ANALYTICAL FRAMEWORK FOR ESTIMATING ANTIMICROBIAL RESISTANCE GENE ABUNDANCE IN METAGENOMIC SAMPLES OF ANIMAL AGRICULTURE ORIGIN

By

Leland K. Ackerson IV

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Animal Science – Master of Science

2023

ABSTRACT

Antimicrobial resistance (AMR) has become an apex global public health threat that requires a multifaceted One Health approach. According to the CDC, 2.8 million antimicrobial resistant infections occur in the United States each year, resulting in more than 35,000 deaths. Although the development of AMR is incredibly intricate, it is widely recognized that the employment of antibiotics is one of the largest selective pressures of AMR. In many countries, antimicrobial consumption in animal agriculture surpasses that of human usage, and it is estimated that nearly 73% of global antibiotics can be attributed to livestock. Monitoring AMR emergence and historical data on a global scale is crucial when working towards the large-scale mitigation of this public health threat. One tool that can contribute to monitoring AMR is shotgun metagenomics, which entails comprehensive evaluation of the genetic material extracted from all the organisms in a complex sample. This subsequently gives genomic insights into the microorganisms residing in the sample of interest. The Sequence Read Archive (SRA) is a public repository housed by the National Center for Biotechnology Information (NCBI) containing extensive sequence data from metagenomic samples in animal agriculture, as well as the associated spatiotemporal attributes. Here we proposed to develop analytical framework to leverage the SRA and estimate relative antimicrobial resistant gene abundances across animal agriculture on a global scale from publicly available metagenomic sequence information. The developed analytical framework was then employed to evaluate metagenomic samples from cattle and swine housed in the SRA. Estimated abundances are utilized as a proof of concept for evaluating AMR characteristics on a global scale using publicly available, highly heterogenous data. The resulting abundance estimation will offer insights into AMR emergence and dynamics as well as inform further development of mitigation strategies.

This thesis is dedicated to family, the
most important thing in this world.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude and appreciation to the individuals who have played significant roles in the completion of this thesis and my development as a scientist. Their unwavering support, guidance, and encouragement have been invaluable throughout this challenging journey.

First and foremost, I would like to thank Dr. Wen Huang for going above and beyond both my expectations and his responsibilities as a mentor. Dr. Huang is a remarkable scientist and researcher who genuinely cares for each and every individual in his lab. From day one, I have appreciated Dr. Huang's mentoring philosophy which is embodied by the ancient quote "*Give a man a fish, and you feed him for a day. Teach a man to fish, and you feed him for a lifetime.*". As a young researcher, I feel that learning how to think critically and developing the skills to independently find answers to difficult questions is vital to my success, both of which are encouraged by the aforementioned mentoring philosophy. I have received, and retained, a myriad of invaluable advice and guidance from Dr. Huang in the past couple years; all of which had my best interest in mind. I am incredibly fortunate to have Dr. Huang as my mentor, and I am ecstatic to return to his lab in pursuit of my PhD following the conferral of my MS.

I would also like to thank the rest of my guidance committee members, Dr. Cedric Gondro, Dr. Pamela Ruegg, and Dr. Paul Coussens for their advice and support throughout the progression of the project. Each of them served an integral and unique role on my well-rounded guidance committee which made work efficient and enjoyable. I was fortunate enough to take multiple classes with Dr. Gondro, and we could often be found catching up and discussing new ideas in his office. This greatly influenced my positive attitude towards research and education despite any discouraging hardships encountered along the way. Dr. Ruegg and Dr. Coussens were both always more than willing to offer a helping hand, and with their expertise in their respective fields acting as guidance,

my research and aspirations seemed drastically less daunting. Thank you all for your dedication to my research and as mentors on my guidance committee.

Additionally, I am very appreciative of Samuel Snowden – his contributions regarding sample annotation were instrumental in the timely completion of this project. Moreover, many thanks to my close friends and fellow lab members for their consistent moral support throughout my academic endeavors.

I would also like to thank my wonderful girlfriend, Susan Hoffman, for being by my side every step of the way. Thank you for being my rock and for always pushing me to reach new heights. I am extremely grateful to have you in my life.

Finally, I would like to thank my mother and father, Angela Ackerson, and Leland Ackerson III. I would not be where I am today without each of you and your unwavering support. Your words of encouragement, reassurance, and belief in my potential have been instrumental in my academic success and development as an individual. These kind words fail to do my appreciation for you both justice, but from the bottom of my heart – thank you!

TABLE OF CONTENTS

Chapter 1. Background and significance

1.1 Antimicrobial resistance is a global public health threat

Antimicrobial resistance (AMR) has become a serious global public health threat. AMR can be partitioned into two types of resistance: intrinsic and acquired. Intrinsic AMR can be defined as a universally shared resistance within a microbial species that is independent of previous antimicrobial exposure. Alternatively, acquired AMR is defined as the ability of a disease-causing microbe to survive exposure to an antimicrobial agent that was previously an effective treatment. As a result, treating infections caused by microorganisms such as bacteria, viruses, fungi, and parasites gradually becomes increasingly impracticable. The Center for Disease Control (CDC) and the World Health Organization (WHO) have both determined that drug resistance to potentially disease-causing pathogens is amongst the top global health security risks. The WHO in particular has declared that AMR exists within the top 10 global public health threats facing humanity. Predictive statistical models have recently estimated that in the year 2019 alone, 4.95 million deaths globally were associated with bacterial AMR; of which, 1.27 million deaths could be directly attributed to bacterial AMR (Antimicrobial Resistance Collaborators, 2022). Alarmingly, in 2016 a government funded review on the risks associated with drug-resistant infections estimated that annual global deaths associated with AMR could spike to 10 million by the year 2050 (O'Neill, 2016; Antimicrobial Resistance Collaborators, 2022).

While the global burden is shared internationally, there are regional disparities. In 2019, estimated deaths attributable to bacterial AMR was found to be highest in western Sub-Saharan Africa and lowest in Australasia with rates of 27.3 and 6.5 per 100,000 deaths, respectively (Antimicrobial Resistance Collaborators, 2022), reinforcing consistent findings within the literature that the burden of AMR is more prevalent in low-income and low-resource geographical locations.

Nonetheless, the impact of this health concern is far from obsolete in high-income areas. In the United States of America, 2.8 million antimicrobial resistant infections occur each year, resulting in more than 35,000 deaths (CDC, 2019). It is abundantly clear that without the rapid implementation of comprehensive steps towards mitigation, the risks associated with AMR will continue to rise.

1.2 Risks and impact of antimicrobial resistance

The risks and impact of widespread antimicrobial resistance cannot be overstated. The list of directly and indirectly associated risks is extensive and in part includes a) increased mortality, b) economic burdens, c) reduced effectiveness of medical procedures, and d) diminished food security. Increased mortality can result from multiple factors in and of itself, but the primary contributor is that of the loss of effective treatments. Regarding economic burdens there are a multitude of contributors. Individuals with AMR infections may require more time intensive and specialized treatment regimen thus inducing extended hospital stays and consequently higher healthcare costs. Additionally, there are increased costs associated with the development of new antimicrobial agents which subsequently can lead to higher drug and healthcare costs. According to a report by the CDC in 2019, it is estimated that AMR costs the United States $55 billion annually; this can be partitioned into costs associated with healthcare and loss of productivity, surmounting to $20 billion and $35 billion, respectively (Dadgostar, 2019). The decrease in efficacy of antimicrobials and or limitations imposed on their use due to prevalent AMR in animal agriculture would greatly contribute to economic burden as a result of a variety factors such as reduced production, increased livestock mortality and or cull rates, and treatment costs, among others. Furthermore, the use of antimicrobials in animal agriculture can lead to the proliferation of resistant bacteria with potential to enter the food chain, thus subjecting human individuals to additional sources of AMR transmission and compromising food safety (Davies and Wales, 2019). Diminished food security in regard to food availability is an additional concern. The global population is continuously expanding,

demanding increased food production to mitigate malnutrition, and rampant AMR development in animal agriculture could reduce production leading to food shortages.

One indirect risk of AMR development is the potential reduced effectiveness of medical procedures that are highly dependent on the use of antimicrobials. Surgeries such as organ transplantations, joint replacement, and caesarean sections, as well as treatments such as chemotherapy could experience lower success and survivability rates or even become too risky to perform solely as a result of the inefficacy of the antimicrobials for which they are reliant upon (O'Neill, 2016). Therefore, AMR can essentially undermine recent developments in contemporary medicine and handicap medical professionals when developing courses of action for patients experiencing various health related issues. It is abundantly clear that there are a plethora of risks and negative impacts of a large magnitude associated with AMR development and emergence. Additionally, the known list of factors influenced by AMR is arguably incomplete as unknown factors and implications may come to fruition as the severity and comprehensive understanding of this global health concern develops further.

1.3 Emergence of AMR

Although the development of AMR is incredibly intricate, it is widely recognized that the employment of antibiotics is one of the largest selective pressures of AMR. When used appropriately and precisely, antimicrobial utilization can effectively terminate or inhibit microorganisms lacking defense mechanisms; however, microbes containing the intrinsic or acquired mechanisms to resist the action of antimicrobials are able to persist and reproduce. Thus, microbes with the ability to resist antimicrobials have a higher relative fitness than their counterparts when exposed to an antimicrobial rich environment, and over time the resistance phenotype and underlying mutations will become more prevalent in the population. The functional mechanisms of antimicrobial resistance result from the expression of encoded genomic material referred to as antibiotic resistant

genes (ARGs). Antibiotic resistant genes are either produced by means of genetic mutations or acquired via gene transfer. The random and spontaneous occurrences of genomic mutations carry the potential to alter functionality of synthesized proteins resulting from antecedent gene expression (Schwarz et al., 2016). If said mutation is nonsynonymous, and therefore induces alterations in both the genetic and amino acid sequences of the encoded protein, subsequent protein functionality could be altered in terms of efficiency, dissolved completely (loss of function), or even redefined. In particular, the redefinition of protein function can result in the acquisition of antimicrobial resistant mechanisms by microbes. Mutations conferring subsequent antimicrobial resistance are often found in regions associated with the target sites of antimicrobial agents as this is a focal point for a variety of resistance mechanisms. In addition, mutations can further induce resistant phenotypes or lessened vulnerability to antimicrobials as a result of the enhanced expression for separate genes such as those pertaining to efflux.

Alternatively, antibiotic resistant genes can be inherited by the actions of vertical gene transfer (VGT) and or horizontal gene transfer (HGT). Primarily, genetic material is passed down from parent to offspring in a vertical process. Resulting in the transmission of various genes and developed mutations to the inheriting organisms. However, genomic material can also be acquired by means of horizontal gene transfer, a process by which genetic material is transferred between distinct evolutionary lineages (Thomas, 2005; Stokes, 2011; Dunning Hotopp, 2011; Soucy, 2015). HGT is relatively common amongst microbes, and typically manifests between organisms of the same domain (archaea, bacteria, eukarya). HGT is most frequently documented within the bacteria domain. Nonetheless, interdomain HGT such as acquisition by eukaryotes from bacteria also occurs (Dunning Hotopp, 2011). This is exemplified by organelles such as mitochondria and chloroplast transmitting DNA to the nuclear genome of eukaryotes. Mitochondrial transmission in particular is well documented in *Arabidopsis thaliana* (Lin et al., 1999). Regardless of the organisms' residential

domains, each case of HGT can result in the acquisition of novel genes and traits independent of those received via VGT inheritance.

The process of horizontal gene transfer can be split into two independent processes: the transfer of the DNA, and incorporation of the DNA into the recipient organisms' genome (Stokes, 2011). There are several mechanisms that can facilitate the horizontal transfer of DNA and mobile elements, but the three most common and primary mechanisms are conjugation, transformation, and transduction. Mobile elements pertain mostly to plasmids and transposons, plasmids are small circular DNA strands and transposons are genetic elements that consist of repeat sequences and a transposase protein (Thomas, 2005). Mobile genetic elements (MGEs) are capable of carrying and disseminating ARGs. Conjugation entails transmission via physical contact of the donor and recipient bacterium. A pilus is formed to connect the cells, and then utilized as a medium for the exchange of genetic material. The uptake of environmental DNA is classified as transformation and is most common in bacteria and archaea. Transduction is the transfer of genetic material by means of phages and can be both general and specialized. General transduction is classified as the intake of a random segment of the host DNA involving a lysed bacterial cell, whereas specialized transduction is performed by temperate bacteriophages without the utilization of lysis. In specialized transduction, viral DNA is integrated with the host DNA and exists in a prophage stage (Soucy, 2015). There are alternative mechanisms of horizontal gene transfer such as cell fusion, gene transfer agents, and intercellular transfer, but these are uncommon compared to the aforementioned mechanisms. After the genetic material is transferred, it must be stably incorporated into the recipient genome. The mechanisms for incorporation are primarily autonomous replication, transposition, homologous recombination, and site-specific recombination (Stokes, 2011). Autonomous replication pertains primarily to plasmids, whereas transposition pertain mostly to transposons, and site-specific recombination involves integrons. Integrons are assembly platforms

used to produce functional genes and are considered mobile elements that can be associated with transposons and plasmids. Each of the discussed mobile elements can be utilized in parallel with homologous recombination to integrate the new genetic information into the host (Stokes, 2011; van Hoek et al., 2011). Summarized and simply stated, both VGT and HGT present means by which antimicrobial resistant gene uptake is facilitated.

1.4 Classification and mechanisms of antimicrobials

Antimicrobial agents can be classified in a variety of ways. At the highest tier, antimicrobials can be partitioned into groups based on the microorganism they impact such as bacteria (antibiotics), fungi (antifungal), viruses (antiviral), and parasites (antiparasitic). We focus in this review on antibiotics as they are utilized more frequently than the others in human medicine and animal agriculture. It should be noted that the term antibiotic is commonly used to refer to both antibacterial and antifungal drugs, as well as used interchangeably with the term antimicrobial despite not conforming to the exact definitions. Within the faction of antibiotics, agents can be further classified by their type of action: bactericidal and bacteriostatic. Bactericidal refers to antibiotics that kill bacteria, whereas bacteriostatic encompasses antibiotics that inhibit bacterial growth and stall cellular activity without directly killing the bacteria. Alternatively, spectrum-based classification can cluster antibiotics into groups of narrow, broad, or extended spectrum. Spectrum is defined as the specificity and range of microorganism that an antibiotic can negatively impact. Narrow spectrum antibiotics only target a limited number of bacterial species. Extended spectrum refers to antibiotics that have been chemically modified to affect a broader range of bacteria. Broad spectrum antibiotics have an even larger scope and can therefore affect a variety of species and types of bacteria such as both Gram positive and Gram negative. In a clinical setting when treating infections, typically broad-spectrum antibiotics are used when the origin of an infection is unknown and narrow spectrum is utilized when the origin has been isolated or refined to a small group of possibilities with high

confidence. Of note, the employment of broad-spectrum antibiotics is more likely to promote antimicrobial resistance and multidrug resistance in bacterial populations as well as disrupt the microbiome by altering the intestinal flora (Gerber et al., 2018; Grada & Bunick, 2021). Multidrug resistance (MDR) characterizes microbials that are resistant to multiple antimicrobial agents, this occurrence makes clinical treatment or corresponding infections more difficult (Alekshun & Levy, 2007). Nonetheless, most commonly antibiotics are classified in terms of their mechanism of action or their chemical structure.

Classification in terms of chemical structure produces categories that antimicrobials are typically referred to by such as beta-lactams, aminoglycosides, tetracyclines, macrolides and more. For example, beta-lactams are characterized by the presence of a beta-lactam ring in their chemical structure and macrolides are all comprised of a 14, 15, or 16-membered ring. Tetracyclines, the most heavily utilized antimicrobial for animal agriculture domestically in the United States which accounts for 65% of sales, are identifiable by four adjacent cyclic hydrocarbon rings (FDA, 2019). Biologically speaking, the identity of molecular structures directly equates to the function of the molecule. Thus, antimicrobials classified into the same groups by chemical structure primarily exhibit similar function, mechanism of action, and or molecular mechanisms.

Antibiotic mechanisms of action, as shown in Figure 1, primarily fall within three factions: cell wall synthesis, protein synthesis, and nucleic acid synthesis (Kapoor et al., 2017). The cell wall synthesis category pertains to antibiotics that target the cell wall and inhibit biosynthesis. Beta-lactams and glycopeptides are two antibiotic classes that fall within the confines of this mechanism of action. Beta-lactams, and more specifically penicillin, target penicillin-binding proteins and block the production of peptidoglycans which are a significant component of the cell wall. The inhibition of this production results in the weakening of the cell wall and leads to subsequent lysis of the bacteria. Glycopeptides, vancomycin in particular, also inhibits cell wall synthesis associated with

penicillin-binding proteins. However, vancomycin binds to the side chain of peptides that normally bind to penicillin-binding proteins as opposed to the binding proteins themselves, subsequently blocking the binding of the peptide to the binding proteins and peptidoglycan production, achieving cell wall synthesis inhibition. The second mechanism of action, pertaining to inhibition of protein synthesis, is carried out by targeting ribosomal subunits necessary for the translation of RNA into proteins. Antimicrobial classes associated with this mechanism of action include but are not limited to aminoglycosides, tetracyclines, and macrolides which inhibit the 30S, 30S, and 50S subunits of ribosomes in bacteria, respectively. As this mechanism occurs within the cell, antimicrobials residing in this classification are aided by the aforementioned antimicrobials that target the cell wall since they allow easier access into the cell. Nonetheless, the presence of cell wall synthesis inhibitors is not required for the effectiveness of protein synthesis inhibitors. Aminoglycosides are positively charged molecules capable of forming pores in the cellular membrane to allow entrance of the antibiotic. Aminoglycosides inhibit protein synthesis by interacting with the 16S rRNA 30S subunit to induce errors in mRNA translation. 16S rRNA comprises the 30S ribosomal subunit and is required for protein synthesis and the correct codon-anticodon pairing during mRNA translation. Notably, the genes that encode 16S rRNA exist in all bacteria and are both highly conserved and species specific. As a result of these characteristics, 16S rRNA genes are commonly used in phylogenetic studies and for microbial identification, classification, and quantification. Nonetheless, the aminoglycoside induced errors are primarily associated with misreading the mRNA and the untimely termination of the process of translation subsequently resulting in cell death. Tetracyclines accomplish protein synthesis inhibition by directly binding with the 16S rRNA 30S subunit, therefore blocking association with tRNA and preventing a key step in producing the translated amino acid sequences. Migrating from 30S subunit inhibitors to a 50S subunit inhibitor, macrolides target conserved sequences in 23S rRNA resulting in the production of truncated peptide chains. Finally,

antimicrobial agent mechanisms of action can be characterized by inhibitors of nucleic acid synthesis. Representatives of this faction are quinolones and fluoroquinolones, which inhibit DNA gyrase and topoisomerase IV thereby negatively influencing DNA replication. DNA gyrase is an enzyme association with the relaxation of positive supercoils in DNA to allow for the progression of DNA replication by DNA polymerase and topoisomerase IV is primarily involved with separating daughter DNA strands immediately following replication. Inhibition of nucleic acid synthesis regarding either DNA replication or transcription is detrimental to microbial health and is the reason that quinolones are commonly bactericidal.

The final criterion for antimicrobial classification is in terms of medical importance (Gelalcha & Kerro Dego, 2022; FDA, 2023). There are two subclasses associated with medical importance, antimicrobials deemed medically important, and those that are not medically important. Within the medically important antimicrobials (MIAs) faction there are three additional subclasses, critically important antimicrobials (CIAs), highly important antimicrobials (HIAs), and important antimicrobials (IAs). The more important the antimicrobial, the larger the human health risk associated with potential resistance development (FDA, 2023). The ranking criteria used to classify antimicrobials into one of the three classes of importance are as follows. Antimicrobial drugs that are unique or exist within a limited group of drugs with the ability to treat serious human infections are considered CIAs. HIAs encompasses two characteristics, these drugs are either

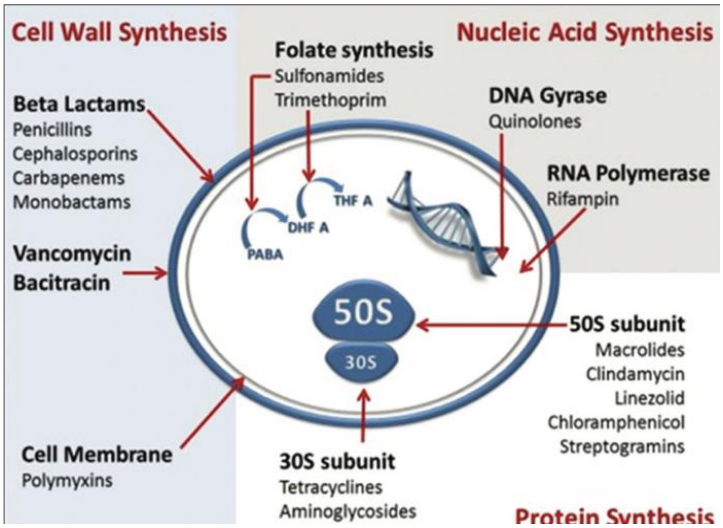

Figure 1 | **The three primary factions of antimicrobial mechanisms of action.** A large subset of antimicrobials can be partitioned into three main classes for mechanism of action: cell wall, nucleic acid, and protein synthesis inhibitors. Adopted from Kapoor et al., 2017.

those that can treat serious human infections but with multiple drug classes available for therapeutic use, or they are one of the limited therapies that can be utilized to treat non-serious infections in humans. IAs are used to treat non-serious human infections when multiple antimicrobial classes of therapeutic drugs are effective and available. Antimicrobials deemed not medically important failed to meet any of the abovementioned ranking criteria and are considered to have less potential impact on human health. Both medically and not medically important antimicrobials are heavily utilized in animal agriculture, which can increase the risk of direct and indirect negative impacts on human health as a result of AMR selection and emergence, reinforcing the notion that AMR is a One Health concern that requires full scale collaboration to aid mitigation.

1.5 Mechanisms of antimicrobial resistance

Effective mitigation of AMR is heavily reliant on the large-scale comprehension of the issue and its means of operation. The molecular mechanisms associated with antimicrobial resistance (Figure 2) are broad, and still require additional research to develop a complete understanding. Discussed here is a limited variety of the understood and common AMR mechanisms including those that a) modify the antibiotic target, b) modify or degrade the antibiotic itself, and c) reduce the intracellular accumulation and or concentration of the antibiotic. Modification of an antimicrobials target can reduce binding affinity for the antimicrobial, thus rendering the drug ineffective and conferring resistance. Alteration of the target protein is typically mediated though enzymatic function and or genomic mutations. Beta lactam and macrolide resistance are commonly developed in this manner through mutations of genes associated with penicillin-binding proteins, and enzymatic methylation of 16S rRNA by methyltransferases, respectively. Inactivation of the antimicrobials themselves can be accomplished though enzymatic modification and degradation of its molecular structure. Degradation of the antimicrobials is carried out by hydrolysis of the molecules functional group; modification typically involves the enzymatic transfer and addition of

chemical groups thus restricting antimicrobial binding to its target. This mechanism of resistance is also common in beta lactams, as beta-lactamases are capable of hydrolyzing the amide bond in the beta lactam ring which results in antimicrobial degradation (Darby et al., 2022; Jeong et al., 2010; Tooke et al., 2019). Modification by the transfer of a chemical group can be exemplified by aminoglycosides, for which a multitude of different enzymes can alter the hydroxyl or amino groups subsequently reducing binding affinity and decreasing efficacy. Lastly, mechanisms pertaining to the reduction of intracellular accumulation and or concentration of antimicrobials has proven an effective method of AMR. These mechanisms are primarily associated with membrane permeability and efflux systems. As previously discussed, many antimicrobials act within the confines of the cell such as nucleic acid and protein synthesis inhibitors, consequently, their effectiveness is reliant upon their ability to enter and persist in the cell. Thus, resistance mechanisms that downregulate transmembrane proteins such as porins or alter the composition of the cellular envelope can directly reduce cell permeability for antibiotics. Additionally, if antibiotics do enter the cell, they can be expelled though efflux activity. Efflux pumps are transmembrane proteins capable of diminishing the concentration of toxic materials such as antibiotics by directly exporting the compounds out of the cell. Efflux is a major contributing mechanism to AMR and multidrug resistance and has

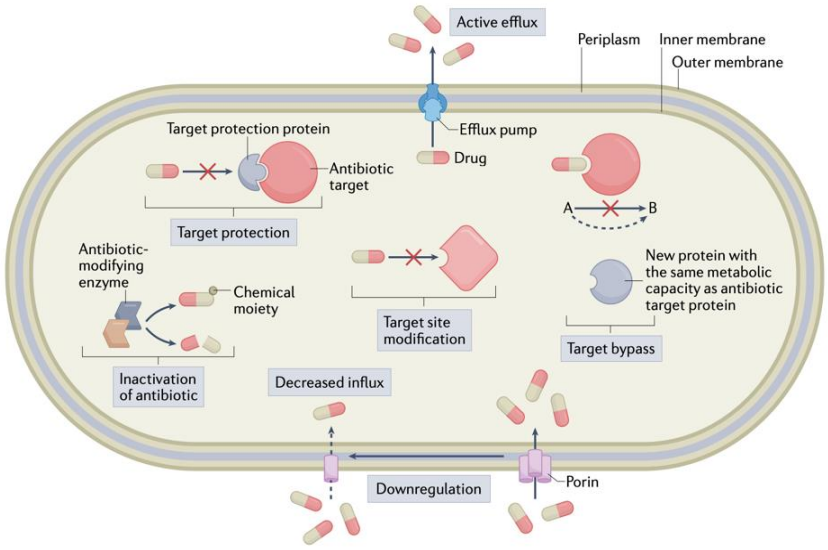

Figure 2 | **Primary mechanisms of antimicrobial resistance.** In order to combat the mechanisms of action associated with antimicrobials, microbes commonly employ these mechanisms of resistance. Mechanisms illustrated here include decreased influx, increased efflux, target protection, bypass, and site modification, as well as antimicrobial inactivation. Adopted from Darby et al., 2022.

been shown to work cooperatively with many of the other resistance mechanisms already discussed (Darby et al., 2022). In depth understanding of mechanisms such as these is crucial to mitigating AMR, and the dynamics of its emergence moving forward.

1.6 Monitoring and surveillance of AMR

The development of global surveillance systems to monitor AMR dynamics in humans, animals, and the environment is critical to fully understand the interface between the three and inform future management strategies. Accomplishing this feat is reliant upon the accurate quantification of antimicrobial resistance in samples. Traditionally, the benchmark for determining antimicrobial resistance has been to selectively culture microbes and measure the exhibited phenotypes when subjected to the presence of various antimicrobials. This methodology is employed globally for multiple surveillance systems; however, it is based on a limited number of culturable pathogens and typically only evaluates antimicrobials relevant to human health. As a result, a plethora of unculturable pathogens and many antimicrobials utilized in alternative sectors such as animal agriculture are not evaluated. In many countries, antimicrobial consumption in animal agriculture surpasses that of human usage, and it is estimated that nearly 73% of global antibiotics can be contributed to livestock (Van Boeckel et al., 2017). It should be noted that is it difficult to draw comparisons between antimicrobial use in humans and animals due to the large disparity in human and animal numbers as well as the dose and duration of antimicrobial administration between species (FDA, 2019). Nonetheless, animal agriculture contributes immensely to total global antimicrobial usage and moreover the selective pressure on AMR development. Unfortunately, surveillance for AMR in farm animals typically relies on passive reporting to gather representative data (Woolhouse et al., 2015). Active surveillance on the other hand provides more accurate and up to date information.

Regarding alternative methodologies for evaluating AMR presence, recent advancements in

high throughput DNA sequencing technologies have made way for cost effective shotgun metagenomic sequencing, allowing for the interrogation of an entire collection of DNA present in a complex system such as animal and environmental samples, including novel and unculturable microorganisms (Quince et al., 2017). The process of shotgun metagenomics from start to finish entails 1) sample collection, 2) DNA extraction, 3) random fragmentation of the genomic material, 4) high throughput sequencing, and 5) subsequent bioinformatic analysis (Quince et al., 2017). This design allows for subsequent data analyses to mine critical information from vast amounts of data including the identification and estimation of antibiotic resistance genes (ARGs). Sequencing is typically performed on the Illumina platform, but advancements in long read sequencing technology from Oxford Nanopore and PacBio has made their utilization viable in the field (Quince et al., 2017). Shotgun metagenomics is commonly used to identify taxonomic composition and function of microbial communities, however, its prevalence in AMR research has grown in recent years. Contemporarily, this methodology has been employed to evaluate ARG presence in samples of human fecal matter origin (urban sewage and toilet waste from airplanes), and animal agriculture (feces from swine, milk filters from dairy farms, and feces from broiler farms). In 2019, a study by Hendriksen et al. sampled untreated urban sewage from 79 cities in 60 countries and subjected them to shotgun metagenomic sequencing to interrogate regional variation of ARG abundance and dissect factors having potential influence. Their findings concluded that there are regional disparities in ARG abundance which are strongly correlated with socioeconomic, health, and environmental factors. Regional differences could be conglomerated into two working groups of similarity: Europe, North America, and Oceania and Africa, Asia, and South America. Of note, the most highly abundant ARGs were associated with macrolides, tetracyclines, aminoglycosides, beta-lactams, and sulfonamides (Hendriksen et al., 2019). This study highlights the ability of metagenomics to characterize AMR development in healthy populations, and more importantly, that there is global

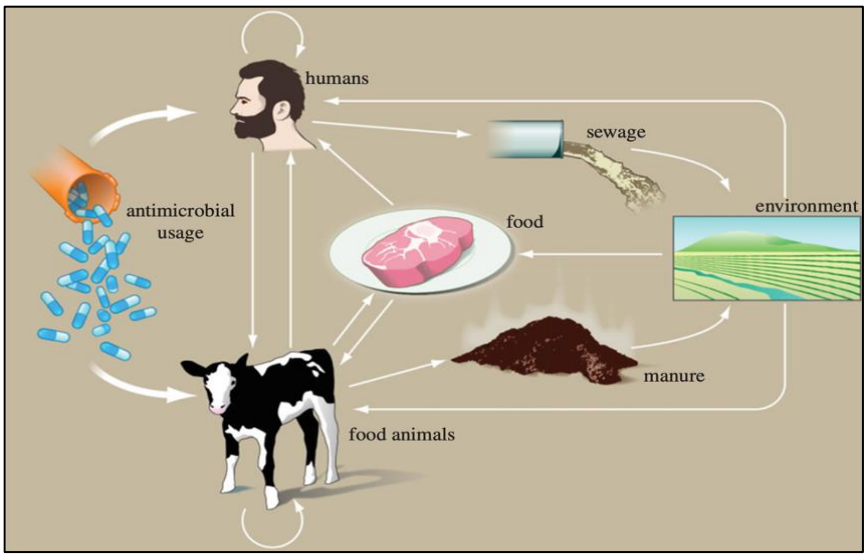

Figure 3 | **The routes of transmission of AMR between livestock, humans, and the environment.** Antimicrobial resistance can be acquired through multiple direct and indirect pathways for each sector. Adopted from Woolhouse and Ward,

disparity and nonuniform dynamics of AMR. Transitioning to animal agriculture associated studies, researchers in Denmark shotgun metagenomically sequenced fecal samples from 181 swine and 178 poultry farms from nine European countries (Munk et al. 2018). Interestingly, upon quantification of the resistome for each species there were significant differences in abundance and composition. Additionally, the study concluded that AMR prevalence was associated with the degree of antibiotic utilization in livestock. Moreover, a country effect on resistome composition was determined significant with variable degrees between species. Showcasing the ability to differentiate AMR profiles based on associated factors in animal agriculture such as species, antibiotic usage, and geographical location. An additional study pertaining to dairy cattle in animal agriculture sought out to evaluate the distribution of ARGs in complex samples such as bulk tank milk filters (Rubiola et al., 2022). The purpose of utilizing an uncommon sampling source such as milk filters is to a) investigate the ability to determine resistomes with raw milk and b) interrogate food security associated risks relevant to AMR. The study concluded that the shotgun metagenomic sequencing of raw milk found in bulk tank filters can successfully determine the associated resistome, and that raw milk can be considered a source of AMR bacteria and genes. Outlining the importance of proper hygiene on dairies, and food safety risks associated with the consumption of raw milk. Once again emphasizing the importance of the

human, livestock, and environment interface. The routes of AMR transmission in this interface are illustrated at a fundamental level in Figure 3 (Woolhouse et al., 2015). Antimicrobial usage in humans and livestock encourages the development of resistant microbials, which can then follow various routes of transmission. Transmission can occur between different species within livestock and humans via close proximity or the consumption of food products from animal agriculture. Additionally, excrement from the living organisms can transmit AMR microbes and genomic material into the environment. Environmental AMR can also be routed back to humans, livestock, and food production, where in the case of the living organisms it can be recycled and spread within the species. It is clear that animal agriculture plays an integral role in AMR development and dissemination, and thus it is vital that rapid and quality livestock surveillance systems be established to aid in the mitigation of this global One Health threat.

1.7 One Health

Despite the encouragement of amicrobial resistance due to selective pressures of antimicrobial use, it should be noted that antimicrobial resistance is not anthropogenic and existed prior to the discovery, development, and utilization of contemporary antimicrobial agents (Zhang et al., 2022). In fact, a recent study in reconstructed ancient microbial genomes from the human gut utilizing over a thousand-year-old preserved palaeofaeces samples identified the presence of antimicrobial resistance genes (D'Costa et al., 2011). Additionally, ARGs exist in the environment and have been discovered in permafrost cores and pristine soil samples from the Yukon territory of Canada and Antarctica, respectively (Zhang et al., 2022; Van Goethem et al., 2018; Wibowo et al., 2021). Each of these studies

Figure 4 | **The relationship between Animal, Human, and Environmental Health.** At the cross section of these three sectors resides the concept of One Health, illustrating the importance for the optimization of each sectors health both individually and as a conglomerate. Adopted from the International Livestock Research Institute.

reinforces that antimicrobial resistance is a natural occurring phenomenon that was not established in response to antimicrobial use, but rather amplified though selective pressures and microbial proliferation. The existence of antibiotic resistant genes in the environment poses a threat to human health as the genes can be acquired by bacterial hosts in humans as well as potential pathogens. The development and emergence of AMR is multifaceted, and the large breadth of influence associated with this global health threat classifies it as a One Health issue. One Health is the idea that human, animal, and environmental health are all intertwined (Figure 4), and the optimization of this kingpin requires a collaborative, multisectoral, and transdisciplinary approach. Regarding AMR, practices employed within each sector heavily influence the health of all three.

1.8 Metagenomic data in public databases

Vast amounts of data produced from published and unpublished studies utilizing shotgun metagenomic sequencing have accumulated in public repositories, most notably in the Sequence Read Archive (SRA) database hosted by the National Center for Biotechnology Information (NCBI). The tens of thousands of metagenomic samples of animal agriculture origin residing in the SRA are incredibly diverse in terms of species, sampling source (fecal, intestinal, tissue), time, and location. Additionally, the metagenomic data is rather unbiased as a result of unintended sampling since the majority of studies were not primarily investigating AMR. That is to say, most studies were not actively pursuing samples where a certain degree of AMR prevalence would be likely. As a result of the aforementioned characteristics associated with these metagenomic samples in the SRA, there exists the ability to capture large scale antimicrobial resistance information in global animal agriculture though the quantification of ARGs. The estimates of antimicrobial resistance gene abundance, and their correlation with a multitude of factors such as national economic indices and antimicrobial usage can then be investigated to offer insights into the dynamics and emergence of AMR as well as mitigation strategies.

Chapter 2. Development of an ARG estimation pipeline from metagenomic data

2.1 Introduction

The development and optimization of a bioinformatic pipeline capable of processing data from raw metagenomic sequences to gene abundance is instrumental in the accurate and efficient quantification of ARGs. Primarily, bioinformatic pipelines such as these employ one of two methodologies: assembly based, or alignment (mapping) based. The process for each method starts identically, microbial DNA is randomly fragmented into many smaller pieces of genomic material and then massively parallel sequenced i.e., shotgun metagenomic sequencing. This results in the production of fastq files which contain a large quantity of reads (nucleic acid sequences). As illustrated in Figure 5, it is here where the divergence of assembly-based and alignment-based methods occur. The alignment-based method entails mapping the prior produced sequence reads to a

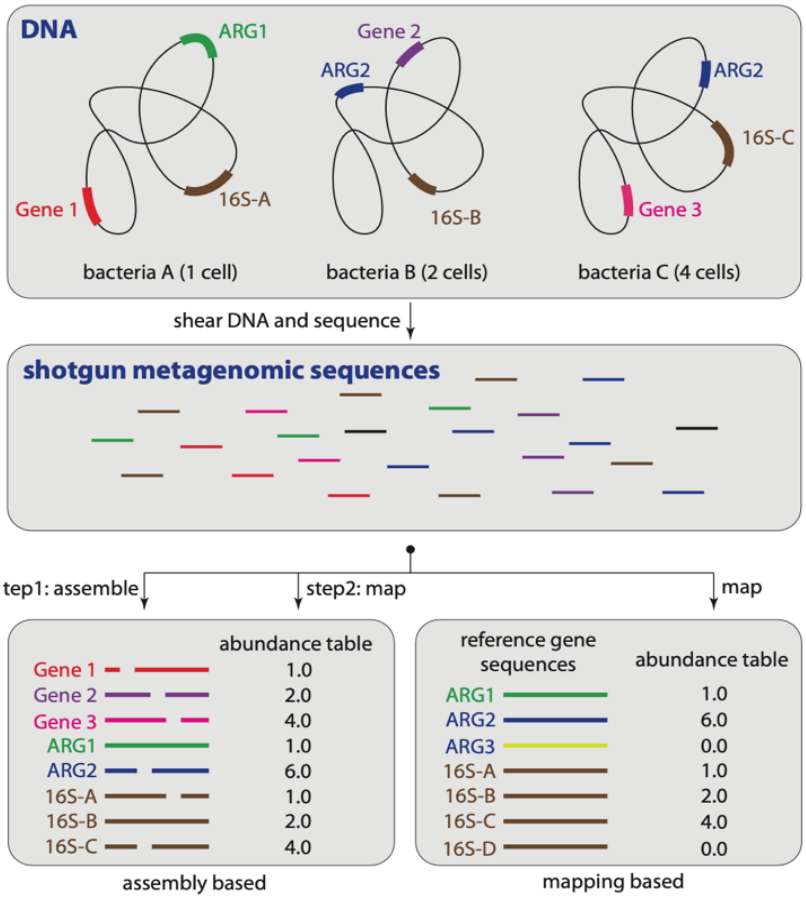

Figure 5 | **Shotgun metagenomic sequencing and ARG abundance estimation.** The procedure associated with shotgun metagenomic sequencing and the estimation of absolute ARG abundance are illustrated. Three distinct hypothetical bacteria are sampled and subjected to shotgun metagenomic sequencing. The produced sequence reads are subsequently utilized to estimate ARG abundance by either a two-step assembly based (left) or mapping based (right) approach.

reference gene catalog, in our case one comprised of existing ARGs. Abundance estimation is subsequently quantified by the number of times each gene in the reference catalog was mapped to from the input query read sequences. Alternatively, the assembly-based methods require a two-step approach. Reads are first computationally assembled into complete genomes via overlaps of similarity between reads. This is made possible due to the high throughput and coverage produced from shotgun metagenomics sequencing. The resulting assemblies can then be functionally classified to produce a reference gene catalog. Finally, abundance can be estimated by mapping the shotgun reads to the assembled gene catalog. In layman's terms, alignment-based methods entail mapping to a known reference gene catalog, whereas assembly-based methods produce their own reference gene catalog and entail mapping to the assembled gene catalog. The utilization of each method comes with its associated benefits and disadvantages when compared with the alternative. Compared to assembly-based methods, alignment-based methods are computationally efficient but may miss unannotated genes and normalization across samples is more challenging. On the other hand, the assembly-based method is more computationally expensive but is better at capturing the total microbiome and may discover novel genes undefined in reference databases. Each method has been shown to effectively capture ARG abundance, nonetheless, due to the sheer volume of data we wish to process, an alignment-based method was employed. The objective of this project was to establish alignment-based analytical framework for estimating ARG abundance in metagenomic samples, that is additionally capable of processing large amounts of bulk genomic sequence data in a feasible time frame.

2.2 Materials and Methods

2.2.1 Metagenomic data

The development and optimization of the constructed bioinformatic pipeline utilized in house published data from the metagenomic sequencing of ileum, cecum, and colon samples (n=18) from 6 distinct pigs (Quan et al., 2020). The objective of the aforementioned study was to compare the metagenomes from different intestinal regions of pigs with contrasting feed efficiency, and an assembly-based bioinformatic approach was utilized for quantification of ARG abundance. The produced data from this study serves here as an evaluation dataset for the development and optimization of a new bioinformatic pipeline as well as a means for comparison between assembly and alignment based methodologies.

2.2.2 Reference databases

Databases utilized for alignment include the Comprehensive Antibiotic Resistance Database (CARD) for antimicrobial resistant genes, and the GreenGenes database for 16S rRNA genes which are later used for taxonomic purposes. The genes comprising each respective database were clustered based on sequence identity to remove any redundancy using CD-HIT. More precisely, CD-HIT-EST (W. Li & Godzik, 2006) was utilized with the following parameters: local sequence identity (as opposed to global), alignment coverage of 90%, clustering to the most similar representative (as opposed to the first cluster to meet clustering criteria), and either a 95% or 99% sequence identify threshold. The CD-HIT-EST algorithm works as follows: 1) input sequences are sorted in order of decreasing length, 2) the longest sorted sequence becomes the representative of the first cluster, 3) the following sequences of shorter length are compared to the representatives of the existing clusters. The sequence of interest will be partitioned into the first representative sequence for which it meets the clustering criteria defined above (sequence identity, alignment coverage, etc.). Therefore,

if a sequence of interest meets the clustering criteria for multiple clusters, it will be partitioned into the cluster with the longest representative sequence. If a sequence fails to meet all of the clustering criteria for all of the existing clusters, a new cluster is formed with the sequence of interest serving as the cluster representative. Once clustering was finished, the representative sequences were retained and used moving forward. Each database was clustered twice, once with 95% and once with 99% sequence identity, to inform optimization of the pipeline discussed later in this chapter. Upon clustering at 95% sequence identity, the CARD and GreenGenes databases were filtered from 4,605 genes to 1,323 genes, and 1,075,170 genes to 52,161 genes, respectively. At a sequence identity threshold of 99%, the CARD and GreenGenes databases were filtered from 4,605 genes to 1,986 genes, and 1,075,170 genes to 175,724 genes, respectively. Gene length distributions for the four individual nonredundant databases are illustrated in Figure 6. Each clustered database was then indexed using BWA-INDEX as required for subsequent mapping to the gene catalogs. This process is performed once and is not repeated for each sample processed by the constructed bioinformatic pipeline.

All sequence processing and abundance estimation was performed on the Michigan State University (MSU) High Performance Computing Cluster (HPCC) hosted by the MSU Institute for Cyber-Enabled Research (ICER).

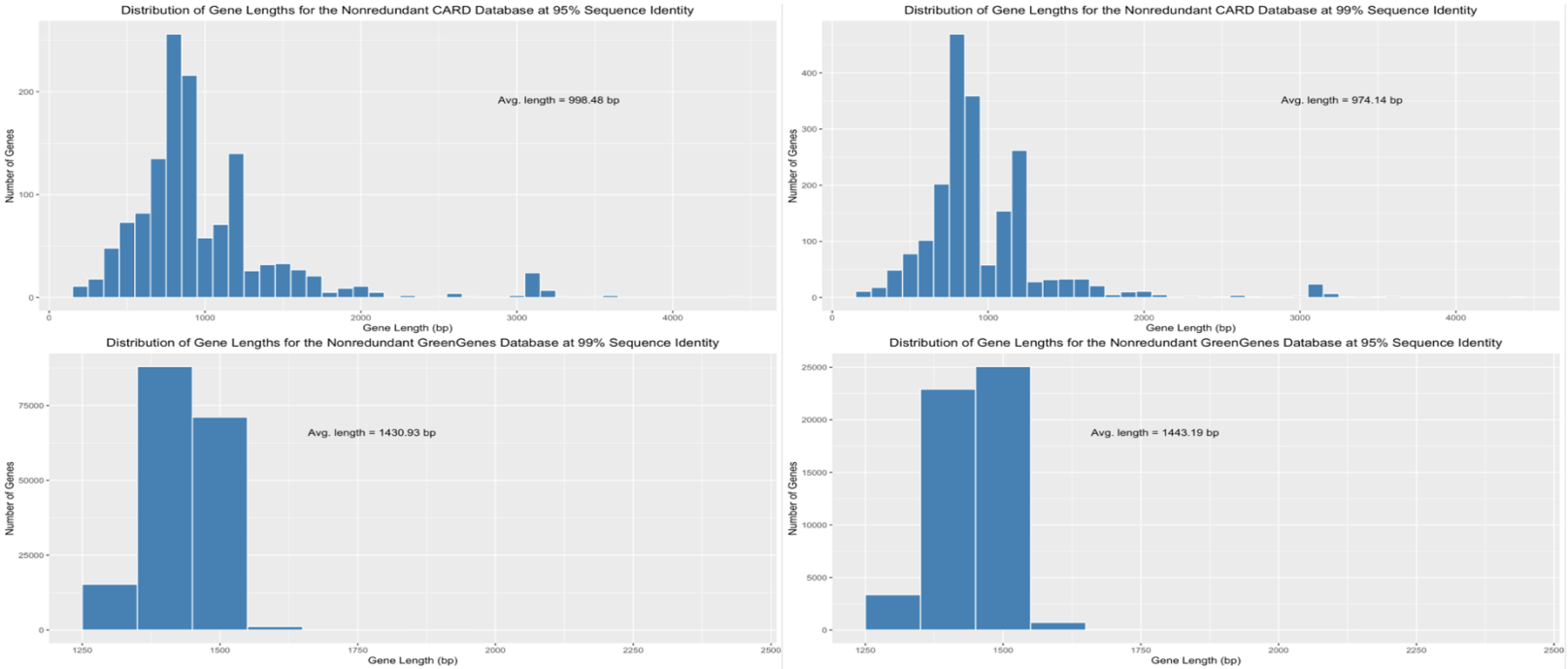

Figure 6 | **Disparity in Gene Length Between Each Nonredundant Reference Database at Variable Sequence Identity Clustering Thresholds**. Each Database (CARD and GreenGenes) were clustered at 95% and 99% sequence identity. The distribution of each nonredundant databases gene lengths, and the average gene length are illustrated. Of note, the 99% sequence identity clustering equated to a relatively smaller average gene length as due to the parameters, less genes met the clustering criteria for each large cluster representative sequence and instead formed their own cluster. Consequently, more clusters with a lower average gene length were produced in the 99% sequence identity reference databases compared to that of the 95% sequence identity reference databases.

2.2.3 Sequence processing

2.2.3.1Fastq quality control

First and foremost, the raw sequence data housed in fastq files are subjected to quality control. This primarily pertains to adapter removal and quality trimming of the shotgun reads, which is performed using BBduk (Bushnell, 2014). A variety of alternative tools exist, such as Cutadapt and Trimmomatic, but BBduk has been shown to be extremely fast, scalable, and memory-efficient while still retaining relatively comparable accuracy to the competition (Guzman & D'Orso, 2017). Additionally, BBduk handles both single-end and paired-end reads very smoothly. Regarding parameterization, a quality threshold of 20 on the Phred scale (99% base calling accuracy) and a minimum length of 50 bases was employed. Filtering of reads based on length is performed after adapter removal and quality trimming. Quality control metrics such as these are a balance between sensitivity and specificity. If the parameters are too relaxed, then filtering is considered sensitive as the retention rate of truly positives is higher. Alternatively, if the parameters are rather strict (specific) then a higher proportion of the true negatives are accurately discarded. When selecting quality control thresholds, it is desirable to optimize a middle ground between the two ends of the spectrum so that reads with high confidence of accuracy off the sequencer are retained and potentially accurate data with lesser confidence aren't massively removed. A quality trimming threshold of 20 resides in the middle of this spectrum, and a minimum length of 50 bases errs on the side of specificity; both of which generally agree with parameters commonly used in the literature. While not included presently, future development of this quality control step can allow for the filtering of host reads. This serves two primary purposes, to validate the sample species origin and remove contaminants in the data.

Validation of the quality control step is performed using FastQC on both the raw sequence data and the retained clean data. FastQC performs various basic quality control (QC) metrics

including sequence quality, read length distribution, duplication levels, GC content, and adapter content. By subjecting both the raw and clean sequence data to FastQC, any QC metrics flagged for concern in the raw data can be cross-checked with the associated clean data QC metrics to validate that the issues have been resolved. The utilization of this technique allows for the wrangling of additional erroneous data that escaped quality control filtering prior to further downstream analysis.

2.2.3.2Alignment

A two-pronged alignment is then performed to query the clean sequence data against the nonredundant 16S rRNA and ARG gene sets. The BWA-MEM algorithm was employed to map the input reads against the references (Figure 7), despite the availability of its contemporary successor BWA-MEM2. BWA-MEM2 produces an identical alignment to that of BWA-MEM but is significantly faster (up to 3.1x). However, the cost of this rapid alignment is skyrocketed memory costs, and due to the high throughput nature of shotgun metagenomic sequencing the utilization of BWA-MEM2 became infeasible with the available resources. The Burrows-Wheeler Alignment (BWA) (Li & Durbin, 2009) is a read alignment tool contingent upon the Burrows–Wheeler Transform (BWT) (Burrows and Wheeler, 1994). In the case of BWA-MEM, alignments are seeded with maximal exact matches (MEMs) and then seeds are extended with the affine-gap Smith-Waterman algorithm (SW) (H. Li, 2013). The algorithm handles sequencing errors very well and works with a wide range of sequence lengths. The execution of this algorithm given reference and query sequences will result in the production of SAM (Sequence Alignment Map) files containing the query sequences and the

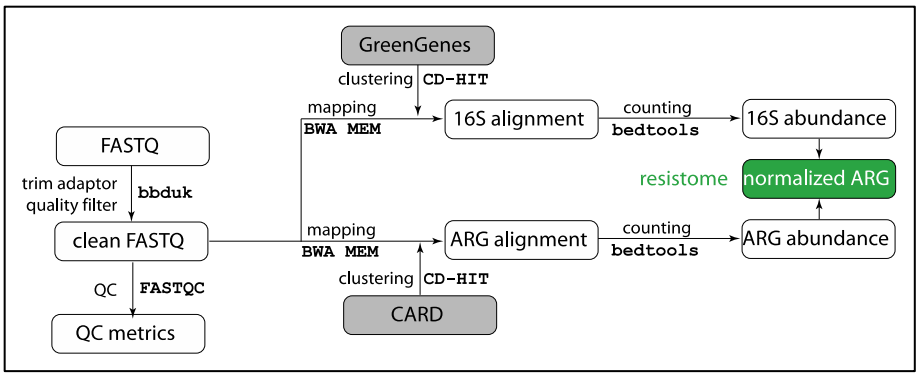

Figure 7 | **Overview of the ARG abundance estimation pipeline.**

reference sequence they aligned to (if applicable). Upon the completion of BWA-MEM alignment with default parameters, each output SAM file was converted to BAM (binary) format with SAMtools software to conserve storage space on the MSU HPCC. The outputs of the alignment step are two BAM files for each input clean fastq file, individually containing mapping information for the input reads to either the 16S rRNA or ARG references. These files contain both mapped and unmapped reads from the clean data.

2.2.4 Gene quantification

2.2.4.1 Absolute abundance estimation

In order to estimate the absolute abundances for the ARG and 16S rRNA genes of interest, the unmapped reads from the previously produced alignment files must be filtered out and discarded. This is accomplished based on each reads associated MAPQ scores produced by BWA-MEM during alignment. MAPQ scores for BWA-MEM range from zero to sixty [0,60] on the Phred scale, a score of zero means that the read could not meet the mapping criteria for any of the genes in the reference catalog (Figure 8). MAPQ scores from one to sixty [1,60] express a level of confidence in the mapping of an input read to a particular gene in the catalog and is directly related to the number of genes for which a read meets the required mapping criteria. A MAPQ score of 60 equates to extreme confidence by the alignment algorithm that the read maps to a single particular gene in the reference. As the number of genes a read can map to increases, the MAPQ score decreases as the aligner is less confident in which gene is the accurate mapping for the read. Any reads that have a MAPQ score of zero are therefore unmapped to any genes in the reference catalog and are discarded from further quantification steps in the analysis. At this point, quantification is temporarily partitioned into two different bioinformatic processes for single and paired end reads, respectively.

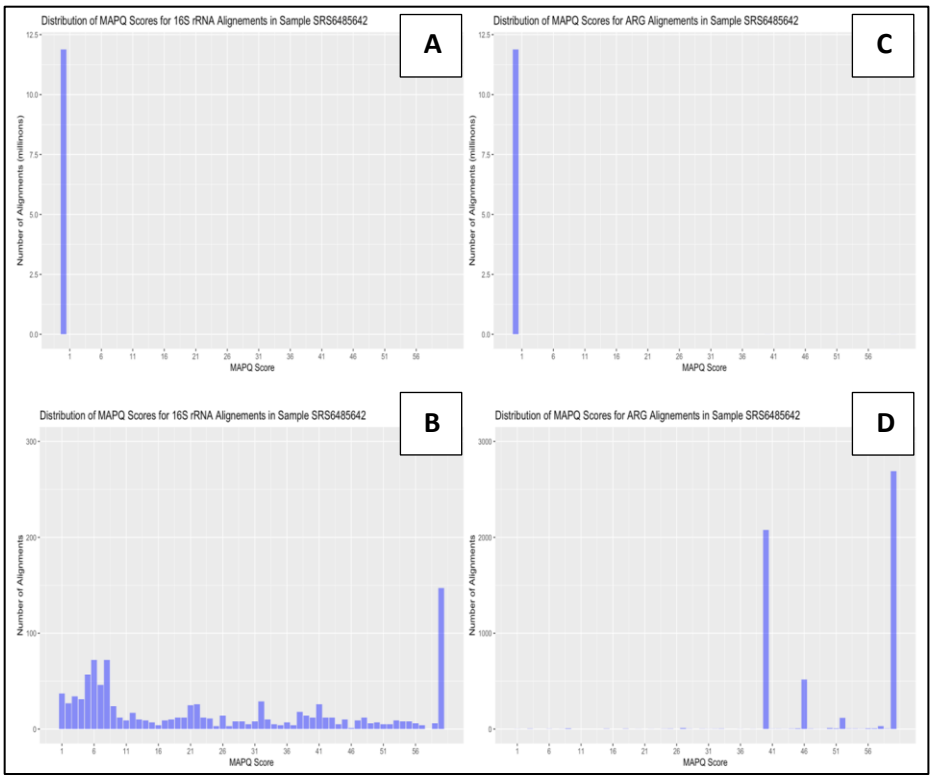

Figure 8 | **MAPQ Score Distribution Produced by Alignment to Each Reference.** The distribution of MAPQ scores associated with the alignment of one individual sample to the 99% sequence identity clustered 16S rRNA and ARG gene catalogs is illustrated. The sample utilized (SRS6485642), was randomly selected from the aforementioned pig intestinal feed efficiency study. A and B both represent 16S rRNA sequence alignment MAPQ scores, however, B illustrates solely MAPQ scores ranging from [1,60]. Similarly, C and D both represent ARG sequence alignment MAPQ scores, with D illustrating solely MAPQ scores ranging from [1,60]. A and C serve to highlight the prevalence of unmapped reads (MAPQ=0), as the distribution is so disproportionate that none of the alternative MAPQ scores [1,60] can be depicted on the same scale.

Single end analysis is relatively straight forward. The number of mappings for each read in the alignment file are summed, and a three-column table is produced with columns 1, 2, and 3, equating to read name, mapped gene name, and number of mappings for the read in question, respectively. To combat multimapping for genes in the reference catalog, the value in column 3 is then inverted to weight each mapping (i.e., 3 gene mappings by one read is inverted to 1/3 of a count for each gene). This ensures that multiple mappings for a single read does not equate to multiple gene counts, as this would overestimate abundance.

Paired end analysis requires additional considerations since the two paired end reads may not necessarily confer identical gene alignments. Consequently, each paired end alignment file must be spliced into two alignment files containing the left end and right end reads, respectively. Each of

these files are then used to produce a table comprising of three columns where columns 1, 2, and 3, represent read name, mapped gene name, and MAPQ score, respectively. Subsequently, the resulting tables are merged into a five-column table for which columns 1, 2, 3, 4, and 5, equate to read name, left end gene mapping name, right end gene mapping name, left end gene mapping MAPQ score, and right end gene mapping MAPQ score, respectively. Here, one of the paired end reads (left end or right end) are selected as the representative for which gene the read was mapped. Selection of this representative was determine using the following procedure: 1) if only one read mapped to a gene in the reference, the mapped read is retained as the representative, 2) if each read was mapped to an identical gene the left end read is selected, 3) if each read mapped to different genes, whichever mapping boasts a higher MAPQ score is selected, 4) if each read mapped to different genes with identical MAPQ score, the left end read is selected. Because of the random nature of left and right assignment, the last (4) selection effectively randomly assigns reads to one of the two genes. Once representatives for the paired end reads have been determined, the remaining reads are processed identically to the single end analysis previously described to produce a three-column table containing read name, mapped gene name, and weighted mapping count, respectively.

At this point, the bioinformatic analysis of single end and paired end reads reconvene. For each file, the weighted mappings for each gene are summed subsequently producing a two-column table comprised of reference gene name and associated summation of weighted mappings, respectively. Of note, there are often multiple sequencing runs and produced fastq files for a single sample. Now that each of these files have been processed from raw data to a table of absolute weighted gene counts, they need to be pooled into a single file representing the original sample. This is performed by summing the gene counts for each file and representing them in a new table with the identical format of the input tables. The resulting table represents the absolute abundance of each gene in the reference catalog for the shotgun metagenomically sequenced sample.

2.2.4.2Normalization

Quantification was performed four times using gene alignments to the nonredundant CARD and GreenGenes databases at 95% and 99% sequence identity for both. Resulting in the construction of four large tables formatted as sample by reference gene with values equating to the absolute abundance of each gene within each sample. However, this abundance estimation can be heavily biased due to disparities in microbial abundance between samples which is the purpose for 16S rRNA quantification. The total abundance of 16S rRNA genes serves as a proxy for the total microbial abundance in the metagenomic samples and is used as a means for normalizing quantified ARG abundance. Relative ARG abundance is calculated using the formula found in Equation 1:

$$\text{ARG abundance} = \frac{N_{ARG}/L_{ARG}}{\sum N_{16S}/L_{16S}}$$

Equation 1 | **ARG Abundance Normalization.**

where N_{ARG} is the number of reads mapped to an ARG, L_{ARG} is the length of the ARG gene, N_{16S} is the number of reads mapped to a 16S rRNA gene and L_{16S} is the length of the 16S rRNA gene. After normalization, the relative abundances for each ARG are enveloped in a sample by ARG table which will serve as the basis for all subsequent bioinformatic analysis.

2.3 Results

2.3.1 Correlation between alignment and assembly based methods

As previously mentioned, the sequence data utilized in this chapter originates from publicly available and published data from Quan et al., 2020. As shown in the publication, ARG abundances were estimated utilizing an assembly-based bioinformatic approach (Quan et al., 2020). Thus providing available benchmarks for comparison between the authors assembly-based approach, and the alignment-based method detailed in this chapter. To evaluate concordance between the developed framework utilizing the alignment-based approach, and the existing assembly-based method, correlations between the concluded ARG abundances for each method using the same dataset were calculated. Furthermore, as to determine which sequence identity threshold should be used for reference database clustering, two correlations were calculated. The first of which is the correlation between the estimated ARG abundances from the

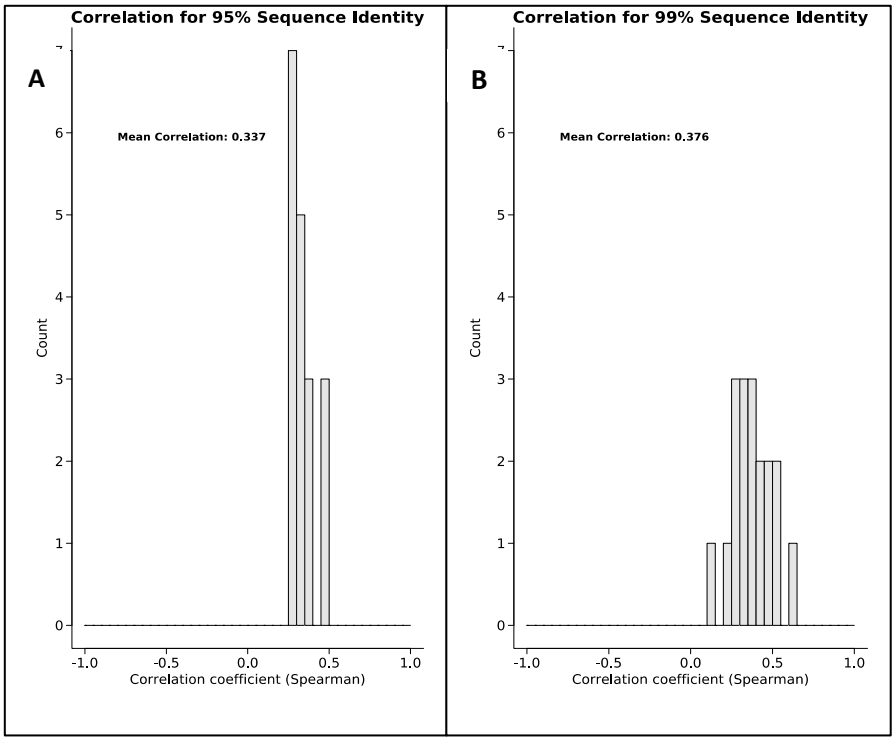

Figure 9 | **Correlation of ARG abundance estimation between the alignment-based and assembly-based methodologies.** Figure A represents the correlation between the alignment-based approach using the 95% sequence identity database and the assembly-based method. Figure B represents the correlation between the alignment-based approach using the 99% sequence identity database and the assembly-based method. The mean correlation of the estimate relative ARG abundance for all samples in A and B are 0.337 and 0.376, respectively.

95% sequence identity reference database alignment-based approach, and the assembly-based ARG abundances from the publication. Secondly, correlations were calculated between the estimated ARG abundances from the 99% sequence identity reference database alignment-based approach, and the same aforementioned assembly-based ARG abundances from the publication. It should be noted that the assembly-based approach is by no means the gold standard for determining ARG abundance, but there is merit to concluding comparable results when the alignment-based approach is substantially more computationally efficient. As shown in Figure 9, the average correlation for the 99% sequence identity reference database (0.376) was marginally higher than its 95% counterpart (0.337). Moreover, the sample with the highest correlation for the 99% reference database was 0.620, significantly higher than 0.497, which is the highest sample correlation from the alternative 95% database. As a result, further analyses with the alignment-based approach will utilize the nonredundant CARD and GreenGenes databases clustered at 99% sequence identity. Furthermore, the results from the alignment-based approach generally agree with the concluded ARG abundances produced via the assembly-based approach.

2.3.2 Computational processing time

When it comes to computational efficiency, CPU time spent on gene quantification and normalization is negligible and nearly instantaneous. However, sequence data trimming and alignment can prove taxing when it comes to large scale data processing. As the pig intestinal data processed above comprises of merely 18 samples, computational burden analysis was performed using the data set discussed later in chapter 3 which contains 126 samples and provides more accurate estimations. As shown in Figure 10, the average time spent to process 1 million reads for trimming, 16S alignment, and ARG alignment, was 0.08, 55.09, and 54.60 minutes, respectively. Therefore, the average CPU time required to process 1 million reads is 109.77 minutes. The relationship between number of reads and time spent for each step generally illustrates a positive

and linear relationship. However, as exhibited in Figure 10, components b and c, there is clustering of files containing over ~600 million reads for the alignment steps that do not agree with the concluded linear relationship. This artifact is a result of variable node efficiency on the MSU HPCC, as some nodes are newer and faster than others and the processing of the samples included in these clusters took place on these particular nodes. Thus, partially explaining the contradictory nature of these clusters to the generally exhibited correlation between number of reads and CPU processing time. This process can be accelerated to a certain extent through optimized multithreading and parallelization in the future.

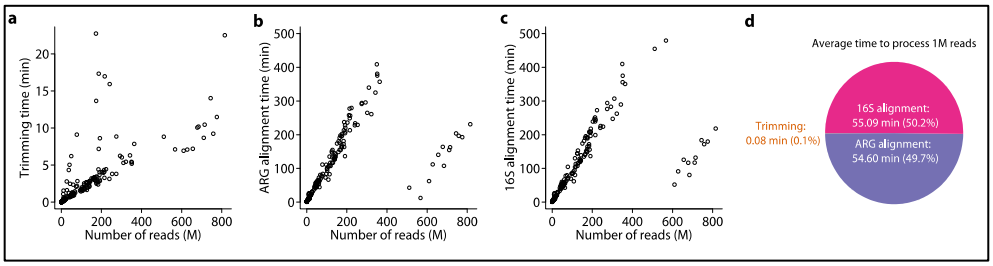

Figure 10 | **Computational Burdens Associated with Read Processing.** CPU time allocation for read processing is not equal proportioned. The average time to process 1 million reads is 109.77 minutes, 99.9% of which can be attributed to alignment. Trimming (a) time is negligible. As shown in a, b, and c, number of reads is directly related with the time spent processing.

2.4 Conclusion

In conclusion, the outlined analytical framework for estimating antimicrobial resistance gene abundance in metagenomic samples of animal agriculture origin is capable of processing vast amounts of genomic sequence data in an incredibly short amount of time. Moreover, the processing time can be accelerated to an even shorter time duration with the incorporation of optimized multithreading and parallelization on HPCC's moving forward. Furthermore, the concluded ARG abundances generally correlate with those of which produced by the alternative assembly-based methodology. This makes large scale application to bulk datasets of genomic sequence data feasible, opening the door for a variety of future explorative studies regarding the identification of ARG presence in animal agriculture.

Chapter 3. Application of developed framework on metagenomic samples

3.1 Introduction

Vast amounts of shotgun metagenomic sequence data pertaining to six particular animal species in animal agriculture (pig, cattle, chicken, sheep, goat, and horse) exist in the Sequence Read Archive (SRA) hosted by NCBI. The SRA is heavily intertwined with the BioSample and BioProject databases, also hosted by NCBI. The BioProject database houses information pertaining to each study at large and is therefore applicable to multiple samples. The BioSample database contains metadata for each unique sample collection. Curation of characterized sequence data from the SRA requires the utilization of information stored in each of the three databases. Moreover, the public repository has exhibited continual growth in terms of volume for sequence data and number of metagenomic samples and is still actively growing. Notably, the number of metagenomic samples submitted the last two years (2021, 2022) has surpassed the previous maximum, and 2022 in particular experienced the largest growth of sample submissions to date. As the cost of sequencing in modern day continues to drop, it can be anticipated that growth will continue moving forward. As of November 2022, the number of shotgun metagenomic samples in the SRA

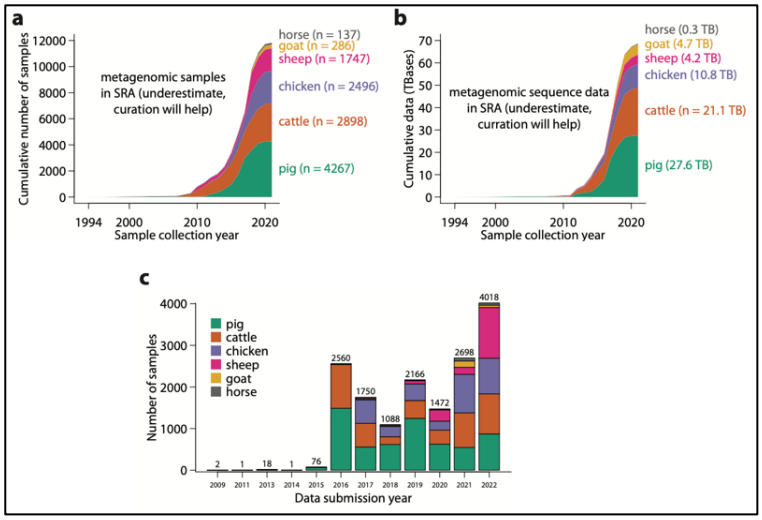

Figure 11 | **Rapid growth of shotgun metagenomic data of animal agriculture origin in the SRA.** (a) Cumulative number of metagenomic samples for six species in the SRA over the past three decades. (b) Cumulative size of metagenomic sequence data in the SRA over time, split by six animal species. The size is expressed in T (10^{12}) Bases. (c) Number of samples submitted to the SRA in each year for each species, yearly totals stated above the stacked bars.

was estimated to be 15,850 by filtering the metadata with associated keywords. It's important to note that this number comprises only shotgun metagenomic samples and does not include 16S amplicon sequencing which is a commonly used methodology in the field of metagenomics. 16S amplicon sequencing by nature fails to offer insights into the genome at large, and therefore is of no use for ARG abundance estimation. Illustrated in Figure 11 are trends and characterization for metagenomic samples and sequence data submitted over the years. Currently, the two species with the largest amount of metagenomic samples are swine and cattle with 6,009 and 4,379 samples respectively. Of particular importance, samples and their associated sequence data in the SRA are accompanied by their respective metadata. The associated metadata varies tremendously in terms of detail, but ideally each samples metadata comprises of sample species, collection data, and sampling location, at a minimum. Upon inspection, 11,831 of the 15,850 (~75%) samples contained at least sample collection year information, and collection dates ranged from 1994 to modern day. Moreover, of the 15,850 samples, 15,085 (95%) contained at least country information representing 37 countries in all six continents other than Antarctica. However, sample distribution is uneven as three counties (USA, China, New Zealand) comprise 55% of the total number of samples. Figure 12 illustrates the distribution of samples on a global scale as well as the dynamics of sample contribution over the past decade. As shown, sample submission is growing globally but additional sampling and sequencing is required, especially in low-income countries. Nonetheless,

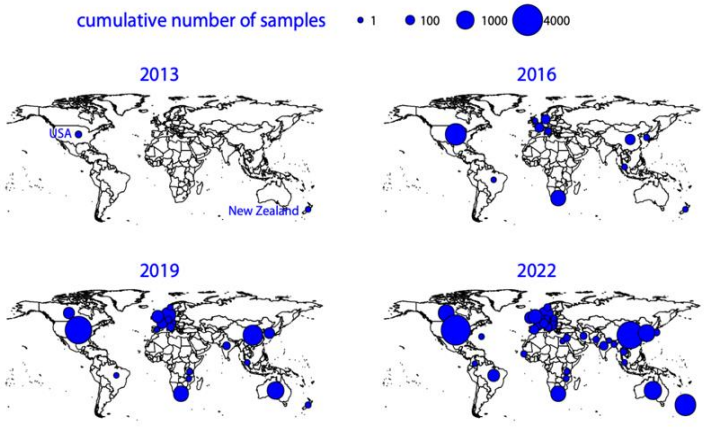

Figure 12 | **Spatiotemporal distribution of metagenomic samples from animal agriculture in the SRA.** Snapshots of the cumulative number of SRA samples from animal agriculture are illustrated for the years 2013, 2016, 2019, and 2022. Significant growth in number of metagenomic samples over the past decade is apparent.

the significant amount of shotgun metagenomic sequence data of animal agriculture origin present in the SRA could prove useful for evaluating spatiotemporal disparity associated with ARGs. The objective of this project was to apply the previously discussed analytical framework for estimating ARG abundance to a subset of metagenomic samples of animal agriculture origin from the SRA, and explore trends associated with the geographical distribution of ARGs on a global scale.

3.2 Materials and methods

3.2.1 Experimental design

An experiment was designed to evaluate regional variation of ARG abundance and serve as a *proof of concept* for further large-scale analysis of the SRA. Shotgun metagenomic samples from a variety of BioProjects, countries, time points, and species housed in the SRA were hand curated to produce a working data set representing a subset of global metagenomic samples from animal agriculture. The curated samples were then subjected to in-depth annotation, and the associated sequence data was downloaded and pushed though the bioinformatic pipeline established in chapter 2 to produce relative ARG abundance estimations. In conjunction with total absolute and relative ARG abundance estimation, ARG composition in terms of associated antimicrobial drug class was also explored. Subsequently, ARG variation between and within various geographical regions and BioProjects were evaluated, and conclusions were drawn.

3.2.2 Sample curation

As previously mentioned, cattle and swine comprise the largest portions of submitted metagenomic samples in the SRA. Therefore, we focused on these two species when curating a working data set as this allowed for the inclusion of more geographical regions within the same species group. The goal was to accumulate a diverse set of samples representing 6 continents (exclusion of Antarctica), and a multitude of countries, for both cattle and swine. When selecting

samples, the availability of a publication was deemed a prerequisite as to 1) ensure that quality data is obtained and utilized, and 2) allow for confident, accurate annotation of the associated samples. Additionally, as an attempt to keep the study relatively balanced, 12 samples were selected for each country included in this study. Of note, availability of quality shotgun metagenomic sequence data (not 16S amplicon sequencing) in Africa was incredibly limited, and available data was refined to one available publication in Malawi which comprised of 31 samples. Of these 31 samples, only 10 were sequenced via shotgun metagenomics and usable for this study. Thus, the selected representative samples from Malawi, Africa (10) did not fit the mold of 12 samples per country despite the availability of both cattle and swine samples. However, for countries with a surplus of available data, curated samples were scattered amongst the geographical landscape within the country. For example, in China, samples were split amongst the cities of Hangzhou and Yunfu, representing the eastern and southern regions, respectively. In the USA, samples were split amongst the states of Pennsylvania, Texas, Colorado, and South Dakota resulting in latitudinal and longitudinal variation for the curated representative samples. Selection of samples within BioProjects were hand selected to represent diverse collection times and geographical locations. For example, if a large study sampled continuously for three years, a relatively equal number of samples were selected from each year. However, within each block of year and location, the utilized samples were randomly selected. Table 1 highlights the number of samples allocated to each country, BioProject, and species for the final working group of curated samples. The selected samples represent 15 unique BioProjects, 11 different countries, and 6 continents; of the 4581 samples attributable to these projects, 130 samples were retained. These 130 curated samples, and their associated publications then progressed to annotation and bioinformation analysis.

3.2.3 Sample annotation

Due to the highly heterogeneous nature of the SRA, selected samples were annotated to confirm available data and or fill in missing data. Additionally, annotation allows for the standardization of the metadata fields. This can be exemplified by collection date heterogeneity as the date October 3rd, 1998, can be expressed as 10/3/1998, 3/10/1998, 10/3/98, etc. Following sample annotation, variation in metadata formatting such as the collection date example can be refined to a singular format which will aid downstream analyses. Prior to sample annotation, the hand curated samples and their associated metadata housed within the BioSample, BioProject, and SRA databases are conglomerated into a local database so that each sample is accompanied by a plethora of information including species, collection date, geographical location, and various accession IDs (SAMN, PRJNA, SRS, SRR).

The curated sample metadata is then subjected to full scale annotation by multiple independent reviewers. The sample annotation procedure is rather detailed, and is solely summarized here, additional information can be found in the appendix. First and foremost, the presence of any missing data in the local database is evaluated and noted. Then, the associated accession IDs for each sample are searched in the three online NCBI databases, and any disparity between the online and local databases is evaluated. Further annotation from this point is reliant upon access to a publication for the sample of interest, which can be identified within the BioProject metadata or through internet exploration. BioProject IDs are typically found within the data availability section of publications, thus providing confirmation once a publication is isolated for a sample of interest. If publications fail to be found during exploration, the submitting authors of the data are contacted via email to acquire missing data or confirmation for the curated metadata. Due to the experimental design defined above, all samples curated for this study are accompanied by detailed publications which were heavily utilized for annotation by the reviewers. Each reviewer read the publication,

identified key information, and annotated the hand curated samples in the local database. Each curated sample was annotated for sample collection date (DD:MM:YYYY format), host species (cattle or swine), animal breed, geographical location (Farm:City:State/Providence:Country format), and source of sample (feces, intestinal, drinking water, etc.). Additionally, variation within BioProjects such as number of species sampled, different sources of sampling, and multiple methods for sequencing (16S amplicon and shotgun metagenomics) are explicitly stated in the annotations. Following annotation of the previously mentioned metadata, reviewers were encouraged to provide additional notes as well as a confidence level in their produced annotations. Confidence levels range from 0% to 100% and are a direct function of specificity in the evaluated publications, if each piece of the necessary metadata was explicitly stated in the publication, then confidence levels were very high. Lastly, concordance between reviewers was evaluated in an all or nothing binary sense. Reviewers either completely agreed with every aspect of the samples' metadata, or they disagreed, regardless of the degree of disagreement.

Following the conclusion of local database annotation, a diverse dataset of shotgun metagenomic sequence data in terms of geographical location and collection time was constructed and validated. The lowest confidence level allocated to a samples' metadata by a reviewer was 70%, and the average confidence level amongst all reviewers for all samples was approximately 91%. Additionally, the annotations for each sample between reviewers was in complete concordance as no disparity existed between annotations. As a result, every sample curated in the local database progressed to the next stage, sequence data download and processing.

SRA Project Accession ID	Country of Origin	Sample Collection Year	Species	Study Sample Size	Number of Selected Samples	Publication	Citation
PRJNA526405	Australia	2017	Swine	877	12	https://doi.org/10.1093/gigascience/giab039	(Gaio et al., 2021)
PRJNA823879	New Zealand	2017	Swine	45	12	https://doi.org/10.3389/fnut.2022.1002369	(Young et al., 2022)
PRJEB26961	Denmark	2014,2015,2016	Swine	218	9	https://doi.org/10.1016/j.prevetmed.2019.104853	(Andersen et al., 2020)
PRJEB31650	Denmark	2016	Swine	277	3	https://doi.org/10.1128/spectrum.00090-22	(Poulsen et al., 2022)
PRJNA390551	USA	2016	Cattle	30	3	https://doi.org/10.1038/s41598-017-12481-6	(Thomas et al., 2017)
PRJNA563872	USA	2013,2014,2015	Cattle	34	3	https://doi.org/10.1089/fpd.2019.2768	(Haley et al., 2020)
PRJNA292471	USA	2013,2014	Cattle	87	6	https://doi.org/10.7554/eLife.13195	(Noyes et al., 2016)
PRJNA684454	Brazil	2020	Cattle	138	12	https://doi.org/10.1128/spectrum.00565-22	(de Carvalho et al., 2022)
PRJNA631951	Iran	2018	Cattle	23	12	https://doi.org/10.1038/s41396-020-00837-2	(Gharechahi et al., 2021)
PRJNA420682	Canada	2014	Cattle	24	12	https://doi.org/10.1038/s41598-018-24280-8	(Zaheer et al., 2018)
PRJEB43305	Switzerland	2019	Cattle	28	12	https://doi.org/10.1038/s41598-021-01031-w	(Y. Li et al., 2021)
PRJEB23561	France	2009,2010,2012	Cattle	82	12	https://doi.org/10.1093/gigascience/giaa057	(J. Li et al., 2020)
PRJNA575543	China	2015	Swine	18	6	https://doi.org/10.3389/fmicb.2020.00032	(Quan et al., 2020)
PRJNA597489	China	2017	Cattle	33	6	https://doi.org/10.1186/s40168-022-01228-9	(Xue et al., 2022)
PRJEB42019	Malawi	2015	Cattle,Swine	31	10	https://doi.org/10.1038/s41564-022-01266-x	(Shaffer et al., 2022)
Total:	**15**	**11**		**1945**	**130**		

Table 1 | **Anatomy of Curated Samples.** The selected SRA samples are associated with 15 BioProjects, 11 distinct countries, and either cattle or swine. Filtering to isolate only samples related to animal agriculture and shotgun metagenomic sequencing resulted in 1,945 samples comprising these 15 projects, for which a subset of 130 was hand selected for further analysis. Collection dates for each sample is shown. The publication DOI and citation for each study is provided.

3.2.4 Sequence data download and processing

From the original 130 metagenomic samples selected, 126 samples were successfully downloaded onto the MSU HPCC as 4 samples were problematic on the SRA end. The associated sequence data for each of the 126 samples were then processed using the developed bioinformatic pipeline characterized in chapter 2. Estimated relative abundances of ARG for each sample were then reassociated with their respective annotated metadata, and conclusions were drawn from the produced abundances in each geographical location. Figure 13 illustrates the employed procedure for sample annotation, downloading the sequence data, and estimating global antimicrobial resistance gene abundance.

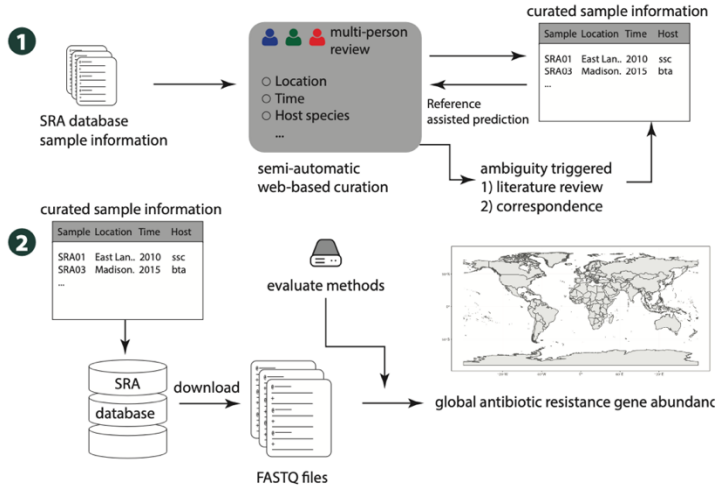

Figure 13 | **Framework for sample annotation, download, and processing.** (1) illustrates the muti-person annotation of curated SRA data, (2) highlights the general schema of the analytical framework to progress from curated sample information and download to global antibiotic resistance gene abundance estimation.

3.3 Results and discussion

Although the number of samples utilized in the study was limited and therefore did not provide enough power to perform formal statistical analyses, some interesting trends in the ARG abundance data could be teased out. Disparities in ARG abundance is evaluated on the BioProject level, however, geographical location such as country of sampling location is indicated to facilitate interpretation.

3.3.1 Analysis of ARGs in metagenomic samples of cattle origin

In cattle, the largest average total relative ARG abundance was found in PRJEB42019 (Malawi), followed by PRJNA420682 (Canada), PRJNA292471 (USA), and PRJNA684454 (Brazil), respectively. Of note, although PRJEB42019 (Malawi) had the highest average ARG abundance, this could be an artifact of the limited number of samples found in the study. Moreover, the individual sample with the highest total relative ARG abundance was found in PRJNA684454 (Brazil). This sample has nearly twice the total ARG abundance than the second highest sample. Additionally, as illustrated in Figure 14a, there is disparity between BioProjects in terms of within BioProject variation. PRJNA684454 (Brazil) has the largest within BioProject variation, which is significantly wider than that of BioProjects such as PRJNA420682 (Canada), PRJNA597489 (China), and PRJEB23561 (France). Regarding the antimicrobial class composition of the estimated ARGs in cattle (Figure 15), there is clearly clustering based on the geographical location of sampling. Samples residing in the same country exhibit highly similar compositions, whereas composition between countries varies rather significantly. Across the

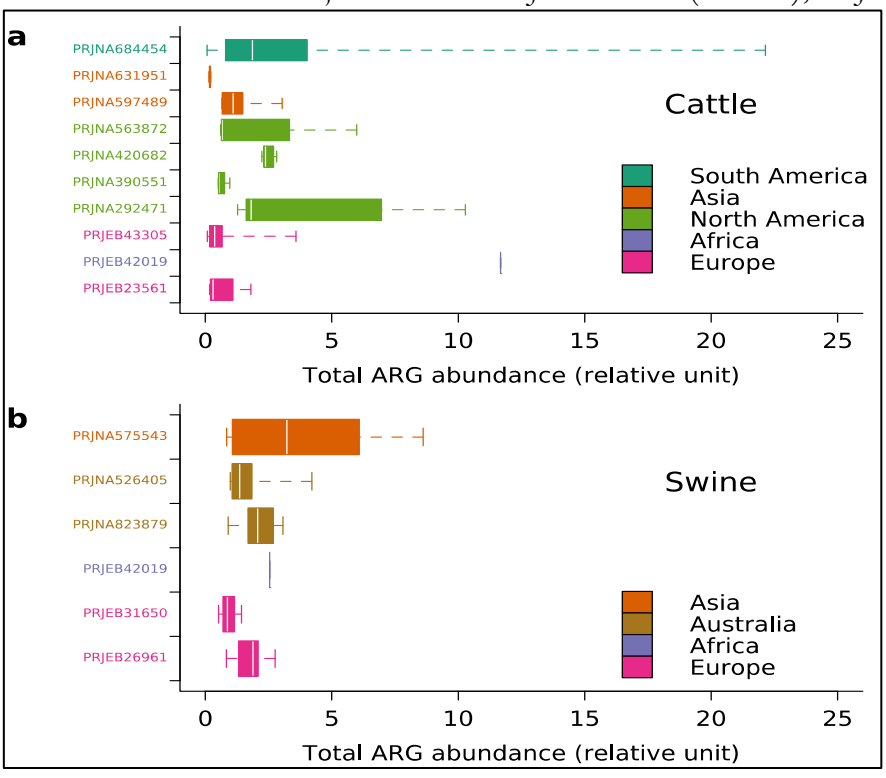

Figure 14 | **Regional Disparities in ARG Abundance for Cattle and Swine.** Total relative ARG abundance is represented for each BioProject, which are split into two subsets: cattle (A) and swine (B). Within BioProject variation is represented by the box and whisker plots. BioProjects are color coded by continent as indicated by the legends in the lower right corners.

board, most commonly ARGs correlated to tetracycline comprise the largest proportion. For PRJNA631951 (Iran), PRJNA420682 (Canada), and PRJEB23561 (France) in particular, ARGs associated with tetracycline comprise over 50% of the total ARG abundance. Contrastingly, the majority of ARGs in PRJNA292471 (USA), PRJNA390551 (USA), PRJNA563872 (USA), PRJNA684454 (Brazil), and PRJEB42019 (Malawi) are associated with classes other than macrolides, streptogramins, lincosamides, aminoglycosides, and tetracyclines. According to a report from the FDA in 2021, of all the domestic sales for antibiotics, an estimated 79% of cephalosporins, 45% of sulfonamides, 52% of aminoglycosides, and 43% of tetracyclines were intended for use in cattle. That being said, in terms of mass (kg) for antimicrobials utilized in food producing animals, cephalosporins and aminoglycosides are rather negligible while tetracyclines are the clear number one (FDA 2021). Due to the large volume and proportion of tetracyclines utilized in the USA, the identification of vast amounts of ARGs associated with tetracyclines in the composition analysis (Figure 15) is far from perplexing. However, no formal conclusion regarding the influence of antimicrobial use on AMR development could be extrapolated in this study.

3.3.2 Analysis of ARGs in metagenomic samples of swine origin

In the curated swine samples, the largest average total relative ARG abundance was found in PRJNA823879 (New Zealand), followed by PRJEB42019 (Malawi) and PRJNA526405 (Australia), respectively. Similar to cattle, as illustrated in Figure 14b, there is disparity between BioProjects in terms of within BioProject variation for ARGs in metagenomic samples for swine. The largest amount of variation, and the sample with the highest total relative ARG abundance are both found in PRJNA823879 (New Zealand). Moreover, while there is variation between BioProjects, the degree of variation is not as dramatic as the results from metagenomic samples of cattle origin. Regarding ARG composition (Figure 15), tetracyclines once again comprised the largest proportion of ARGs by drug class. Clustering of drug class proportions based on each samples' associated

country allows for differentiation to a minimal extent, a larger sample size is necessary to further differentiate countries by ARG composition for swine. Of note, the proportion of ARGs associated with tetracyclines, aminoglycosides, and lincosamides in PRJNA526405 (Australia) are consistent across samples. PRJNA823879 (New Zealand), PRJEB42019 (Malawi), and PRJNA575543 (China) all vary significantly in terms of ARG composition. In PRJEB31650 and PRJEB26961 (both Denmark), composition of ARGs in each sample generally agree, but to a lesser extent than that of PRJNA526405 (Australia). Interestingly, the prevalence of ARGs associated with lincosamides appears to be greater in PRJNA526405 (Australia) than the alternative BioProjects.

3.3.3 Limitations of the developed tool

As samples for both species were not curated for each country, it is impossible to draw any conclusion pertaining to disparity between cattle and swine on the global stage. One significant limitation imposed by the utilized experimental design. Moreover, a larger sample size is required to determine temporal dynamics of AMR at the country level. Additionally, it is important to note that the results from large-scale application of this pipeline to the SRA should merely be used for exploration. The metadata, sequence data, and sampling methods associated with samples in the SRA are highly heterogenous, which limits highly confident ARG quantification. Furthermore,

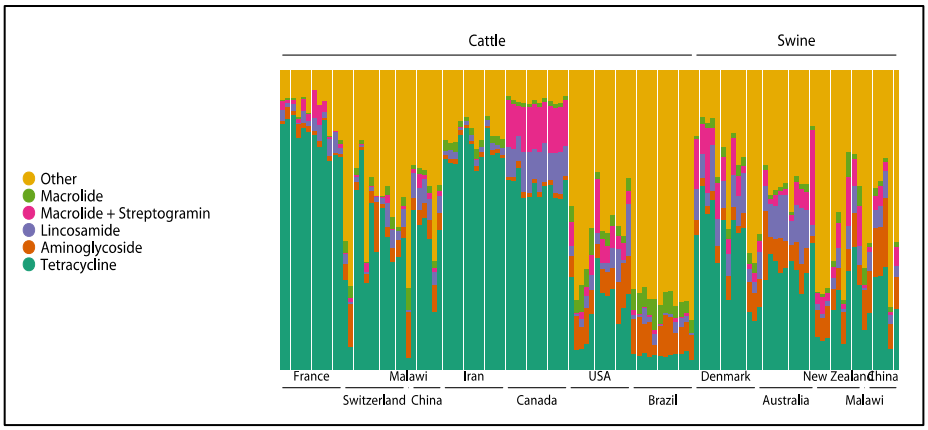

Figure 15 | **Composition of total estimated ARG abundance by drug class.** Samples from each country and species are clustered together. Composition for each antimicrobial drug class as a percent of total ARG abundance for each sample is illustrated. The most prevalent drug classes (tetracyclines, aminoglycosides, lincosamides, macrolides and streptogramins, and macrolides) are color coordinated with the graphic, any abundance of genes not partitionable into the aforementioned classes are pooled together are deemed as 'other'.

comparisons and conclusions should primarily be made at the BioProject level as larger, randomly selected, sample sizes are needed for the experiment to summarize the results at the country level.

3.4 Conclusion

In conclusion, estimated total ARG abundance and ARG composition varied between BioProjects for metagenomic samples of both cattle and swine origin. Furthermore, the degree of variation within BioProjects for both ARG abundance and ARG composition is not consistent. While the estimated abundances cannot be interpreted as absolute, this tool could prove useful to inform future studies as far determining sampling locations and exploring regional trends. Moreover, this tool could identify potentially high priority areas where actions can be made to aid the mitigation of AMR. The primary conclusion of this study is that large-scale application of the developed analytical framework for estimating ARG abundance in metagenomic samples of animal agriculture origin is feasible, and when employed conjointly with publicly housed genomic sequence data such as that found in the SRA, the spatiotemporal distribution of ARGs on a global scale can be deduced.

BIBLIOGRAPHY

Alekshun M. N. & Levy S. B. (2007). Molecular mechanisms of antibacterial multidrug resistance. Cell 128(6) 1037–1050. https://doi.org/10.1016/j.cell.2007.03.004

Andersen V. D. Aarestrup F. M. Munk P. Jensen M. S. de Knegt L. V. Bortolaia V. Knudsen B. E. Lukjancenko O. Birkegård A. C. & Vigre H. (2020). Predicting effects of changed antimicrobial usage on the abundance of antimicrobial resistance genes in finisher' gut microbiomes. Preventive Veterinary Medicine 174 104853. https://doi.org/10.1016/j.prevetmed.2019.104853

Antimicrobial Resistance Collaborators. (2022). Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. Lancet (London England) 399(10325) 629–655. https://doi.org/10.1016/S0140-6736(21)02724-0

Burrows, M., & Wheeler, D.J. (1994). A Block-sorting Lossless Data Compression Algorithm.

Bushnell Brian. (2014) BBMap: A Fast Accurate Splice-Aware Aligner. United States.

D'Costa V. M. King C. E. Kalan L. Morar M. Sung W. W. L. Schwarz C. Froese D. Zazula G. Calmels F. Debruyne R. Golding G. B. Poinar H. N. & Wright G. D. (2011). Antibiotic resistance is ancient. Nature 477(7365) 457–461. https://doi.org/10.1038/nature10388

Dadgostar P. (2019). Antimicrobial Resistance: Implications and Costs. Infection and Drug Resistance 12 3903–3910. https://doi.org/10.2147/IDR.S234610

Darby E. M. Trampari E. Siasat P. Gaya M. S. Alav I. Webber M. A. & Blair J. M. A. (2022). Molecular mechanisms of antibiotic resistance revisited. Nature Reviews. Microbiology. https://doi.org/10.1038/s41579-022-00820-y

Davies R. & Wales A. (2019). Antimicrobial Resistance on Farms: A Review Including Biosecurity and the Potential Role of Disinfectants in Resistance Selection. Comprehensive Reviews in Food Science and Food Safety 18(3) 753–774. https://doi.org/10.1111/1541-4337.12438

de Carvalho F. M. Valiatti T. B. Santos F. F. Silveira A. C. de O. Guimarães A. P. C. Gerber A. L. Souza C. de O. Cassu Corsi D. Brasiliense D. M. Castelo-Branco D. de S. C. M. Anzai E. K. Bessa-Neto F. O. Guedes G. M. de M. de Souza G. H. de A. Lemos L. N. Ferraz L. F. C. Bahia M. de N. M. Vaz M. S. M. da Silva R. G. B. … Gales A. C. (2022). Exploring the Bacteriome and Resistome of Humans and Food-Producing Animals in Brazil. Microbiology Spectrum 10(5) e00565-22. https://doi.org/10.1128/spectrum.00565-22

Dunning Hotopp J. C. (2011). Horizontal gene transfer between bacteria and animals. Trends in Genetics: TIG 27(4) 157–163. https://doi.org/10.1016/j.tig.2011.01.005

Gaio D. DeMaere M. Z. Anantanawat K. Eamens G. J. Liu M. Zingali T. Falconer L. Chapman T. A. Djordjevic S. P. & Darling A. E. (2021). A large-scale metagenomic survey dataset of the post-weaning piglet gut lumen. GigaScience 10(6) giab039.

https://doi.org/10.1093/gigascience/giab039

Gelalcha B. D. & Kerro Dego O. (2022). Extended-Spectrum Beta-Lactamases Producing Enterobacteriaceae in the USA Dairy Cattle Farms and Implications for Public Health. Antibiotics (Basel Switzerland) 11(10) 1313. https://doi.org/10.3390/antibiotics11101313

Gerber J. Ross R. Bryan M. Localio A. R. Szymczak J. Fiks A. Barkman D. Odeniyi F. Conaboy K. Bell L. Zaoutis T. & Wasserman R. (2018). Comparing Broad- and Narrow-Spectrum Antibiotics for Children with Ear Sinus and Throat Infections. Patient-Centered Outcomes Research Institute (PCORI). http://www.ncbi.nlm.nih.gov/books/NBK582423/

Gharechahi J. Vahidi M. F. Bahram M. Han J.-L. Ding X.-Z. & Salekdeh G. H. (2021). Metagenomic analysis reveals a dynamic microbiome with diversified adaptive functions to utilize high lignocellulosic forages in the cattle rumen. The ISME Journal 15(4) 1108–1120. https://doi.org/10.1038/s41396-020-00837-2

Grada A. & Bunick C. G. (2021). Spectrum of Antibiotic Activity and Its Relevance to the Microbiome. JAMA Network Open 4(4) e215357. https://doi.org/10.1001/jamanetworkopen.2021.5357

Guzman C. & D'Orso I. (2017). CIPHER: A flexible and extensive workflow platform for integrative next-generation sequencing data analysis and genomic regulatory element prediction. BMC Bioinformatics 18(1) 363. https://doi.org/10.1186/s12859-017-1770-1

Haley B. J. Kim S.-W. Salaheen S. Hovingh E. & Van Kessel J. A. S. (2020). Differences in the Microbial Community and Resistome Structures of Feces from Preweaned Calves and Lactating Dairy Cows in Commercial Dairy Herds. Foodborne Pathogens and Disease 17(8) 494–503. https://doi.org/10.1089/fpd.2019.2768

Hendriksen R. S. Munk P. Njage P. van Bunnik B. McNally L. Lukjancenko O. Röder T. Nieuwenhuijse D. Pedersen S. K. Kjeldgaard J. Kaas R. S. Clausen P. T. L. C. Vogt J. K. Leekitcharoenphon P. van de Schans M. G. M. Zuidema T. de Roda Husman A. M. Rasmussen S. Petersen B. … Aarestrup F. M. (2019). Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nature Communications 10(1) 1124. https://doi.org/10.1038/s41467-019-08853-3

Jeong J. Song W. Cooper W. J. Jung J. & Greaves J. (2010). Degradation of tetracycline antibiotics: Mechanisms and kinetic studies for advanced oxidation/reduction processes. Chemosphere 78(5) 533–540. https://doi.org/10.1016/j.chemosphere.2009.11.024

Kapoor G. Saigal S. & Elongavan A. (2017). Action and resistance mechanisms of antibiotics: A guide for clinicians. Journal of Anaesthesiology Clinical Pharmacology 33(3) 300–305. https://doi.org/10.4103/joacp.JOACP_349_15

Li J. Zhong H. Ramayo-Caldas Y. Terrapon N. Lombard V. Potocki-Veronese G. Estellé J. Popova M. Yang Z. Zhang H. Li F. Tang S. Yang F. Chen W. Chen B. Li J. Guo J. Martin C. Maguin E. … Morgavi D. P. (2020). A catalog of microbial genes from the bovine rumen unveils a

specialized and diverse biomass-degrading environment. GigaScience 9(6) giaa057. https://doi.org/10.1093/gigascience/giaa057

Li Y. Kreuzer M. Clayssen Q. Ebert M.-O. Ruscheweyh H.-J. Sunagawa S. Kunz C. Attwood G. Amelchanka S. & Terranova M. (2021). The rumen microbiome inhibits methane formation through dietary choline supplementation. Scientific Reports 11(1) 21761. https://doi.org/10.1038/s41598-021-01031-w

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: Genomics.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658–1659. https://doi.org/10.1093/bioinformatics/btl158

Lin X. Kaul S. Rounsley S. Shea T. P. Benito M. I. Town C. D. Fujii C. Y. Mason T. Bowman C. L. Barnstead M. Feldblyum T. V. Buell C. R. Ketchum K. A. Lee J. Ronning C. M. Koo H. L. Moffat K. S. Cronin L. A. Shen M. … Venter J. C. (1999). Sequence and analysis of chromosome 2 of the plant Arabidopsis thaliana. Nature 402(6763) 761–768. https://doi.org/10.1038/45471

Munk P. Andersen V.D. de Knegt L. Jensen M.S. Knudsen B.E. Lukjancenko O. Mordhorst H. Clasen J. Agersø Y. Folkesson A. et al. (2017). A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. J. Antimicrob. Chemother. 72 385–392.

Munk P. Knudsen B. E. Lukjancenko O. Duarte A. S. R. Van Gompel L. Luiken R. E. C. Smit L. A. M. Schmitt H. Garcia A. D. Hansen R. B. Petersen T. N. Bossers A. Ruppé E. EFFORT Group Lund O. Hald T. Pamp S. J. Vigre H. Heederik D. … Aarestrup F. M. (2018). Abundance and diversity of the faecal resistome in slaughter pigs and broilers in nine European countries. Nature Microbiology 3(8) 898–908. https://doi.org/10.1038/s41564-018-0192-9

Noyes N. R. Yang X. Linke L. M. Magnuson R. J. Dettenwanger A. Cook S. Geornaras I. Woerner D. E. Gow S. P. McAllister T. A. Yang H. Ruiz J. Jones K. L. Boucher C. A. Morley P. S. & Belk K. E. (2016). Resistome diversity in cattle and the environment decreases during beef production. ELife 5 e13195. https://doi.org/10.7554/eLife.13195

O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. London: Review on Antimicrobial Resistance 2016.

Poulsen C. S. Ekstrøm C. T. Aarestrup F. M. & Pamp S. J. (2022). Library Preparation and Sequencing Platform Introduce Bias in Metagenomic-Based Characterizations of

Microbiomes. Microbiology Spectrum 10(2) e00090-22. https://doi.org/10.1128/spectrum.00090-22

Quan J. Wu Z. Ye Y. Peng L. Wu J. Ruan D. Qiu Y. Ding R. Wang X. Zheng E. Cai G. Huang W. & Yang J. (2020). Metagenomic Characterization of Intestinal Regions in Pigs With Contrasting Feed Efficiency. Frontiers in Microbiology 11 32. https://doi.org/10.3389/fmicb.2020.00032

Quince C. Walker A. W. Simpson J. T. Loman N. J. & Segata N. (2017). Shotgun metagenomics from sampling to analysis. Nature Biotechnology 35(9) 833–844. https://doi.org/10.1038/nbt.3935

Rubiola S. Macori G. Chiesa F. Panebianco F. Moretti R. Fanning S. & Civera T. (2022). Shotgun metagenomic sequencing of bulk tank milk filters reveals the role of Moraxellaceae and Enterobacteriaceae as carriers of antimicrobial resistance genes. Food Research International (Ottawa Ont.) 158 111579. https://doi.org/10.1016/j.foodres.2022.111579

Schwarz S. Shen J. Kadlec K. Wang Y. Brenner Michael G. Feßler A. T. & Vester B. (2016). Lincosamides Streptogramins Phenicols and Pleuromutilins: Mode of Action and Mechanisms of Resistance. Cold Spring Harbor Perspectives in Medicine 6(11) a027037. https://doi.org/10.1101/cshperspect.a027037

Shaffer J. P. Nothias L.-F. Thompson L. R. Sanders J. G. Salido R. A. Couvillion S. P. Brejnrod A. D. Lejzerowicz F. Haiminen N. Huang S. Lutz H. L. Zhu Q. Martino C. Morton J. T. Karthikeyan S. Nothias-Esposito M. Dührkop K. Böcker S. Kim H. W. … Zhang S. (2022). Standardized multi-omics of Earth's microbiomes reveals microbial and metabolite diversity. Nature Microbiology 7(12) 2128–2150. https://doi.org/10.1038/s41564-022-01266-x

Soucy S. M. Huang J. & Gogarten J. P. (2015). Horizontal gene transfer: building the web of life. Nature Reviews Genetics 16 (8) 472-482.

Stokes H. W. & Gillings M. R. (2011). Gene flow mobile genetic elements and the recruitment of antibiotic resistance genes into Gram-negative pathogens. FEMS microbiology reviews 35 (5) 790-819.

Thomas C. M. & Nielsen K. M. (2005). Mechanisms of and barriers to horizontal gene transfer between bacteria. Nature reviews microbiology 3 (9) 711-721.

Thomas M. Webb M. Ghimire S. Blair A. Olson K. Fenske G. J. Fonder A. T. Christopher-Hennings J. Brake D. & Scaria J. (2017). Metagenomic characterization of the effect of feed additives on the gut microbiome and antibiotic resistome of feedlot cattle. Scientific Reports 7(1) 12257. https://doi.org/10.1038/s41598-017-12481-6

Tooke C. L. Hinchliffe P. Bragginton E. C. Colenso C. K. Hirvonen V. H. A. Takebayashi Y. & Spencer J. (2019). β-Lactamases and β-Lactamase Inhibitors in the 21st Century. Journal of Molecular Biology 431(18) 3472–3500. https://doi.org/10.1016/j.jmb.2019.04.002

US Centers for Disease Control and Prevention. Antibiotic resistance threats in the United States 2019. Atlanta GA: US Department of Health and Human Services 2019.

US Food and Drug Administration Center for Veterinary Medicine. Antimicrobials Sold or Distributed for Use in Food-Producing Animals 2019. US Department of Health and Human Services 2019.

US Food and Drug Administration Center for Veterinary Medicine. Antimicrobials Sold or Distributed for Use in Food-Producing Animals 2021. US Department of Health and Human Services 2021.

US Food and Drug Administration Center for Veterinary Medicine. Evaluating the Safety of Antimicrobial New Animal Drugs with Regard to their Microbiological Effects on Bacteria of Human Health Concern 2023. Rockville MD: US Department of Health and Human Services 2023.

Van Boeckel T.P. Glennon E.E. Chen D. Gilbert M. Robinson T.P. Grenfell B.T. Levin S.A. Bonhoeffer S. and Laxminarayan R. (2017). Reducing antimicrobial use in food animals. Science 357 1350–1352.

Van Goethem M. W. Pierneef R. Bezuidt O. K. I. Van De Peer Y. Cowan D. A. & Makhalanyane T. P. (2018). A reservoir of 'historical' antibiotic resistance genes in remote pristine Antarctic soils. Microbiome 6(1) 40. https://doi.org/10.1186/s40168-018-0424-5

van Hoek A. H. A. M. Mevius D. Guerra B. Mullany P. Roberts A. P. & Aarts H. J. M. (2011). Acquired antibiotic resistance genes: An overview. Frontiers in Microbiology 2 203. https://doi.org/10.3389/fmicb.2011.00203

Wibowo M. C. Yang Z. Borry M. Hübner A. Huang K. D. Tierney B. T. Zimmerman S. Barajas-Olmos F. Contreras-Cubas C. García-Ortiz H. Martínez-Hernández A. Luber J. M. Kirstahler P. Blohm T. Smiley F. E. Arnold R. Ballal S. A. Pamp S. J. Russ J. … Kostic A. D. (2021). Reconstruction of ancient microbial genomes from the human gut. Nature 594(7862) 234–239. https://doi.org/10.1038/s41586-021-03532-0

Woolhouse M. Ward M. van Bunnik B. & Farrar J. (2015). Antimicrobial resistance in humans livestock and the wider environment. Philosophical Transactions of the Royal Society of London. Series B Biological Sciences 370(1670) 20140083. https://doi.org/10.1098/rstb.2014.0083

Xue M.-Y. Xie Y.-Y. Zhong Y. Ma X.-J. Sun H.-Z. & Liu J.-X. (2022). Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. Microbiome 10(1) 32. https://doi.org/10.1186/s40168-022-01228-9

Young W. Maclean P. Dunstan K. Ryan L. Peters J. Armstrong K. Anderson R. Dewhurst H. van Gendt M. Dilger R. N. Dekker J. Haggarty N. & Roy N. (2022). Lacticaseibacillus rhamnosus HN001 alters the microbiota composition in the cecum but not the feces in a piglet model. Frontiers in Nutrition 9 1002369. https://doi.org/10.3389/fnut.2022.1002369

Zaheer R. Noyes N. Ortega Polo R. Cook S. R. Marinier E. Van Domselaar G. Belk K. E. Morley P. S. & McAllister T. A. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. Scientific Reports 8(1) 5890. https://doi.org/10.1038/s41598-018-24280-8

Zhang Z. Zhang Q. Wang T. Xu N. Lu T. Hong W. Penuelas J. Gillings M. Wang M. Gao W. & Qian H. (2022). Assessment of global health risk of antibiotic resistance genes. Nature Communications 13(1) 1553. https://doi.org/10.1038/s41467-022-29283-8

APPENDIX

SRA Sample Annotation Procedure

1. Evaluate the presence of any missing data for the sample in the local database.
    1. Collection Date (NOT submission date), Organism, Location (State/City), Source of Sample (Fecal, Intestinal, etc.), additional Accession IDs
2. Access the SRA website, search associated Accession ID, and locate samples or project.
    1. Can search any or multiple of the accession numbers (SRS, SRP, PRDJ, etc.)
    2. Information on meaning of different IDs can be found here:
3. Jump between BioProject, SRA, and BioSample hyperlinks to confirm that you have found the exact samples and or project of interest.
    1. Note similar and or contradicting information between online metadata and our local metadata. Update local database as needed.
4. If there is still missing / unconfirmed / contradicting data in our local file:
    1. Is there an explicit and exact publication stated in the SRA metadata?
        1. YES
            1. Navigate to the publication online, and review the paper:
                1. What to Find:
                    1. Sample Collection Date
                        1. Format:
                            1. DD:MM:YYYY
                        2. If there is a range of dates all in same year, just record the year:
                            1. YYYY
                        3. If it is a multi-year study, annotate each sample individually with a collection date. If per sample resolution is not available just record the more recent year (YYYY).
                    2. Host Species
                    3. Geographical Location
                        1. Format:
                            1. Farm:City:State/Province/Country
                                1. If any missing replace with "NA":
                                2. i.e. NA:NA:Michigan:USA
                    4. Source of Sample (fecal, nasal, etc.)
                    5. Additional (these can be indicated in notes section):
                        1. Were all samples collected roughly in the same area for the study? If they vary significantly then we will need resolution down to each sample for that project.
                        2. Were all samples taken from the same species? If not, are samples collected from an alternative species annotated correctly? AND are these alternative species samples included in our local database?

3. Are there duplicated sample IDs in our spreadsheet for this project? This likely indicates 16s rRNA amplicon sequencing.

b. Order for Evaluation:
    1. Title
    2. Materials and Methods
    3. Abstract
    4. Supplemental Figures and Information
    5. Data Availability
        1. Use this to check alternative areas that data information may be housed if the above four sections fail to offer insight.

ii. NO
1. Search the internet for publications tied to authors and data that may be available but not explicitly stated in the SRA metadata.
    1. Copy and paste #PRJxxxxx from SRA into google.
    2. Copy and paste abstract and or title or alternative Accession IDs into google.
    3. Once a paper has been located and confirmed use the same procedure as identified above in 4.a.i.1.
2. If still can't find publication:
    1. Send to Lee Ackerson or Dr. Wen Huang
        2. They will email authors for confirmation or discovery of the associated data.
        3. An executive decision may be made to remove the PRJ from our study due to low confidence metadata.
            1. Factors Affecting this decision:
                1. Size of study (#samples)
                2. Completeness and detail of SRA annotation by sample submitter

5. Lastly, sign initials in column saying that you finished the annotation and record any notes you feel may be significant in the designated column. Also state your confidence level in your annotation from 0%-100% in the designated column. If you are the second reviewer, note your annotation concordance (agree or disagree) with the first reviewer; if disagreement, note what deciphered metadata differs.