Universities Council on Water Resources
Journal of Contemporary Water Research & Education
Issue 160, Pages 42-59, April 2017

Searching for Antibiotic Resistance Genes in a Pristine Arctic Wetland

Krystalle S. Diaz[1,4], Virginia I. Rich[2], *Jean E. McLain[3]

[1]University of New Hampshire, Molecular, Cellular, and Biomedical Sciences, [2]Ohio State University, Department of Microbiology, [3]University of Arizona, Water Resources Research Center, [4]Work conducted at The University of Arizona, Department of Soil, Water, and Environmental Science
**Corresponding author*

Abstract: Antibiotics are the foundation of modern medicine, and the sustainability of antibiotic use depends on overcoming the challenge of antibiotic resistance (AR). This may be extremely difficult, given that antibiotic production by environmental microbiota is a natural process and AR is widespread in nature. AR has been shown to spread readily and quickly, as well as to exist in the absence of anthropogenic forcing (i.e., the application of antibiotics by humans). This study compared a range of approaches to identify AR gene sequences in microbial community genomes in three habitats (Palsa, Bog, and Fen) across a permafrost thaw gradient in Abisko, Sweden. Depending on the search algorithm used, we found a range of 0.72-1.29% AR protein-coding genes across all habitats, with model-based search algorithms resulting in greater sensitivity than traditional identity-based approaches for identifying AR sequences. Though each habitat within the permafrost gradient showed different cell densities, pH, bioavailable carbon, water content, nutrient availability, and other factors that impact microbial life, the difference in overall AR protein abundance between habitats was not significant. Examination of AR at this location will contribute to our understanding of AR presence in natural environments and how AR may be impacted by permafrost thawing related to global warming.

Keywords: *metagenomics, microbial communities, climate change, soil resistome, permafrost*

In recent years, there has been rising global concern regarding the emergence and increase of antibiotic resistant pathogens (Allen et al. 2010; D'Costa et al. 2011; World Health Organization 2014). While these concerns have been centered in health care settings, there is growing awareness that human use of antibiotics in medical, veterinary, agricultural, and wastewater treatment settings may lead to environmental spread of antibiotic resistance (AR) that is anthropogenic in origin. The spread of AR via reclaimed wastewater irrigation and its potential subsequent spread into the human health chain is an especially pressing concern in arid and semiarid ecosystems, where such irrigation is common. In the Southwest United States, reclaimed wastewater is used for environmental remediation, irrigation of golf courses and other recreational areas, and also irrigation of food crops. While bacterial indicator levels allowed in reclaimed water for different uses are carefully controlled, the levels of antibiotics and AR genes are wholly unregulated. Only the most modern municipal wastewater treatment plants are able to eliminate emerging microparticles such as antibiotics (McLain and Williams 2014), and thus antibiotics present in reclaimed wastewater could, in theory, induce survival of resistant bacteria (Wright 2007; Martinez 2009; Allen et al. 2010; Martinez 2012; Li et al. 2015). However, establishing the potential for antibiotic usage to induce AR in surrounding environments requires a clear understanding of natural, "default" AR levels and diversity existing prior to anthropogenic forcing.

Resistance to antibiotics is ancient, and evolutionary mechanisms that confer resistance in the wild are often similar to clinical variants (D'Costa et al. 2011). AR has been found in natural

environments such as caves, deep and ancient ice cores, and permafrost (reviewed in D'Costa et al. 2011). It is hypothesized that pressures such as resource competition, microbial signal trafficking, and/or detoxification may facilitate AR spread in natural environments (Dantas et al. 2008; Davies and Davies 2010; D'Costa et al. 2011; Nesme et al. 2014). Thus, assessing the potential for anthropogenically-produced antibiotics to increase resistance in natural environments must take into account the confounding factor of natural AR.

Despite the well-known presence of AR in the natural environment, studies analyzing AR spread in the environment due to anthropogenic activities such as reclaimed wastewater use or agricultural antibiotic use often lack control sites unimpacted by human activities. Though this could lead to erroneous conclusions about potential impacts of human-induced AR development, it is becoming increasingly clear that AR levels are often higher in human-impacted sites than in natural environments (Li et al. 2015). A critical research task highlighted by recent work (e.g., Durso et al. 2012; Nesme et al. 2014) is the establishment of a baseline of natural AR abundance. Such information will be crucial in addressing the question of anthropogenic influence on AR spread in natural and human-created environments alike. If AR levels and diversity in human-impacted sites differ from those in natural sites, it follows that there is basis for claims of anthropogenic AR.

A limited number of studies have analyzed AR presence in natural environments using next generation sequencing tools. Next generation sequencing tools allow scientists to analyze microbial community genes (metagenomics) and gene products and to estimate the abundance of target genes with improved accuracy. AR studies which have used metagenomic methods have found AR in clinical, agricultural, veterinary, and treated wastewater samples as well as in natural environments (Durso et al. 2012; Hu et al. 2013; Nesme et al. 2014). These studies use identity-based methods, primarily BLAST (Basic Local Alignment Search Tool), to identify similar regions between protein, DNA, or RNA sequences (Johnson et al. 2008).

Published AR sequences in gene and protein databases are dominated by AR variants from clinical samples, but environmental datasets invariably contain uncharacterized bacteria whose AR variants may prevent their identification through traditional identity-based searches. Newer, model-based approaches increase search sensitivity for metagenomic data, which is particularly useful in datasets whose sequence variants are likely to be divergent from published AR sequences. In general, environmental AR studies use permissive BLAST cutoffs (ex: ≥35% identity to known sequences), as compared to the ≥80% or ≥97% identity cutoffs typically used for classification of sequences. This allows for better identification of sequences in environmental datasets that will differ from their clinical, published counterparts.

Three metagenomics AR studies, two environmental (Durso et al. 2012; Nesme et al. 2014) and one clinical (Hu et al. 2013), have most informed the current work. In these three studies, the range of AR genes in environmental, agricultural, gut, and fecal metagenomes was 0.05-7.4% of overall metagenome content, with the highest diversity of AR found in soil metagenomes. Each of these studies uses identity-based methods. None of them analyze AR spread within or to natural habitats, only AR gene presence within publically available metagenomes. Recently, AR studies using next generation sequencing methods to analyze community AR have become more common (Fouhy et al. 2015; Li et al. 2015; Johnson et al. 2016). Johnson et al. (2016) found that AR genes on swine farms clustered together, leading to multiple AR enriched bacterial communities, regardless of bacterial community composition. The authors state such enrichment is likely due to high selective pressures from farm usage of a subset of chemicals to which the clusters confer resistance, leading to enrichment of already available genes.

Here, we demonstrate a potentially more effective method to identify AR-coding genes in metagenomic datasets than methods used in traditional identity-based approaches. In order to gauge a baseline for AR in a natural wetland, our study mined three arctic wetland metagenomes (Woodcroft et al. in prep) for predicted AR genes and their products. Examination of AR in the Swedish permafrost sites will contribute to understanding AR presence and diversity in natural wetlands. This

is critical before addressing concerns regarding human-induced AR, as human-impacted sites are expected to differ from their natural counterparts. Mere AR presence or similar AR levels would not necessarily mean AR spread, but a difference between natural and impacted AR diversity and level would be expected. The present study tested four algorithms for analyzing metagenomic data, two of them identity-based (BLAST) methods commonly used in AR metagenome studies, and two model-based (via hidden Markov models) methods, to assess the presence of AR genes along a relatively pristine permafrost thaw gradient in Abisko, Sweden. For the purposes of this paper, "pristine" includes natural sites impacted by climate change. Abisko is the site of a research station studying the impact of climate change in the region. Changes in climate have induced permafrost thaw globally (Intergovernmental Panel on Climate Change 2014), and its impact on habitat greenhouse gas emissions (McCalley et al. 2014), soil organic matter (Hodgkins et al. 2014), and microbiota (Mondav et al. 2014).

Methods

Study Site

The study site is Stordalen Mire, a thawing permafrost mosaic system near the Abisko Research Station in Abisko, Sweden. The Mire is relatively pristine, with minimal recorded human impacts, and is composed of three main habitats spanning a permafrost thaw gradient (Figure 1). These habitats are Palsa (intact permafrost which is ~30-40 cm below the surface, dry, dominated by ericaceous shrubs); Bog (intermediate permafrost thaw which is 60-90 cm below the surface, with a perched rainwater-fed water table, and dominated by *Sphagnum spp.* peat moss); and Fen (fully thawed with no detectable permafrost in the upper meter, inundated, continuous with groundwater, and dominated by sedges). In addition to thaw and water content gradients, there is also a nutrient content and microbial diversity gradient, with an increase in both from the Palsa to the Fen (McCalley et al. 2014; Mondav et al. 2014). Material was collected from triplicate cores from each of the habitats in June 2012, from three depths (Table 1) through the active surface layer of the soil above the permafrost (where present) subject to annual thaw.

DNA Extraction, Sequencing, and Metagenome Processing

DNA was extracted using a variation of the PowerMax DNA Isolation Kit #12966-10 (Mo Bio Laboratories, Inc., USA), retaining the preservation solution during lysis, as in Mondav et al. (2014). This kit extracts genomic DNA as well as plasmids, which house many known AR genes (Martinez 2008; Allen et al. 2010; D'Costa et al. 2011; McArthur et al. 2013). Bulk DNA extracts were sequenced on an Illumina HiSeq (Illumina Inc., USA), with paired-end reads. All sequences from each habitat were pooled and assembled with the Australian Centre for Ecogenomics (ACE) Assembly Pipeline (as in Mondav et al. 2014), resulting in assemblies with 567,790 contigs (overlapping sequence data reads ≥500 bp) in the Palsa, 400,114 contigs in the Bog, and 749,828 contigs in the Fen (Table 1). Metagenome annotation and protein-coding sequence prediction were carried out with Prokka (v.1.11) (Seemann 2014), using the *--metagenome* flag and otherwise default parameters. These data are summarized in Table 1, and are part of a larger dataset that describes the shifting microbiology over space and time in this ecosystem (Woodcroft et al., in preparation). The sequences recovered in this study are available at https://github.com/ksdiaz/AR_project/.

Overall Workflow

To locate AR genes in the habitat datasets, we used (i) BLAST for comparison to published studies, and (ii) the HMMer (v3.0) software package (Finn et al. 2015). HMMer employs a model-building method that compares known variants of a specific gene or protein and models conserved versus variable sites, with more weight given to matches for the conserved sites. Profile HMMs (hidden Markov models) in HMMer are statistical models of sequence alignments based on consensus of protein structure and conservation. Although many AR genes evolve naturally in soils, clinical AR sequences are likely to have diverged from their wild counterparts. However, the active sites of the proteins still accomplish the same biological tasks and are likely conserved. Thus,

environmental AR genes may be identified by homology-based approaches when they cannot be identified using identity-based measures, making an HMM-based method a useful complementary approach. Other studies have lowered BLAST cutoffs to permissive thresholds (≥35% and ≥40% identity) to circumvent the likely divergence of known AR from environmental sequences; to our knowledge, only Pehrsson et al. (2016) have targeted functionally homologous sequences.

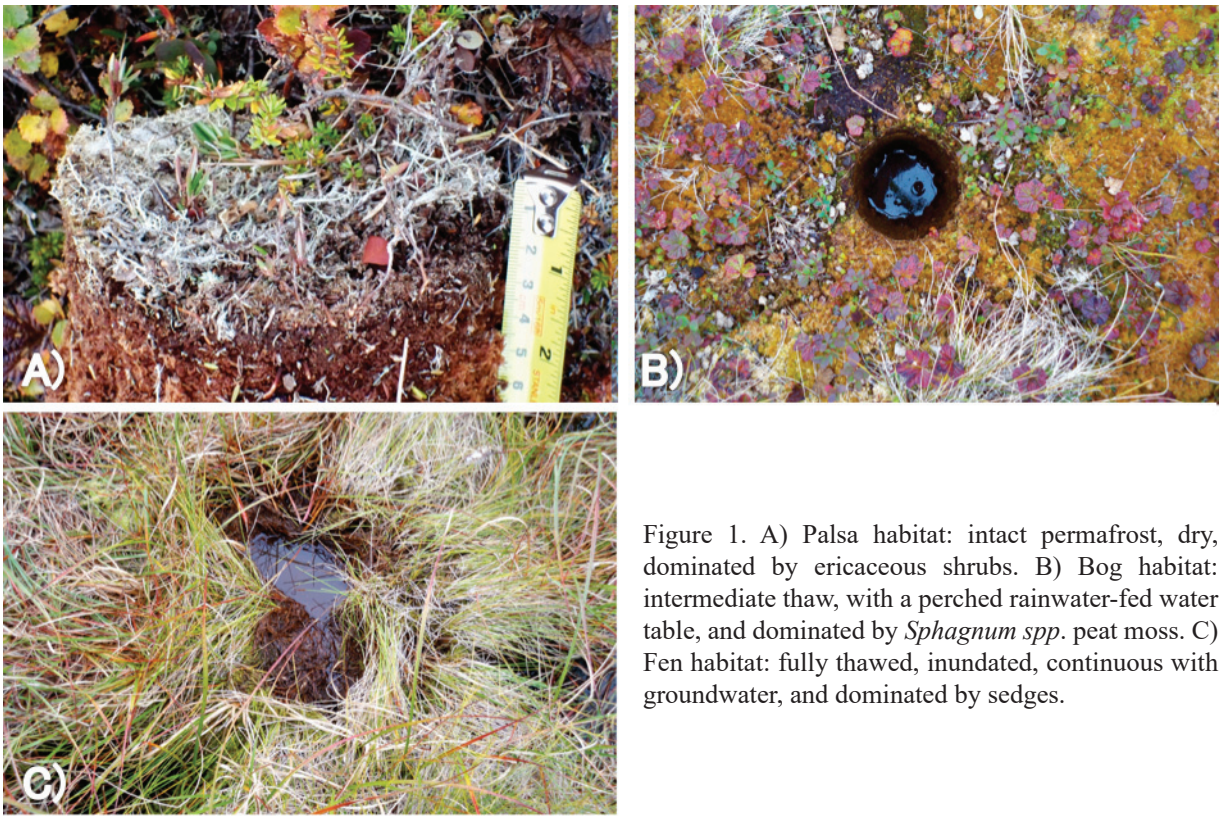

Figure 1. A) Palsa habitat: intact permafrost, dry, dominated by ericaceous shrubs. B) Bog habitat: intermediate thaw, with a perched rainwater-fed water table, and dominated by *Sphagnum spp*. peat moss. C) Fen habitat: fully thawed, inundated, continuous with groundwater, and dominated by sedges.

Table 1. Comparison of metagenome outputs to the original assemblies, generated from paired-end reads. All samples were collected in the field on 3 June 2012. The multiple sequencing depths were assembled into a single metagenome.

Habitat	Depth (name)	Depth (cm)	Megabases in raw reads	Contigs	Megabases in contigs	Prokka-predicted proteins
Palsa	Surface	1-4	1,741	567,790	289	228,334
	Middle	7-10	1,491			
	Deep	14-17	1,385			
Bog	Surface	1-4	1,994	400,114	472	685,553
	Middle	4-7	1,556			
	Deep	14-17	1,869			
Fen	Surface	1-4	1,653	749,828	995	2,038,823
	Middle	5-8	1,317			
	Deep*	11-14	15,029			

*The Fen Deep assembly was sequenced to a higher depth.

Resistance Database

The Comprehensive Antibiotic Resistance Genes Database (CARD, http://arpcard.mcmaster.ca/) contains 4,120 genes in 24 antibiotic classes for antibiotic biosynthesis, resistance, and inhibition of resistance, with 3,111 genes specific to resistance in 18 antibiotic classes (Table 2) (McArthur et al. 2013). The polypeptide and nucleotide versions of the CARD were downloaded 3 February 2015. To streamline workflow and improve search speed, the nucleotide version of the CARD was dereplicated using UCLUST's *derep_fulllength* command to remove exact copies via full-sequence matching. The dereplicated nucleotide CARD was then translated in all six reading frames using *transeq* in EMBOSS (Rice et al. 2000), and appended to the polypeptide version of the database to create a concatenated CARD with all possible representations of a given sequence (Figure 2, C1 & D1). This translation step was conducted because the metagenomes were annotated with Prokka, which predicts proteins from a given metagenome, and these would be searched against a BLAST algorithm that translates gene sequences in all six reading frames (*BLASTx*) before searching them against the provided database. This *BLASTx* algorithm was used in Hu et al. (2013) and Nesme et al. (2014). The translation step allowed for more direct comparability of HMMer results to *BLASTx*. Finally, this concatenated CARD was dereplicated again for a total of 14,808 AR polypeptide sequences. Thus, a non-redundant database with all possible versions of a known AR sequence was created. Further edits were made downstream for use in HMMer searches.

BLAST Searches: Identity-Based Methods

Identity-based searches were performed using two different BLAST algorithms (NCBI's BLAST+ v. 2.2.30) to search the metagenome assemblies for AR coding genes. Both algorithms search a protein database. *BLASTp*, is for a protein query while *BLASTx*, used for comparison to published studies (Hu et al. 2013; Nesme et al. 2014), uses a nucleotide query that it translates prior to conducting searches and assigning likelihood scores. Stricter filtering was used for *BLASTp* for a more accurate, narrower identification of potential AR-coding sequences with higher identity-based confidence, as a representation of a basic BLAST search with these data.

Lenient parameters: *BLASTx* (Figure 2, A) searched the concatenated polypeptide CARD against the prokka-predicted genes from the metagenomes, using an E-value cutoff of <1e-5, in the range of what is typical in the literature. Results were filtered for the top BLAST match for each query sequence, with percent identities ≥35% and bitscores ≥60. This level of leniency for positive matches allowed comparability to Hu et al. (2013) and Nesme et al. (2014).

Strict parameters: *BLASTp* (Figure 2, B) searched the concatenated polypeptide version of the CARD against the translations of the Prokka-predicted genes from the metagenomes. An E-value cutoff of <1e-5 was again used, but outputs were filtered to ≥80% identity, ≥70% query coverage, and bitscore ≥60 (Hu et al. 2013; Nesme et al. 2014). Unlike *BLASTx*, *BLASTp* uses input sequences that are already in protein format, so it does not translate the query into proteins in all six reading frames as it searches, thus the matches it makes are more direct and less inclusive than *BLASTx*. This search assumes the correctness of the Prokka-predicted genes in the metagenome.

Hidden Markov Models with HMMer 3.0

Model-based searches were performed using two HMM-based methods in the HMMer software suite (v3.0). At its simplest, HMMer identifies conserved versus variable regions when analyzing proteins and is designed to be a homology-based approach. Such a model may also be able to effectively capture functional similarities due to convergent evolution, and shared evolutionary histories. However, because clinical AR variants often have an environmental source (Martinez 2009; Martinez et al. 2009), targeting homology is a key approach in identifying AR analogs in the environment.

The *hmmsearch* algorithm (Figure 2, C) searches protein HMM profiles, built from a multiple sequence alignment, against a target query sequence database (Finn et al. 2011). In order to best target homologous sequences, protein profile HMMs require relatedness or functional similarity in the input protein sequences. To build the profile HMMs, the concatenated polypeptide

CARD was split into 18 antibiotic classes (Figure 2, C1 & D1; Table 2). These AR classes were grouped according to the ontology provided in the CARD download files. Only 18 of the available 24 classes denoted in the ontology files were associated with resistance genes, while the other six classes were associated with antibiotic target or biosynthesis sequences. All sequences with annotations indicating resistance to more than one antibiotic class were regrouped into the multiple

Table 2. Antibiotic classes used to create the HMMs from AR-coding and AR-associated sequences in the CARD.

Profile HMM name	Polypeptides used to build profile HMM	AR genes used to build profile HMM (CARD descriptions)
aminocoumarin	282	Point mutations to DNA gyrase subunit B (gyrB) can result in resistance to aminocoumarins. In some cases, expression of parY(R), which encodes an aminocoumarin resistant topoisomerase IV, can also confer aminocoumarin resistance.
aminoglycoside	1,170	Genes conferring resistance to aminoglycoside antibiotics.
beta-lactam	8,058	Genes conferring resistance to beta-lactams. Genes that directly or indirectly modulate beta-lactam resistance.
chloramphenicol	54	Genes conferring resistance to chloramphenicol.
ethambutol	42	Genes conferring resistance to ethambutol.
fluoroquinolone	1,236	Genes conferring resistance to fluoroquinolones.
fosfomycin	78	Genes conferring resistance to fosfomycin.
glycopeptide	408	Genes conferring resistance to glycopeptides.
lincosamide	78	Genes conferring resistance to lincosamides.
macrolide	186	Genes or their products conferring resistance to macrolide antibiotics.
multiple[a,b]	2,040	Resistance to antibiotics is often conferred by single nucleotide polymorphisms (SNPs) and other mutations in target genes. Efflux proteins that pump antibiotic out of a cell to confer resistance. Enzyme that catalyzes the inactivation of an antibiotic resulting in resistance. Inactivation includes chemical modification, destruction, etc. Genes that directly or indirectly change rates of antibiotic efflux. Enzymes that confer resistance by modifying antibiotic targets. Genes involved in restructuring of the cell wall.
polymyxin	18	Genes conferring resistance to polymyxins.
rifampin	132	Genetic determinants which confer resistance to rifampin.
streptogramin	492	Genes conferring resistance to streptogramin antibiotics.
sulfonamide	300	Genes conferring resistance to sulfonamide antibiotics.
tetracycline	66	Genes conferring resistance to tetracycline antibiotics.
tunicamycin	6	Genes that confer resistance to tunicamycin.
unspecified[a]	162	A gene whose product inactivates an antibiotic by formation of a complex, preventing interaction of the antibiotic with its target. Genes that are involved in conferring self resistance to antibiotics. Genes directly or indirectly reducing overall permeability to antibiotics.

[a]Searched against UniProt if not linked to a specific antibiotic class. Gene products that led back to a single, specific class were re-assigned their category before building the profile HMM.
[b]Multiple includes any gene product associated with two or more distinct classes of antibiotics.

resistance category, and those without specificity to an antibiotic class but annotated as "antibiotic resistance" in CARD were classed as unspecified AR sequences. *Hmmsearch* requires an alignment from which to build profile HMMs, so each CARD subset database was aligned using the algorithms Muscle, Mafft, and Clustal Omega, so that there were three multiple sequence alignments (Figure 2, C2). Muscle creates pairwise alignments between protein sequences, scores them, and chooses the best alignment before repeating the step until it has found the best match, then moves on to the next set of alignments (Edgar 2004). Mafft aligns on the basis of sequence similarity, but builds evolutionary information into its alignment strategy (Katoh and Standley 2013). Clustal Omega uses HMMs to build its alignments (Sievers et al. 2011). Using these three aligners minimizes the chance of false positives.

Mafft was run on each of the CARD subset databases using the parameters *mafft --localpair --maxiterate* 1000, as recommended for <200 sequences in the Mafft manual. For Muscle, each CARD subset database was run with defaults and the *-maxiters 2* for large datasets and better accuracy and speed. Default parameters were

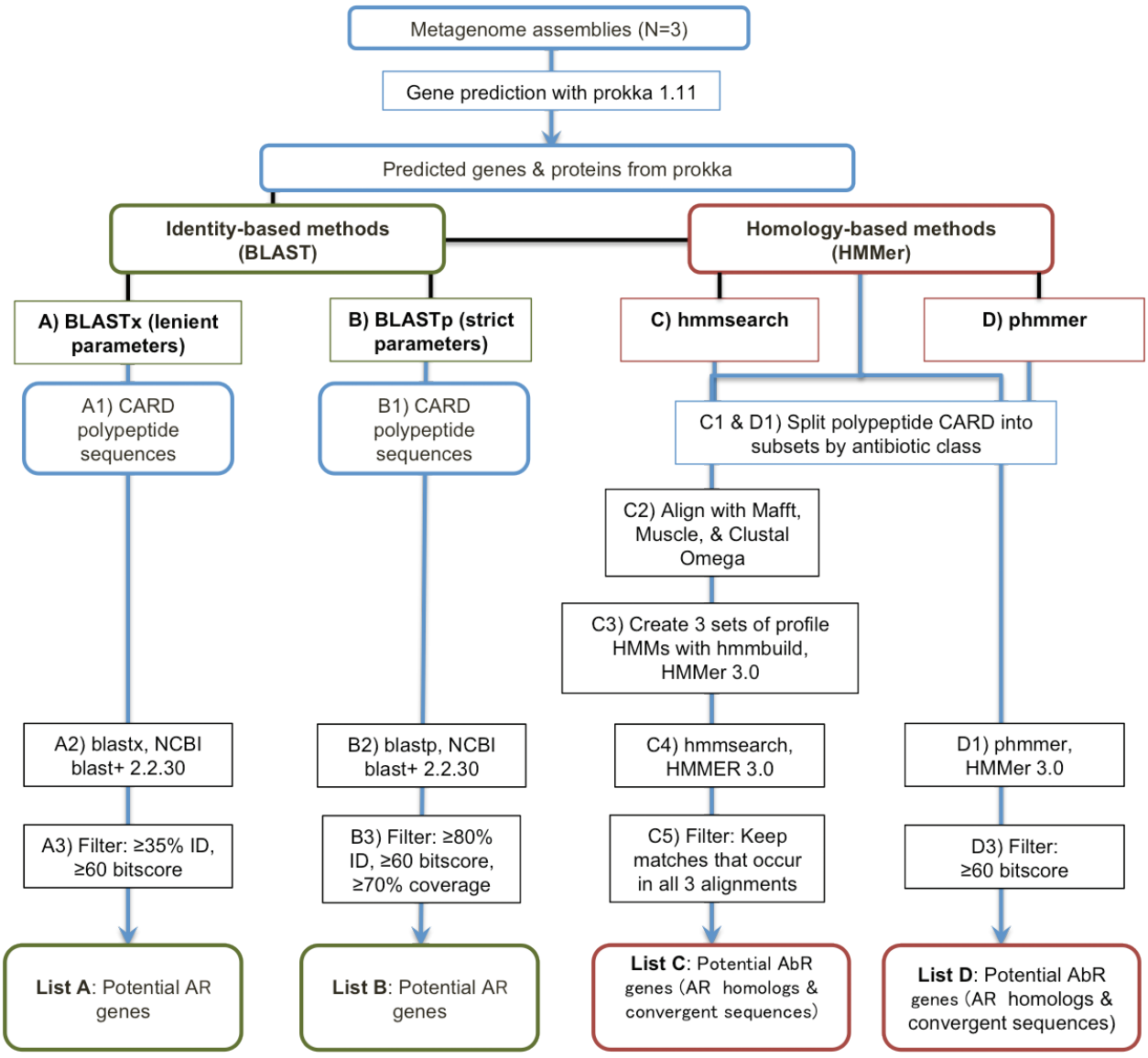

Figure 2. Bioinformatic workflow to identity potential AbR-coding sequences and homologs in wetland metagenomes.

used for Clustal Omega. Following the initial runs, all three AR alignment sets were run through HMMer's *hmmbuild* ("profile HMM construction from multiple sequence alignments") to create three profile HMMs, one from each aligner, for each set of AR proteins (Figure 2, C3). These profile HMMs were then used to search the metagenomes in *hmmsearch*, using command *hmmsearch -o <logfile.txt> --tblout <output_table.txt> -E 1e-5 --cpu 2 <query hmm file> <target seqfile>*. *Hmmsearch* outputs two bitscores, one for the sequence-HMM match and one for the best-scoring domain in the sequence, both of which we filtered to ≥60. In order to maintain a conservative dataset, only sequences that matched CARD profile HMMs from all three aligners were retained (Figure 2, C5).

Similar to *BLASTp*, *phmmer* (Figure 2, D) takes protein sequences as a query and searches them against a database, but *phmmer* converts the query sequences into profile HMMs as it searches for matches. The polypeptide CARD sequences were used to build the profile HMMs during the *phmmer* search (Figure 2, D1). Metagenomes were utilized as a database when running *phmmer*, as this results in searches of the metagenome for homologs to known AR sequences and iteratively generates metagenome-specific HMMs used to then collect more matches. The alternative of scanning known AR sequences for homologs to protein-coding sequences in an environmental dataset could lead to a higher frequency of spurious identifications.

Five Key Resistance Proteins

In order to gain a clearer idea of genetic diversity of microorganisms within the habitats, four frequently occurring AR-conferring proteins and an additional protein conferring resistance to vancomycin and teicoplanin were chosen for more detailed analysis (Table 3). These were DHPS (dihydropteroate synthase), DNA gyrA (DNA gyrase subunit A), MdtC (multidrug resistance protein MdtC), BepG (efflux pump membrane transporter BepG), and VanA (vancomycin/teicoplanin A-type resistance protein VanA). Vancomycin and teocoplanin are two broad-spectrum glycopeptide class antibiotics used in clinics to combat *Clostridium* infections and methicillin-resistant *Staphylococcus aureus*.

Results

In the literature, the range of AR abundances in gut, fecal, agricultural, and environmental metagenomes ranges from 0.05-7.4% of total gene content, with a trend of lower abundance but higher diversity in environmental samples, and higher abundance but predictable diversity in agricultural and fecal samples (Durso et al. 2012; Hu et al. 2013; Nesme et al. 2014). In our study, *hmmsearch* (model-based; Figure 2, List C) and *BLASTp* (identity-based; Figure 2, List B) returned results consistently below 0.05%, but *BLASTx* (identity-based; Figure 2, List A) and *phmmer* (model-based; Figure 2, List D) searches returned abundances between 0.72-0.75% and 0.84-1.29%, respectively, consistent with the current literature.

With both identity- and model-based methods, no significant differences were found in overall AR abundance between habitats (one-way ANOVA, n=3 habitats, p=0.96). There were, however, differences in AR class and gene types between habitats. The Bog metagenome yielded more instances of two efflux pump-related genes, BepG and MdtC, than did the other habitats, at 0.03% overall gene abundance for MdtC and 0.01% for BepG as calculated by *phmmer*. The other habitats do not display this trend (Supplemental Figure 1). Search method significantly impacted the identification of potential AR sequences, with *phmmer* consistently returning the highest AR diversity and abundance (ANOVA, $p<0.001$).

In *BLASTx*, the AR class in the highest proportion, out of total Prokka-predicted genes, was multiple antibiotic resistance (average=0.33%), followed by aminocoumarin (average=0.09%), macrolide (average=0.07%), and glycopeptide (average=0.05%) resistance (Figure 3). *Phmmer* returned a similar distribution of AR classes for multiple (average=0.56%), aminocoumarin (average=0.09%), macrolide (average=0.12%), and glycopeptide (average=0.07%) resistance. Too few matches were obtained from *BLASTp* and *hmmsearch* to identify any trends. Despite similar overall AR, protein types within AR classes (multiple resistance, aminocoumarin resistance, macrolide resistance, etc.) in the *phmmer* and *BLASTx* results had differing protein identifications within their AR matches in all of the habitats

Table 3. Five key resistance-conferring proteins and the antibiotic or antibiotic class they confer resistance to. DHPS, DNA gyrA, and MdtC are ordinarily targets of the antibiotics in the table, but these proteins are resistant mutants. All information from UniProt (http://www.uniprot.org/) and InterPro (http://www.ebi.ac.uk/interpro/) databases.

Abbreviation	Protein name	Resistance to	Function	Description
DHPS	Dihydropteroate synthase	sulfonamides	cellular metabolic process	Enzyme that catalyzes the second to last step in folic acid biosynthesis.
DNA gyrA	DNA gyrase subunit A	fluoroquinolones	DNA topological change, ATP binding, DNA binding	DNA gyrase negatively supercoils closed circular double-stranded DNA in an ATP-dependent manner and also catalyzes the interconversion of other topological isomers of double-stranded DNA rings, including catenanes and knotted rings.
MdtC	Multidrug resistance protein MdtC	novobiocin (aminocoumarin), deoxycholate, acriflavine, beta-lactams, chloramphenicols	efflux pump, multidrug resistance	The MdtABC tripartite complex confers resistance against novobiocin and deoxycholate. Multidrug resistance proteins are molecular pumps that participate in a low energy shock adaptive response in some bacteria. The MdtBC complex is iInvolved in the resistance of gram negative bacteria to novobiocin, deoxycholate and beta-lactams.
BepG	Efflux pump membrane transporter BepG	deoxycholate, nalidixic acid (fluoroquinolone), other	efflux pump, transporter activity	May contribute to resistance to some drugs, such as deoxycholate, sodium dodecyl sulfate, and nalidixic acid, in the absence of BepD and BepE.
VanA	Vancomycin/ teicoplanin A-type resistance protein VanA	glyocpeptides (vancomycin & teicoplanin)	peptidoglycan biosynthesis, ATP binding, DNA binding	Required for high-level resistance to glycopeptide antibiotics. D-Ala- -D-Ala ligase of altered specificity which catalyzes ester bond formation between D-Ala and various D-hydroxy acids; produces a peptidoglycan which does not terminate in D-alanine but in D-lactate, thus preventing vancomycin or teicoplanin binding.

(Supplemental Figure 3), with an average of 39% shared protein identifications (Supplemental Figure 2).

Four Top AR Classes

In order to analyze the completeness of pathways with AR involvement within the site metagenomes, we focused a more detailed analysis on the four most abundant AR classes in the data: multiple, macrolide, aminocoumarin, and glycopeptide resistance. The KEGG (Kyoto Encyclopedia of Genes and Genomes) is a resource for understanding high-level functions of microorganisms and their utility to ecosystem function from molecular information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies (Tanabe and Kanehisa 2012). Across the four resistance classes, in all three habitats KEGG molecular pathways from the *phmmer* results were consistently more complete than the *BLASTx* results, with an average of 46.0% of the multiple, 79.2% of the macrolide, 59.5% of the aminocoumarin, and 58.9% of the

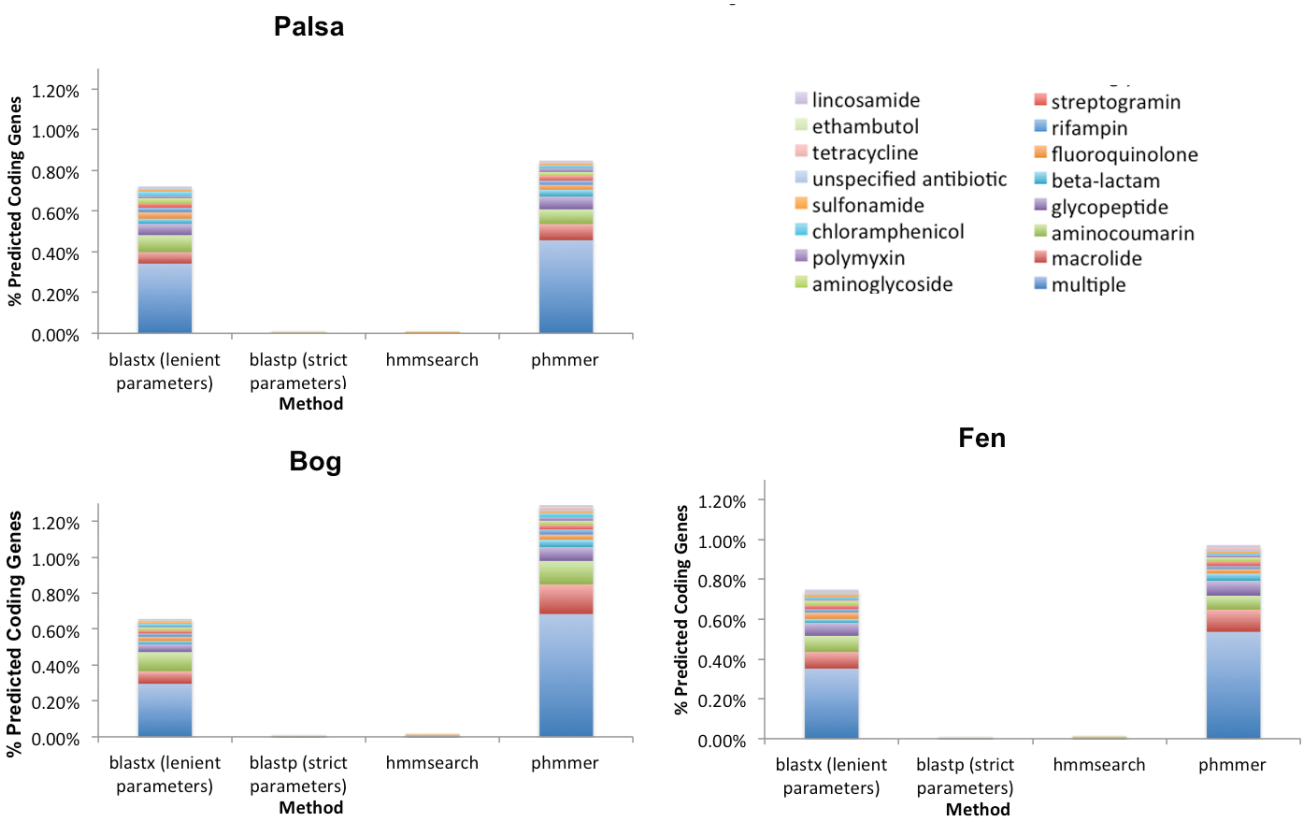

Figure 3. Abundance of hits, by AR detection method, normalized to their respective habitats. Both *BLASTx* and *phmmer* returned results within the ranges expected from the literature, with similar overall abundances to particular antibiotic classes, but their individual hits were quite different (Supplemental Figures 1-6).

glycopeptide *phmmer* matches annotated to a KEGG functional category. In comparison, 38.5% of the multiple, 71.4% of the macrolide, 36.8% of the aminocoumarin, and 34.8% of the glycopeptide *BLASTx* matches were annotated on average (Supplemental Figures 4-7). This is possibly due to the HMM building involved in HMMer. KEGG maps of the pathways in the multiple, macrolide, aminocoumarin, and glycopeptide matches demonstrate some difference in pathways available to pump out antibiotics from a bacterial cell, with more pathways present in the *phmmer* matches than in the *BLASTx* matches, all in either the two-component system or the ABC transporter system. The two-component system is a stimulus-response system, and the ABC transporter system is a protein family that transports microbial substrates. Multiple, aminocoumarin, and glycopeptide resistance matches mapped to the two-component system. Multiple resistance and macrolide matches mapped to pathways in the ABC transporter system

Five Key Resistance Proteins

With the five key Prokka-predicted proteins DHPS, DNA gyrA, MdtC, BepG, and VanA, abundances of each gene chosen were relatively stable among the three habitats for *phmmer* and *BLASTx*, but not for *hmmsearch* or *blastp,* due to the low number of AR identifications. BepG and VanA were not represented in the *hmmsearch* and *BLASTp* hits of any habitat (Supplemental Figure 1). There were more MdtC identifications in *phmmer* for the Bog than for the other habitats and methods (≥0.01% in the Bog, 0.01% in the Palsa and Fen habitats, versus the *BLASTx* results of 0.01% for Palsa and Fen and 0.02% in the Bog). There were two hits to vancomycin resistance in the Palsa with both *BLASTx* and in *phmmer* (≤0.01% abundance). Interestingly, despite the low VanA hits, vancomycin pathways were found in all of the glycopeptide hits for each habitat, although they were less complete in the *BLASTx* hits than in the *phmmer* hits.

Discussion

Although each habitat has different cell densities, pH, bioavailable carbon, water content, nutrient availability, and other factors that impact microbial life, the difference in overall AR proteins between habitats was not significant. Difference in resistance to the four most abundant classes (multiple, macrolide, aminocoumarin, and glycopeptide resistance) was also not significant. In the Abisko permafrost mosaic system, the overall AR gene product abundance was constant across habitats. Multiple resistance was the most abundant category due to efflux pumps, which, although not plasmid-encoded, serve a general purpose in the cell, thus making any change in response to any individual antibiotic a potential shift to resistance to multiple antibiotic classes (Webber & Piddock 2003; Martinez et. al 2009). KEGG functional annotations of the multiple macrolide, aminocoumarin, and glycopeptide resistance matches in each habitat did indicate some difference in pathways available for potential AR responses.

Conclusion

We have demonstrated a more effective method with higher sensitivity than traditional BLAST-based approaches for identifying potential AR genes in metagenomic datasets, particularly for environmental datasets whose sequence variants are most likely to be divergent from the CARD and other gene and protein databases. The *phmmer* algorithm yielded more potential matches to the CARD database, and those matches spanned more complete KEGG pathways than their *BLASTx*-derived counterparts. Results using this methodology provide valuable information about the presence and completeness of AR pathways, but they do not yield information regarding the activity or non-activity of those processes in the metagenome sites. Additional study using metaproteomics or metatranscriptomics would enhance information regarding AR, enabling a contrast between what genes are available and what genes are being expressed, as it is conceivable that AR itself may not be occurring even if AR genes are spread into a habitat due to human activities. Such a case would open another set of public health and environmental health concerns regarding the potential for anthropogenically-induced AR spread.

Further research on the topic would allow for examination of environments where AR is suspected to have spread. Culturing samples for this sort of analysis is not practical, but sequencing small samples to analyze with HMMer, using comparable environmental AR levels and diversities, could establish the presence of AR potentially changed from a prior state. This has practical applications as well. For example, in Arizona and the Southwestern United States, reclaimed wastewater irrigation in agriculture is a common practice to conserve water resources (Arizona Department of Water Resources 2014), and there is concern about AR spreading within an irrigated field and to the surrounding environment via runoff, similar to concerns regarding disposal of medical and veterinary waste. Bacterial indicator levels allowed in reclaimed water for different uses are carefully controlled in the Southwest, but the levels of antibiotics and AR genes are unregulated, and thus antibiotics present in reclaimed wastewater could induce survival of resistant bacteria (Wright 2007; Martinez 2009; Allen et al. 2010; Martinez 2012; Li et al. 2015). Currently, these are only concerns, not all of which have been addressed, and there are currently no standardized tests to analyze any possible effects if antibiotics are present in the environment (Roose-Amsaleg and Laverman 2015). As AR continues to evolve in the clinic, it is time to begin addressing these concerns and anticipate those that will arise.

Acknowledgements

We thank Drs. Gene Tyson and Ben Woodcroft of the Australian Ecogenomics Centre for kindly sharing their unpublished metagenomic data for this study.

Author Bio and Contact Information

Krystalle Diaz is pursuing a Master's Degree in Microbiology at the University of New Hampshire. She has a Bachelor of Science in Environmental Science from the University of Arizona, where she began undergraduate research in the soil microbial ecology laboratory of Virginia Rich, (Ph.D.). This work led to

interest in the microbial community genomics in natural environments experiencing change from anthropogenic influence and their potential recoveries. She may be contacted at ksdiaz17@gmail.com or Gregg Hall 431, University of New Hampshire, 36 Colovos Road, Durham, New Hampshire 03820.

Virginia Rich has been an Assistant Professor of Microbiology at Ohio State since June 2015 after moving from the University of Arizona. Virginia's lab uses molecular microbial ecology techniques and strong partnerships with biogeochemists, informaticians, and modelers to investigate how microbes respond to, and mediate, environments undergoing anthropogenic change. Her team works towards a "systems-level" understanding of microbial community interactions with the environment. Virginia holds a Bachelor of Arts degree from the University of California at Berkeley, with majors in Integrative Biology and in Molecular and Cell Biology: Genetics. She earned a Ph.D. in Microbial Oceanography from the Joint Program of the Massachusetts Institute of Technology and Woods Hole Oceanographic Institute. She may be contacted at virginia.isabel.rich@gmail.com or 105 Biological Sciences Building, 484 W. 12th Ave, Columbus, OH 43210.

Jean McLain has been the Associate Director and an Associate Research Scientist at University of Arizona Water Resources Research Center since 2011 following a 10-year career with the USDA-Agricultural Research Service. Jean's work in environmental microbiology includes managing an international research program focused on establishing the human health and environmental risks of using recycled municipal wastewater for crop irrigation, replenishment of surface water, and groundwater recharge. Jean holds a Bachelor of Science Degree in Forestry from the University of Vermont, and a Master's Degree in Forest Science from Yale University. She received her Ph.D. from Duke University in Microbial Ecology in 2002. She may be contacted at mclainj@email.arizona.edu or 350 N. Campbell Avenue, Tucson, Arizona 85719.

Supplemental Figures (S1-S7)

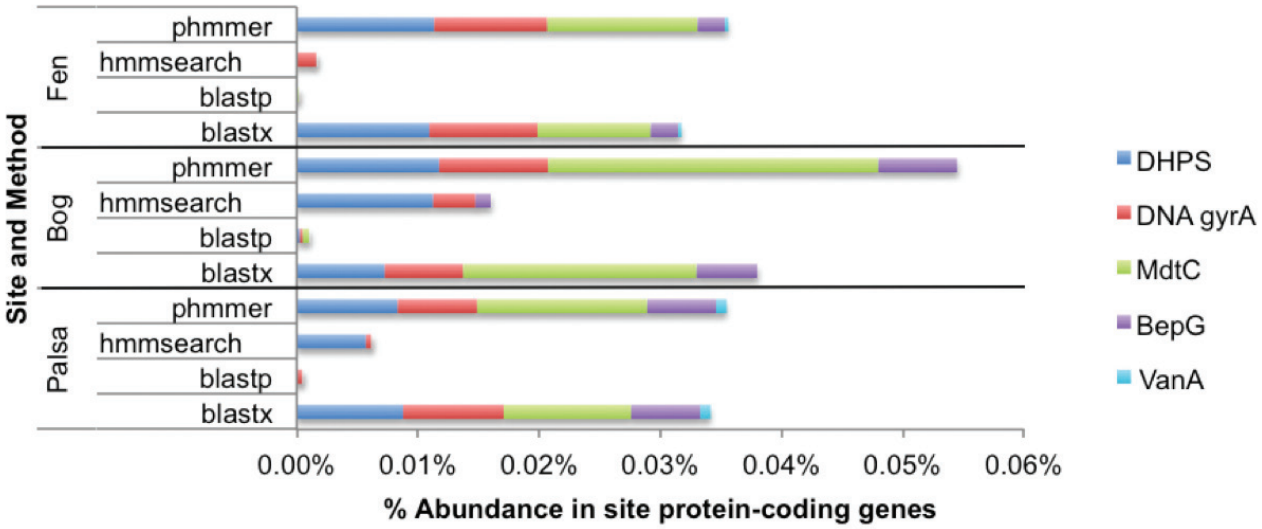

Figure S1. Five key resistance proteins, chosen for their conferring of AR, and their identification in each method and habitat.

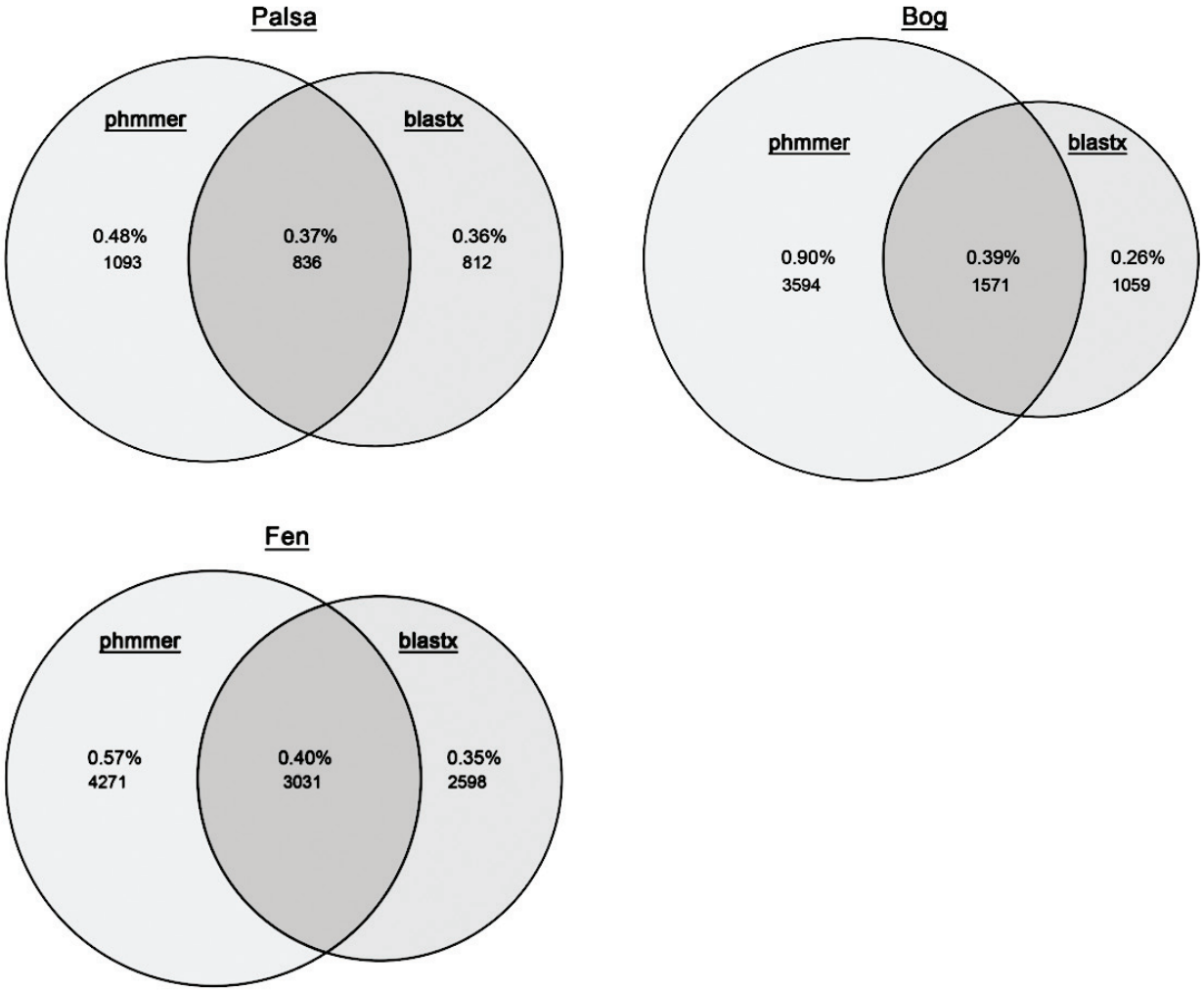

Figure S2. Prokka-predicted proteins caught by *phmmer* and *BLASTx* searches. Percentages are normalizations to the habitat's prokka-predicted proteins, not to the overall AR abundance.

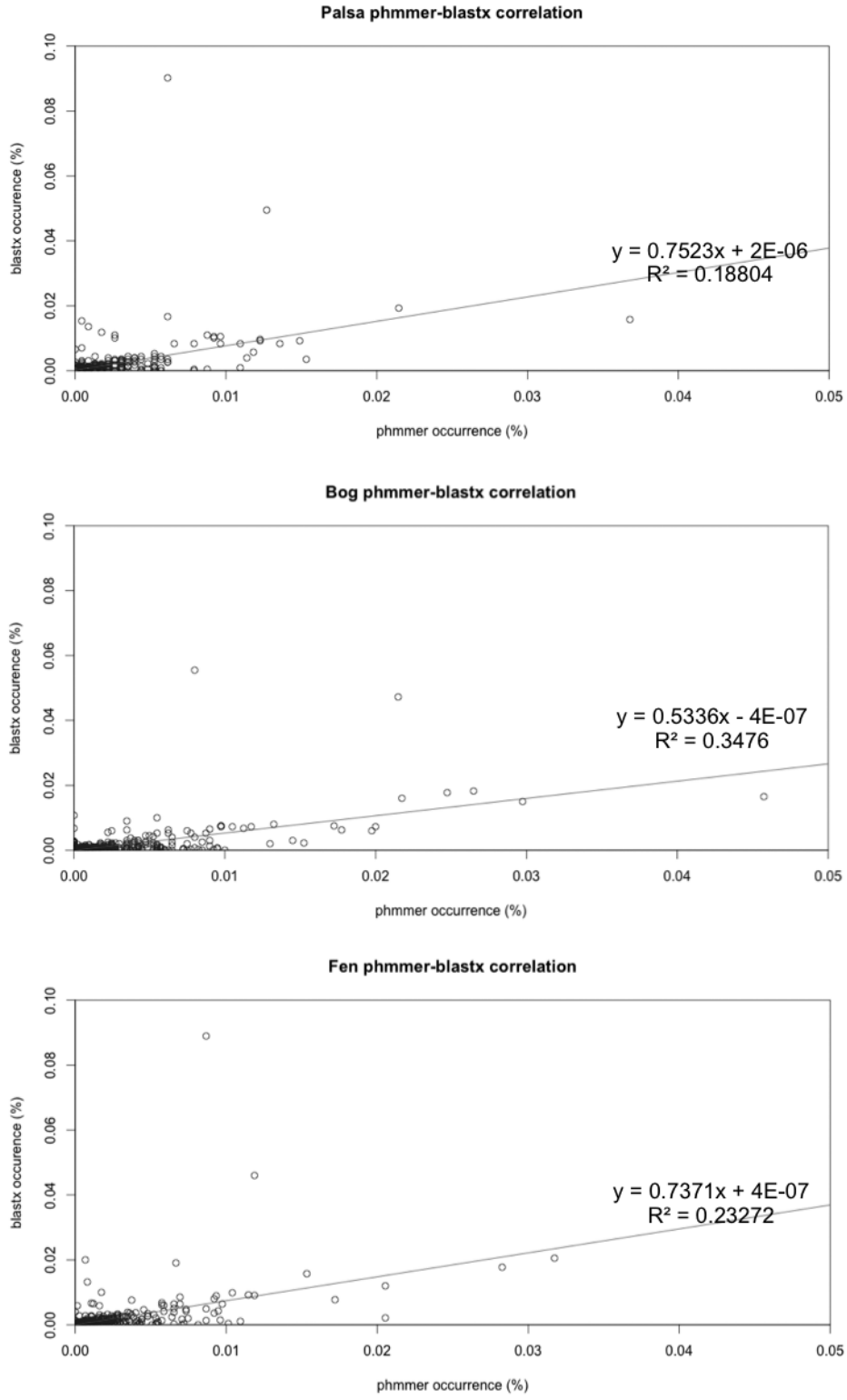

Figure S3. Correlations between *phmmer* and *BLASTx* results for each unique protein in antibiotic classes.

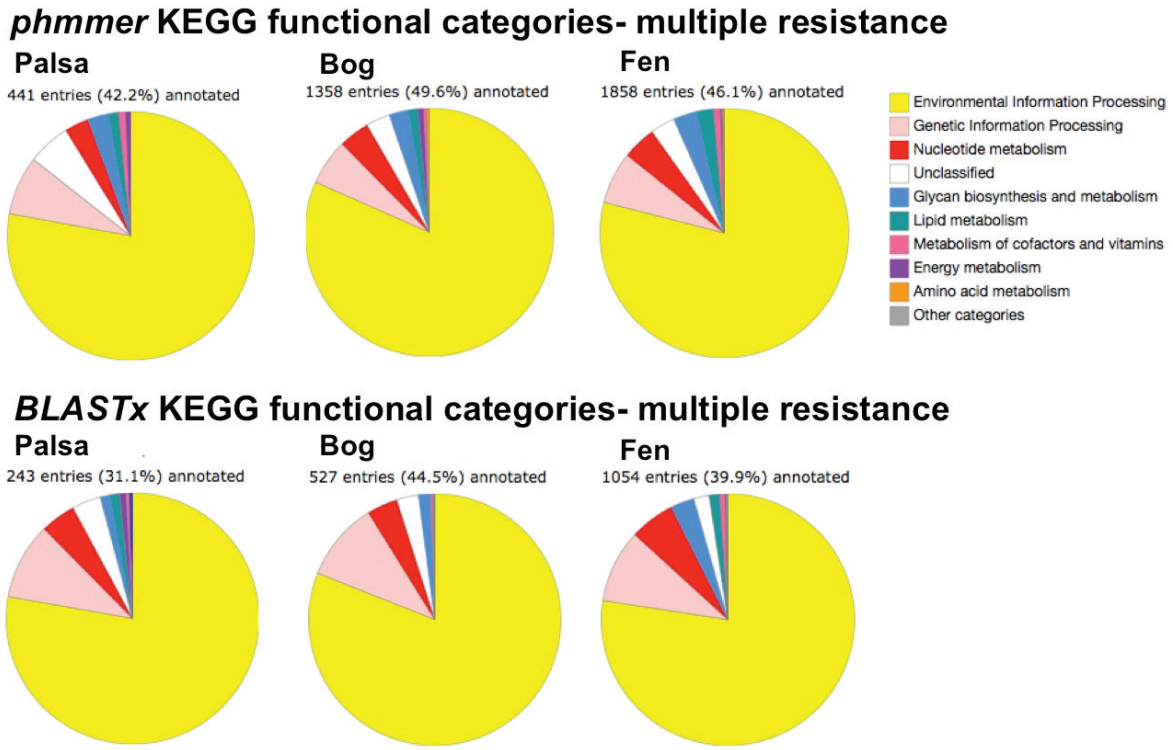

Figure S4. KEGG pie charts, displaying the number of sequences and the percent of hits from that subset of data that KEGG was able to link to a particular function. Note that despite being very similar in overall distribution, a higher percentage of sequences was able to be mapped in *phmmer* than in *BLASTx*.

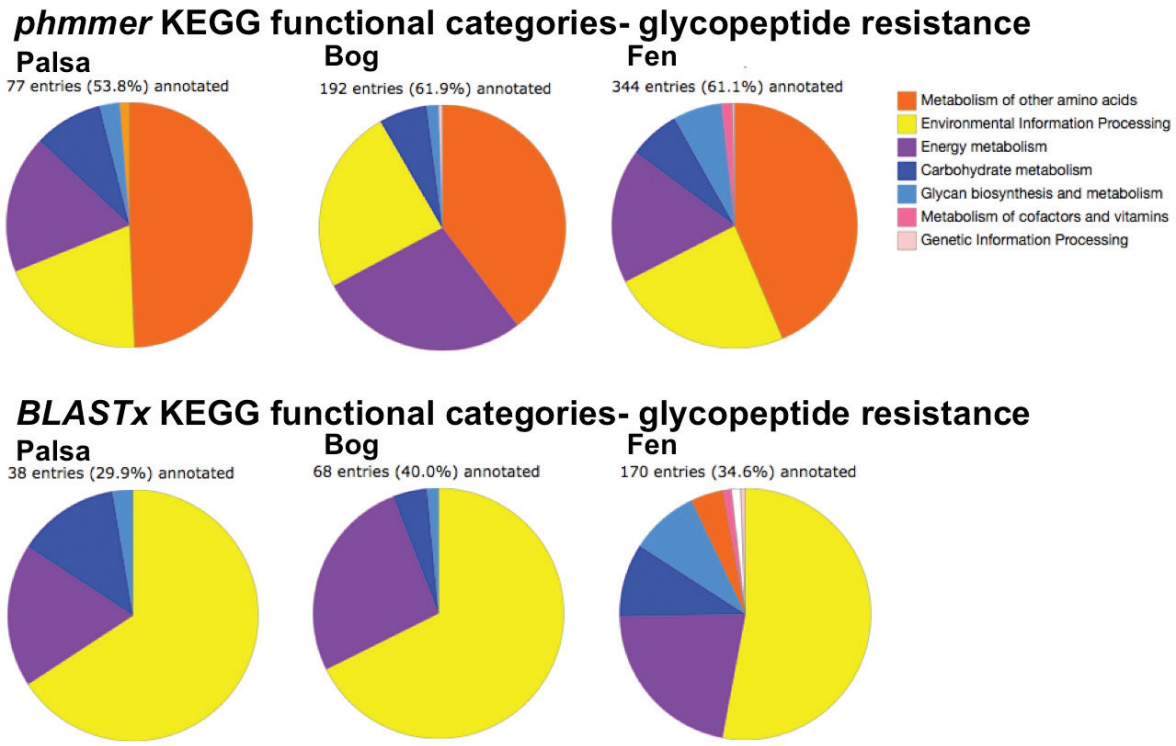

Figure S5. KEGG functionality pie charts, displaying the number of sequences and the percent of hits from that subset of data that KEGG was able to link to a particular function. A higher percentage of sequences was able to be mapped in *phmmer* than in *BLASTx*. Unlike the other antibiotic classes, distribution of functionality between *BLASTx* and *phmmer* hits was quite different in the glycopeptides than in the other classes.

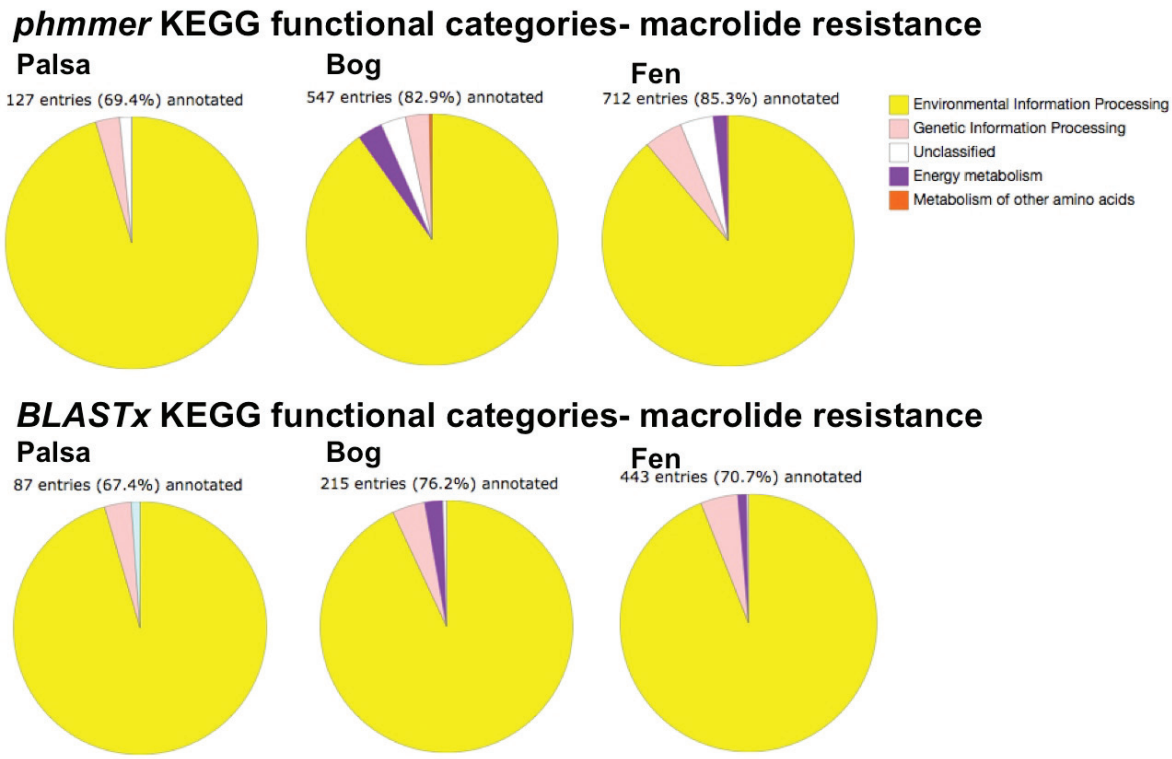

Figure S6. KEGG functionality pie charts, displaying the number of sequences and the percent of hits from that subset of data that KEGG was able to link to a particular function. Note that despite being very similar in overall distribution, a higher percentage of sequences was able to be mapped in *phmmer* than in *BLASTx*.

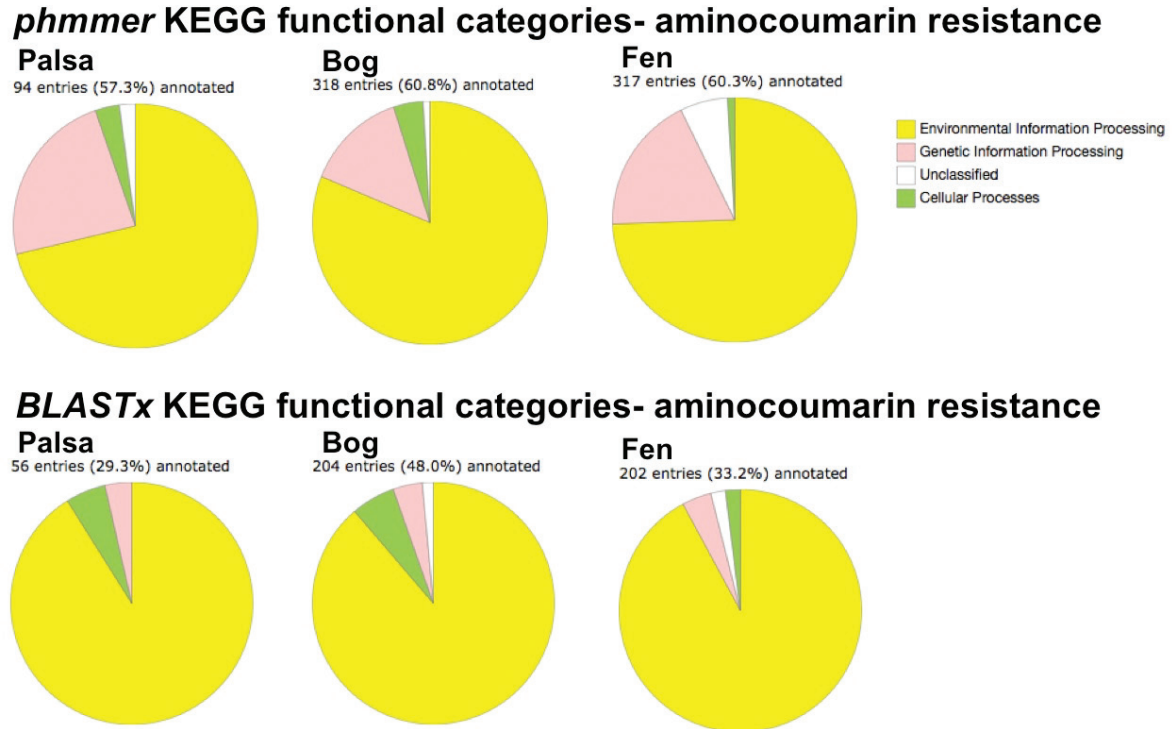

Figure S7. KEGG functionality pie charts, displaying the number of sequences and the percent of hits from that subset of data that KEGG was able to link to a particular function. Note that despite being very similar in overall distribution, a higher percentage of sequences was able to be mapped in *phmmer* than in *BLASTx*.

References

Arizona Department of Water Resources. 2014. Active Management Area Water Supply - Effluent and Contamination Sites. Available at: http://www.azwater.gov/azdwr/StatewidePlanning/WaterAtlas/ActiveManagementAreas/PlanningAreaOverview/WaterSupplyEffluent.htm. Accessed February 21, 2017.

Allen, H.K., J. Donato, H.H. Wang, K.A. Cloud-Hansen, J. Davies, and J. Handelsman. 2010. Call of the wild: Antibiotic resistance genes in natural environments. *Nature Reviews Microbiology* 8(4): 251-259. Available at: http://dx.doi.org/10.1038/nrmicro2312. Accessed February 21, 2017.

D'Costa, V.M., C.E. King, L. Kalan, M. Morar, W.W.L. Sung, C. Schwarz, D. Froese, G. Zazula, F. Calmels, R. Debruyne, G.B. Golding, H.N. Poinar, and G.D. Wright. 2011. Antibiotic resistance is ancient. *Nature* 477(7365): 457-461. Available at: http://dx.doi.org/10.1038/nature10388. Accessed February 21, 2017.

Dantas, G., M.O.A. Sommer, R.D. Oluwasegun, and G.M. Church. 2008. Bacteria subsisting on antibiotics. *Science* 320(5872): 100-103.

Davies, J. and D. Davies. 2010. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews: MMBR 74*(3): 417-433. Available at: http://doi.org/10.1128/MMBR.00016-10. Accessed February 21, 2017.

Durso, L.M., D.N. Miller, and B.J. Wienhold. 2012. Distribution and quantification of antibiotic resistant genes and bacteria across agricultural and non-agricultural metagenomes. *PLoS ONE* 7(11): 1-12.

Edgar, R.C. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* 32(5): 1792-1797. Available at: http://doi.org/10.1093/nar/gkh340. Accessed February 21, 2017.

Finn, R.D., J. Clements, W. Arndt, B.L. Miller, T.J. Wheeler, F. Schreiber, A. Bateman, and S.R. Eddy. 2015. HMMER web server: 2015 update. *Nucleic Acids Research* 43(W1): W30-W38. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkv397. Accessed February 21, 2017.

Finn, R.D., J. Clements, and S.R. Eddy. 2011. HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Research* 39(suppl_2): W29-W37. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr367. Accessed February 21, 2017.

Fouhy, F., C. Stanton, P.D. Cotter, C. Hill, and F. Walsh. 2015. Proteomics as the final step in the functional metagenomics study of antimicrobial resistance. *Frontiers in Microbiology* 6(March): 1-6. Available at: http://journal.frontiersin.org/article/10.3389/fmicb.2015.00172. Accessed February 21, 2017.

Hodgkins S.B., M.M. Tfaily, C.K. McCalley, T.A. Logan, P.M. Crill, S.R. Saleska, et al. 2014. Changes in peat chemistry associated with permafrost thaw increase greenhouse gas production. *Proceedings of the National Academy of Sciences USA* 111: 5819-24.

Hu, Y., X. Yang, J. Qin, N. Lu, G. Cheng, N. Wu, Y. Pan, J. Li, L. Zhu, X. Wang, Z. Meng, F. Zhao, D. Liu, J. Ma, N. Qin, C. Xiang, Y. Xiao, L. Li, H. Yang, J. Wang, R. Yang, G.F. Gao, J. Wang, and B. Zhu. 2013. Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nature Communications* 4: 2151. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23877117. Accessed February 21, 2017.

International Panel on Climate Change. 2014: *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. IPCC, Geneva, Switzerland.

Johnson, M., I. Zaretskaya, Y. Raytselis, Y. Merezhuk, S. McGinnis, and T.L. Madden. 2008. NCBI BLAST: A better web interface. *Nucleic Acids Research* 36(Web Server issue): W5-W9. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2447716/. Accessed February 21, 2017.

Johnson, T.A., R.D. Stedtfeld, Q. Wang, J.R. Cole, S.A. Hashsham, T. Looft, Y.-G. Zhu, and J.M. Tiedje. 2016. Clusters of antibiotic resistance genes enriched together stay together in swine agriculture. *mBio* 7(2): e02214-15. Available at: http://mbio.asm.org/content/7/2/e02214-15.full. Accessed February 21, 2017.

Katoh, K. and D.M. Standley. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution* 30(4): 772-780.

Li, B., Y. Yang, L. Ma, F. Ju, F. Guo, J.M. Tiedje, and T. Zhang. 2015. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *The ISME Journal (9)*: 2490-2502. Available at: http://dx.doi.org/10.1038/ismej.2015.59. Accessed February 21, 2017.

Martínez, J.L. 2008. Antibiotics and antibiotic resistance genes in natural environments. *Science* 321(5887): 365-367.

Martinez, J.L. 2009. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proceedings. Biological Sciences / The Royal Society* 276(1667): 2521-2530.

Martínez, J.L. 2012. Natural antibiotic resistance and contamination by antibiotic resistance determinants: The two ages in the evolution of resistance to antimicrobials. *Frontiers in Microbiology* 3(JAN): 2010-2012.

Martínez, J.L., M.B. Sánchez, L. Martínez-Solano, A. Hernandez, L. Garmendia, A. Fajardo, and C. Alvarez-Ortega. 2009. Functional role of bacterial multidrug efflux pumps in microbial natural ecosystems. *FEMS Microbiology Reviews* 33(2): 430-449.

McArthur, A.G., N. Waglechner, F. Nizam, A. Yan, M.A. Azad, A.J. Baylay, K. Bhullar, M.J. Canova, G. De Pascale, L. Ejim, L. Kalan, A.M. King, K. Koteva, M. Morar, M.R. Mulvey, J.S. O'Brien, A.C. Pawlowski, L.J.V. Piddock, P. Spanogiannopoulos, A.D. Sutherland, I. Tang, P.L. Taylor, M. Thaker, W. Wang, M. Yan, T. Yu, and G.D. Wright. 2013. The comprehensive antibiotic resistance database. *Antimicrobial Agents and Chemotherapy* 57(7): 3348-3357.

McCalley, C.K., B.J. Woodcroft, S.B. Hodgkins, R.A. Wehr, E.H. Kim, R. Mondav, P.M. Crill, J.P. Chanton, V.I. Rich, G.W. Tyson, and S.R. Saleska. 2014. Methane dynamics regulated by microbial community response to permafrost thaw. *Nature* 514(7523): 478-481. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25341787. Accessed February 21, 2017.

McLain, J. and C. Williams. 2014. Sustainability of water reclamation: Long-term recharge with reclaimed wastewater does not enhance antibiotic resistance in sediment bacteria. *Sustainability* 6(3): 1313-1327. Available at: http://www.scopus.com/inward/record.url?eid=2-s2.0-84897046113&partnerID=tZOtx3y1. Accessed February 21, 2017.

Mondav, R., B.J. Woodcroft, E.H. Kim, C.K. McCalley, S.B. Hodgkins, P.M. Crill, J. Chanton, G.B. Hurst, N.C. VerBerkmoes, S.R. Saleska, P. Hugenholtz, V.I. Rich, and G.W. Tyson. 2014. Discovery of a novel methanogen prevalent in thawing permafrost. *Nature Communications* 5: 3212. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24526077. Accessed February 22, 2017.

Nesme, J., S. Cécillon, T.O. Delmont, J.-M. Monier, T.M. Vogel, and P. Simonet. 2014. Large-scale metagenomic-based study of antibiotic resistance in the environment. *Current Biology* 24(10): 1096-1100.

Pehrsson E.C., P. Tsukayama, S. Patel, M. Mejía-Bautista, G. Sosa-Soto, and K.M. Navarrete. 2016. Interconnected microbiomes and resistomes in low-income human habitats. *Nature* 533: 212-216.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European molecular biology open software suite. *Trends in Genetics* 16(6): 276-277.

Roose-Amsaleg, C. and A.M. Laverman. 2015. Do antibiotics have environmental side-effects? Impact of synthetic antibiotics on biogeochemical processes. *Environmental Science and Pollution Research*. Available at: http://link.springer.com/10.1007/s11356-015-4943-3. Accessed February 22, 2017.

Seemann, T. 2014. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 30(14): 2068-2069.

Sievers F., A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, and W. Li. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 7. e-pub ahead of print DOI: 10.1038/msb.2011.75.

Tanabe, M. and M. Kanehisa. 2012. Using the KEGG database resource. *Current Protocols in Bioinformatics* (SUPPL.38): 1-54.

Walsh, F., S.G.B. Amyes, and B. Duffy. 2013. Challenging the concept of bacteria subsisting on antibiotics. *International Journal of Antimicrobial Agents* 41(6): 558-563. Available at: http://dx.doi.org/10.1016/j.ijantimicag.2013.01.021. Accessed February 22, 2017.

Webber, M.A. and L.J.V. Piddock. 2003. The importance of efflux pumps in bacterial antibiotic resistance. *Journal of Antimicrobial Chemotherapy* 51(1): 9-11.

World Health Organization. 2014. *Antimicrobial Resistance. Global Report on Surveillance 2014*. World Health Organization, Geneva, Switzerland.

Woodcroft B., C. Singleton, P. Evans, R. Wilson, S. Hodgkins, J. Chanton, S. Saleska, V.I. Rich, and G.W. Tyson. Microbial community response to permafrost thaw across a natural gradient via shallow metagenome profiling. In preparation.

Wright, G.D. 2007. The antibiotic resistome: The nexus of chemical and genetic diversity. *Nature Reviews. Microbiology* 5(3): 175-186.