene fragments from streptococcal mites. Microb. Genom. 7
40. Casen, C. et al. (2021) Microbial contamination in hospital environment has the potential to colonize preterm newborns’ nasal cavities. Pathogens (Basel) 10, 615
56. Reijr, P. et al. (2022) Microbial DNA extraction of high-host content and low biomass samples: optimized protocol for nasopharynx metagenomic studies. Front. Microbiol. 13, 1038120


