FEMS Microbiology Ecology, 94, 2018, fiy079

doi: 10.1093/femsec/fiy079
Advance Access Publication Date: 26 April 2018
Research Article

RESEARCH ARTICLE

MetaCompare: a computational pipeline for prioritizing environmental resistome risk

Min Oh[1], Amy Pruden[2], Chaoqi Chen[3], Lenwood S. Heath[1], Kang Xia[3] and Liqing Zhang[1,*]

[1]Department of Computer Science, Virginia Tech, Blacksburg, VA 24061, USA, [2]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, VA 24061, USA and [3]Department of Crop and Soil Environmental Sciences, Virginia Tech, Blacksburg, VA 24060, USA

*Corresponding author: Department of Computer Science, Virginia Tech, 2160K Torgersen Hall, Blacksburg, Virginia 24060, USA. Tel: 540-231-9413, E-mail: lqzhang@cs.vt.edu

One sentence summary: A computational pipeline ranking resistome risk for given environmental samples based on shotgun metagenomic sequencing data was developed.

Editor: Edward Topp

†Min Oh, http://orcid.org/0000-0002-0938-9046

ABSTRACT

The spread of antibiotic resistance is a growing public health concern. While numerous studies have highlighted the importance of environmental sources and pathways of the spread of antibiotic resistance, a systematic means of comparing and prioritizing risks represented by various environmental compartments is lacking. Here, we introduce MetaCompare, a publicly available tool for ranking 'resistome risk', which we define as the potential for antibiotic resistance genes (ARGs) to be associated with mobile genetic elements (MGEs) and mobilize to pathogens based on metagenomic data. A computational pipeline was developed in which each ARG is evaluated based on relative abundance, mobility, and presence within a pathogen. This is determined through the assembly of shotgun sequencing data and analysis of contigs containing ARGs to determine if they contain sequence similarity to MGEs or human pathogens. Based on the assembled metagenomes, samples are projected into a 3-dimensionalhazard space and assigned resistome risk scores. To validate, we tested previously published metagenomic data derived from distinct aquatic environments. Based on unsupervised machine learning, the test samples clustered in the hazard space in a manner consistent with their origin. The derived scores produced a well-resolved ascending resistome risk ranking of: wastewater treatment plant effluent, dairy lagoon, and hospital sewage.

Keywords: antibiotics resistance gene; resistome; resistome risk; metagenomics; environmental samples

INTRODUCTION

The acquisition of antibiotic resistance by bacterial pathogens is of global concern, as it undermines available therapeutics for treating and preventing serious infections (World Health Organization 2014). Thus, there is a need to develop means to track and control the spread of antibiotic resistance, with growing interest in understanding the role of environmental sources and pathways by which antibiotic resistance may spread. Recent advances in next-generation DNA sequencing enable capture of the full metagenomic complement of functional genes in complex environments, including antibiotic resistance genes (ARGs)

Received: 25 April 2018; **Accepted:** 25 April 2018

and mobile genetic elements (MGEs) (Wooley, Godzik and Friedberg 2010). ARGs encode various cellular functions that enable bacteria to survive in the presence of antibiotics, while MGEs enable ARGs to be transferred among different bacterial strains. Importantly, ARGs relevant to clinically important bacteria have largely been found to have evolved in and originated from natural environments (Poirel, Kampfer and Nordmann 2002; Poirel *et al*. 2005; Teeling and Glockner 2012). It is now widely known that the phenomenon of antibiotic resistance itself is as ancient as the microbes themselves that produce antibiotics (D'Costa *et al*. 2011). There is no environment to date; including frozen tundra (Allen *et al*. 2010), isolated caves (Bhullar *et al*. 2012) or isolated human populations (Clemente *et al*. 2015) that has been found to be free of ARGs (Martínez 2008). This raises the question of how to judge the extent to which an ARG detected in the environment poses a risk to human health?

Human health risk assessment provides a quantitative framework for estimating the likelihood of illness given a specific exposure route and dose (U.S. Environmental Protection Agency (U.S. EPA) 2012; Ashbolt *et al*. 2013). For example, quantitative microbial risk assessment has been applied to estimate the concentration of *Legionella pneumophila* that is necessary in bulk water during showering to cause the severe pneumonia characteristic of Legionnaires' Disease (Schoen and Ashbolt 2011). Developing such a risk assessment framework tailored antibiotic resistance, however, is much more complex. At the heart of the challenge is the question of exactly which aspects of the hazard are to be modeled. Shotgun metagenomic DNA sequencing is becoming more accessible and widely applied for broadly profiling the full complement of ARGs, or the 'resistome', of a given environmental sample (Bengtsson-Palme 2017; Wright 2007). Reads containing DNA sequences with high similarity to known ARGs can be identified via existing publicly available databases, such as the Comprehensive Antibiotic Resistance Database (CARD) (Jia *et al*. 2017). Such an approach is advantageous, as it is more holistic, circumventing biases associated with culture-based techniques and can be applied to compare various attributes of the resistome that might represent a hazard, for example, the reduction in total ARGs following a given mitigation event. This comparative hazard identification and exposure assessment approach assumes that the actual reduction in risk is proportional to the reduction in total ARGs. However, in reality, each individual ARG varies in its actual impact on risk based on several aspects, including the type of antibiotic resistance it confers, whether it is present in an actual pathogen, and whether it has the potential to be transferred to an actual pathogen (Martinez, Coque and Baquero 2015). Furthermore, aggregate evaluation of putative impacts of individual ARGs provides a more realistic and comprehensive characterization of potential hazards for subsequent risk assessment of a given environment. In this regard, we define 'resistome risk' as the cumulative potential for ARGs to occur on MGEs and in human pathogens, as inferred from metagenomic data. In this manner, evaluating resistome risk provides a means to prioritize mitigation efforts based on relative ranking of specific environments or critical control points in terms of their corresponding likelihood of mobilizing ARGs to pathogens and incurring a negative impact on human health. Unfortunately, a quantitative model that enables comparison of resistome risk is lacking and urgently needed to advance monitoring and mitigation strategies and fight the spread of resistance. To tap into the full potential of shotgun metagenomic data sets and derive information ultimately important to human health risk assessment, it is ideal to be able to identify not only ARGs, but also gene fragments corresponding to MGEs and human pathogens. MGEs, such as transposons, integrons, plasmids and prophages, play a key role in mediating horizontal transfer of resistance determinants from their original bacterial hosts to human pathogens (Forsberg *et al*. 2012). Through bioinformatics techniques, such as local sequence alignment tools and sequence assembly (Boratyn *et al*. 2013; Breitwieser, Lu and Salzberg 2017), it is possible to estimate which ARGs occur on MGEs and whether they are present in pathogens. Such information can aid in judging the potential for resistance to spread, as well as in ranking potential human health hazards represented by individual ARGs and resistomes.

Recently, a conceptual framework for prioritizing the risk of dissemination associated with individual ARGs by a ranking system was proposed (Martinez, Coque and Baquero 2015). This framework suggests criteria for ranking ARGs depending on the likelihood that they present a public health concern. While this is a useful framework, there are challenges to actual implementation. Specific challenges and limitations include: (i) focus on gene-level assessment, which ignores the full power of a shotgun metagenomic data set; (ii) requirement of functional assays to confirm annotations and gene assignments, which is unrealistic from a practical standpoint and (iii) necessity to discriminate between known and unknown mechanisms of resistance based on functional evaluation, which is difficult to accurately differentiate through a computational approach. Although we judge that some of these barriers cannot presently be surmounted, here we develop and demonstrate a system incorporating key criteria for deriving and prioritizing resistome risks from metagenomic data, considering factors such as abundance, mobility and potential association of ARGs with pathogens. Specifically, we introduce MetaCompare, a computational pipeline for prioritizing resistome risk by estimating the potential for ARGs to be disseminated into human pathogens in a given environment based on metagenomic sequencing data. This pipeline is publicly available at https://github.com/minoh0201/MetaCompare.

METHODS

MetaCompare overview

For a given metagenomic data set derived from a sample of interest, MetaCompare estimates resistome risk by identifying and quantifying assembled sequence fragments that are associated with abundance, mobility and presence of ARGs within a pathogen (Fig. 1). To estimate the co-occurrence of key resistome risk 'components', specifically ARGs, MGEs and gene fragments corresponding to human pathogen, the pipeline includes the assembly of metagenomic sequencing reads. After quality control of raw reads with Trimmomatic, high-quality reads are assembled using IDBA-UD with default parameters (Peng *et al*. 2012; Bolger, Lohse and Usadel 2014). Assembled contigs meeting specific conditions are classified into three categories: (i) contigs in which ARG-like sequences are found, (ii) contigs where ARG-like sequences and MGE-like sequences are detected concurrently and (iii) contigs in which ARG-like, MGE-like and human pathogen-like sequences are observed together. The number of assembled contigs corresponding to each category is normalized by the total number of assembled contigs. Based on the normalized values, each sample is projected into 3-dimensional space, which is termed 'hazard space', representing a geometric position of the given sample in that space. Lastly, how closely related each sample is to the theoretical sample

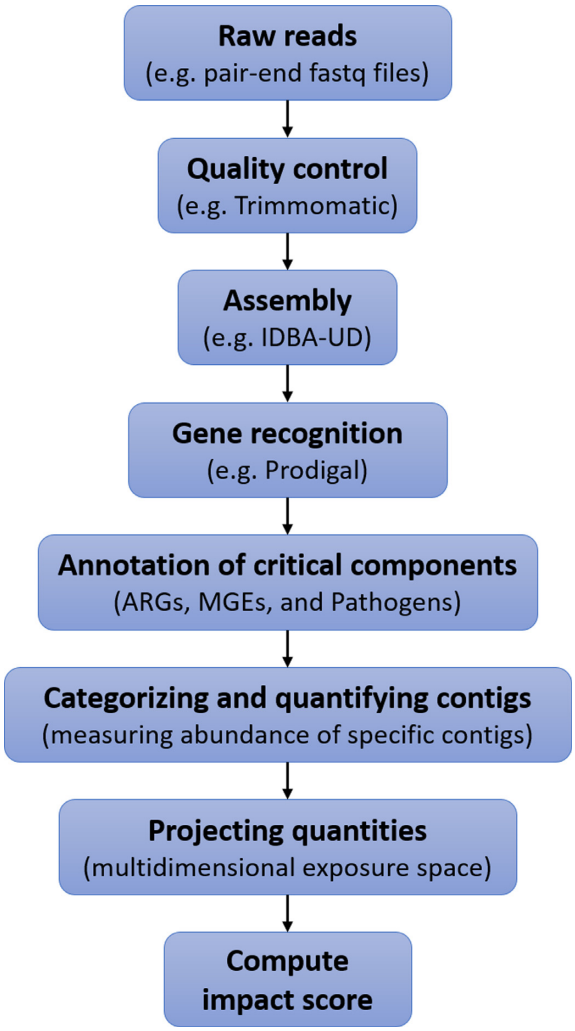

Figure 1. Overview of MetaCompare pipeline for ranking resistome risk based on shotgun metagenomic sequencing data obtained from a given environmental sample.

representing the highest potential resistome risk is calculated by measuring the distance between the point of the given sample and the theoretical boundary in a given hazard space. Lastly, the resistome risk score is derived from the distance measurement, enabling comparison of resistome risk across all samples of interest.

Detecting and quantifying critical contigs

Martinez, Coque and Baquero (2015) postulate a conceptual framework in which an ARG residing on an MGE that is hosted by a human pathogen represents the highest 'risk level,' as defined by the authors. To identify such cases from metagenomic reads, it is necessary to search for clues for coexistence of genetic elements indicative of ARGs, MGEs and pathogens, which herein we define as 'hazards'. MetaCompare utilizes assembled contigs as the basic search unit, because the length of a typical Illumina sequencing read does not sufficiently support detection of co-occurrence of multiple gene sequences. For each contig, protein-coding genes were predicted using Prodigal, a prokaryotic gene recognition tool, to support accurate annotation of functional components (Hyatt *et al.* 2010). The predicted genes were analyzed with BLASTX (E-value < 1E-10, Identity >60%, and Minimum alignment length > 25 amino acids) against the CARD database, which is presently considered to be the most comprehensive repository for ARGs (Jia *et al.* 2017). To identify MGE-like sequences, the ACLAME database, which is comprised of known bacteriophage genomes, plasmids and transposons, was searched using BLASTN (E-value < 1E-10, Identity >60%) (Leplae, Lima-Mendez and Toussaint 2010). To ensure high-quality annotation, we filtered out mapping results yielding less than 90% coverage of the MGE reference. To identify human pathogen-like sequences, BLASTN searches were conducted against the human bacterial pathogen genomes that populate the PATRIC database (E-value <1e−10, Identity >60%, Minimum alignment length >150 bps) (Wattam *et al.* 2017). Note that these parameters can be adjusted depending on depth of coverage and quality of the metagenomic sequencing samples.

We categorized assembled contigs into categories according to their identified critical components (Fig. 2). Let ARG, MGE and $PATH$ be those sets of contigs in which ARG-like sequences, MGE-like sequences, and pathogen-like sequences are detected, respectively. For a given contig c and the entire set of contigs C, the indicator function is as follows.

$$I_{ARG}(c) = \begin{cases} 1, & \text{if } c \in ARG \\ 0, & \text{if } c \in C - ARG \end{cases}$$

Given the total number of assembled contigs $N_{contigs}$, we estimated the relative quantity of contigs where ARG-like sequences were detected as follows.

$$Q(ARG) = \frac{1}{N_{contigs}} \sum_{c} I_{ARG}(c)$$

Likewise, contigs annotated with ARGs and MGE-like sequences were quantified as follows.

$$Q(ARG, MGE) = \frac{1}{N_{contigs}} \sum_{c} I_{ARG \cap MGE}(c)$$

Contigs annotated with pathogen genomes as well as ARGs and MGEs were quantified in a similar manner.

$$Q(ARG, MGE, PATH) = \frac{1}{N_{contigs}} \sum_{c} I_{ARG \cap MGE \cap PATH}(c)$$

Projecting samples into a hazard space and computing resistome risk scores

We define the hazard space to be a 3-dimensional space, where each dimension indicates a resistome risk factor that can be quantified as a real number. Here, three quantities were computed for each metagenomic data set: $Q(ARG)$, $Q(ARG, MGE)$ and $Q(ARG, MGE, PATH)$. By mapping these values to the coordinates, we projected metagenomic data sets into a 3-dimensional hazard space (Fig. 3). Metagenomic data derived from each sample is represented by a single point in hazard space. Euclidean distance between two points in hazard space yields the proximity of each sample pair. To compare resistome risk among multiple samples, we define a theoretical point h that represents the highest resistome risk in the hazard space. The

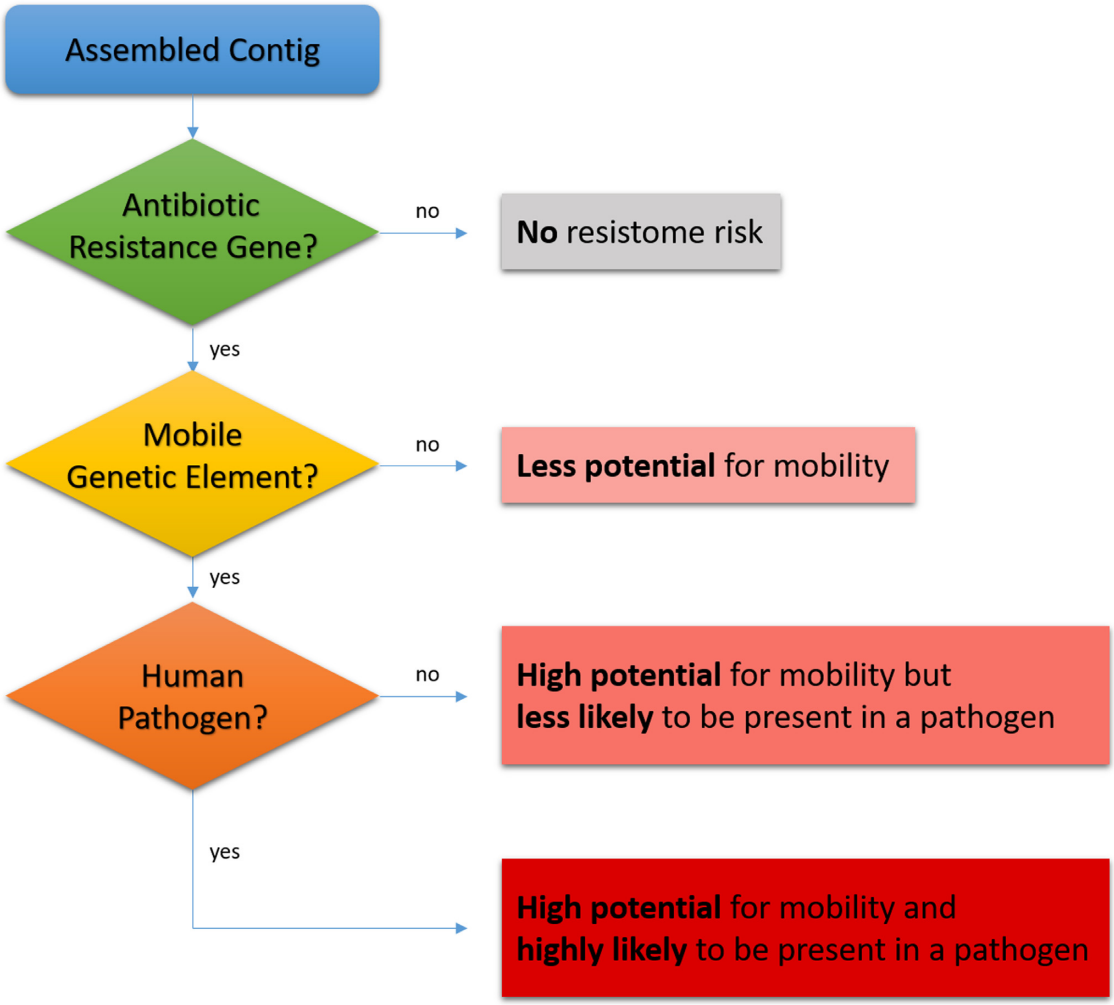

Figure 2. Proposed categorization of an assembled contig according to annotation of its individual critical components.

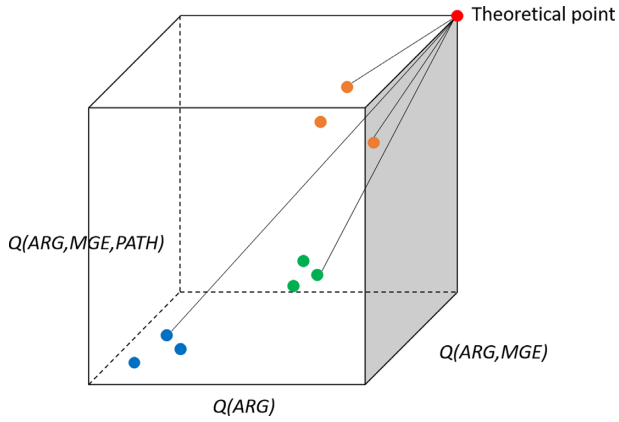

Figure 3. Illustration of the three-dimensional hazard space.

theoretical maximum values are assigned for h as follows:

$$Q_h(ARG) = 0.01$$

$$Q_h(ARG, MGE) = 0.01$$

$$Q_h(ARG, MGE, PATH) = 0.01$$

We conducted a simulation utilizing the prevalence data in CARD database to obtain an empirical maximum boundary for point h (Jia *et al.* 2017). The prevalence data consist of 10318 pathogen genomes, each of which is complete chromosome and complete plasmid sequences or whole genome sequencing assemblies, annotated with ARGs using Resistance Gene Identifier (Jia *et al.* 2017). Those genomes are one of 27 resistant pathogens, including World Health Organization's antibiotic-resistant priority pathogens, multidrug-resistant ESKAPE pathogens (*Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* species). For each genome, 23 contigs were arbitrarily cut from the genomes by randomly picking up the length of the contigs from the length distribution of assembled contigs derived from real data used in this study. In total, 237314 contigs were gathered, which is approximately equivalent to the average number of assembled contigs per sample in the real data. The fraction of the contigs annotated with ARGs out of the total number of contigs was calculated. We repeated this process 1000 times and secured a distribution of the fractions, resulting in obtaining the value 0.0105 that cuts off a right tail to be 1% of

whole area in the fraction distribution. We used the rounded value 0.01 as a maximum boundary for $Q_h(ARG)$ and the following $Q_h(ARG, MGE)$ and $Q_h(ARG, MGE, PATH)$ as $Q_h(ARG) \geq Q_h(ARG, MGE) \geq Q_h(ARG, MGE, PATH)$.

By calculating the distance from the point h to each sample, we can obtain a distance from the point h to each sample for comparison. Suppose s denotes a point in hazard space corresponding to a given sample and $dist(s, h)$ indicates the Euclidean distance between s and h. For convenience, the resistome risk score is computed by inverting the distance value and incorporating logarithmic scaling as follows.

$$\text{score}\,(s) = \frac{1}{(2 + \log_{10} dist\,(s, h))^2}$$

Clustering methods and evaluation

We conducted unsupervised machine learning analysis to assess how accurately the proposed hazard space classified the different metagenomic data sets. Three clustering algorithms, k-means, fuzzy c-means and k-medoid, were applied to the samples within the proposed hazard space and the alternative hazard space derived from basic abundances of reads annotated as ARGs. Information about sample type was excluded in the learning step and used to evaluate how consistent the clustering results were with the origin of the samples.

Given disjoint clusters of data points, silhouette analysis was conducted to validate consistency within clusters (Rousseeuw 1987). For each data point i, suppose $a(i)$ is the average dissimilarity of i with other points within the same cluster and $b(i)$ is the lowest average dissimilarity of i to all points in any other cluster, of which i is not a member. The silhouette of the data point i is as follows:

$$S\,(i) = \frac{b\,(i) - a\,(i)}{\max\,\{a\,(i), b\,(i)\}}$$

The agreement between the clustering result and the actual sample type, i.e. 'dairy lagoon', 'hospital sewage', and 'WWTP effluent', is evaluated by accuracy metrics. Suppose we have N samples $s_1, \ldots, s_N$ and M classes $c_1, \ldots, c_M$. Each sample has an actual class c_i and belongs to an answer set C_i such that all elements in the set have the same class c_i. Based on similarity between samples, clustering method generates M clusters $K_1, \ldots, K_M$ where K_j contains samples supposed to have the same class. For answer sets C and clusters K, the purity is computed by

$$\text{Purity}\;(C, K) = \sum_j \frac{|K_j|}{N} \max_i \frac{|K_j \cap C_i|}{|K_j|}$$

where $|K_j|$ denotes cardinality of K_j. F-measure is computed by

$$F\;(C, K) = \sum_i \frac{|C_i|}{N} \max_j \left(\frac{2\text{Pr}\,(C_i, K_j) \cdot \text{Re}\,(C_i, K_j)}{\text{Pr}\,(C_i, K_j) + \text{Re}\,(C_i, K_j)} \right)$$

where precision and recall are calculated respectively as follows:

$$\text{Pr}\;(C_i, K_j) = \frac{|K_j \cap C_i|}{|K_j|}$$

$$\text{Re}\;(C_i, K_j) = \frac{|K_j \cap C_i|}{|C_i|}$$

We utilized various R packages for cluster analysis. For clustering, the 'stats', 'e1071' and 'cluster' packages were used. For clustering evaluation and visualization, the 'FlowSOM', 'factoextra', 'ggplot2' and 'plotly' packages were utilized.

Significance test for the resistome risk score

To assess the statistical significance of the resistome risk score distribution, permutation tests were performed to determine if the calculated distances from the theoretical point were significantly different from the null hypothesis of random distribution of resistome risk scores. For a given sample, random samples having the same number of contigs and the same number of annotations for critical components, but having randomly varied co-occurrences of critical components, were generated and a P-value for a distance of the observed sample was obtained from the sampling distribution.

Validation

Publicly available environmentally derived shotgun metagenomic data sets were selected for validation and downloaded from the European Nucleotide Archive (Rowe *et al.* 2016; Rowe *et al.* 2017). Three data sets were selected, representing three distinct environments expected to represent distinct resistome risks within the same geographical area (Cambridge, UK): Lagoon water from the University of Cambridge dairy farm (dairy lagoon), combined wastewater effluent from the main wards of the Cambridge University Hospital (hospital sewage), and effluent from a municipal wastewater treatment plant discharging to the River Cam (WWTP effluent). Cambridge University Hospital has a bed capacity of 1000 and reported 1070.48 Defined Daily Doses per 1000 bed days in October 2014. The WWTP effluent is from a conventional secondary wastewater treatment plant serving mainly domestic wastewater generated by a population of ~200000. According to personal communication with the authors of the study from which the data were derived, the dairy lagoon was fed by the runoff from scraped and flushed manure, which was sampled after approximately 2 months of operation and immediately prior to withdrawal for application to agricultural land. Specifically, the dairy lagoon samples were collected from a sampling valve on the feed pipe of the lagoon.

Comparison with common abundance measurement

We constructed an alternative hazard space using common abundance measurements for critical components. Abundances of ARGs and MGEs were calculated using MetaStorm, which computes the normalized ARG/MGE abundance as the copy of a functional gene per copy of 16S rRNA genes as described in Li *et al.* (2015) (Arango-Argoty *et al.* 2016). For putative pathogens, relative abundances of individual species were derived from MetaPhlan2 and cross-referenced to bacterial pathogens housed in the PATRIC database (Truong *et al.* 2015). Silhouette analysis was applied to evaluate tightness of clusters in hazard space and the alternative hazard space.

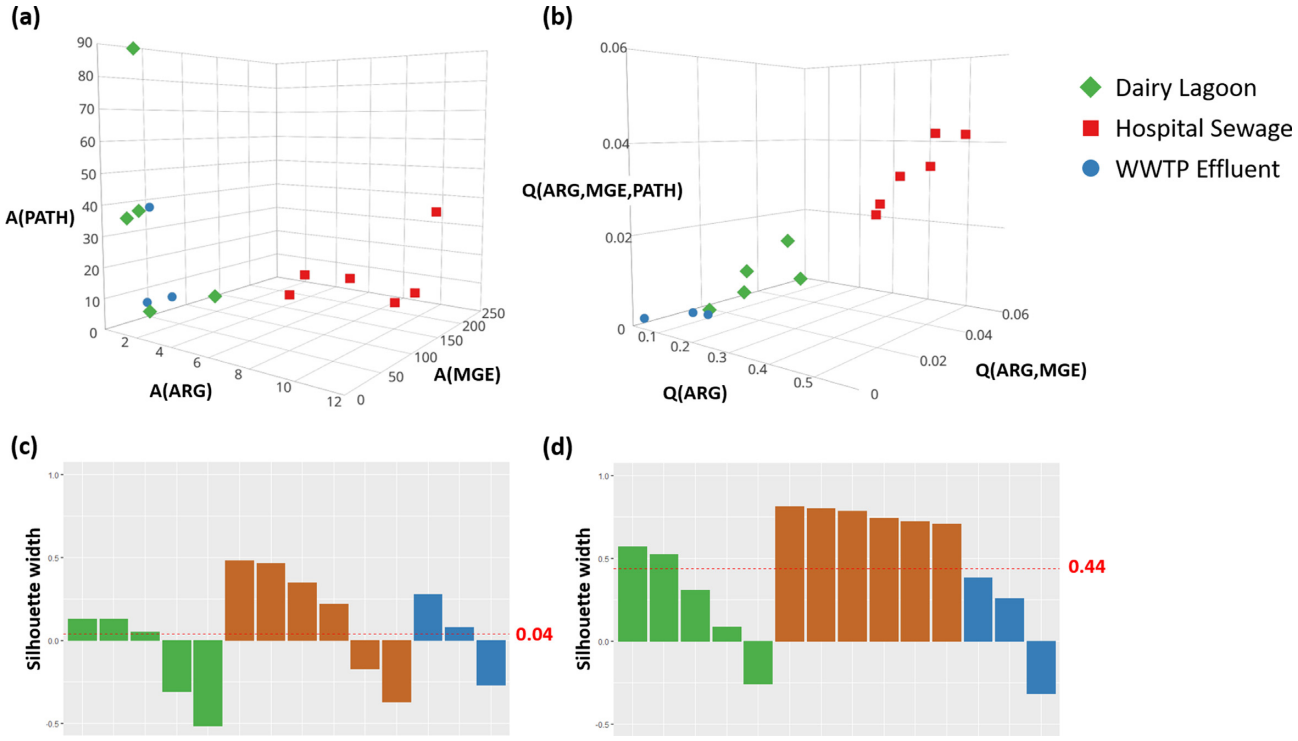

Figure 4. Projection of the test metagenomic data sets into hazard space followed by silhouette analysis to evaluate cohesiveness of members assigned to each cluster. (**a**) Hazard space derived from abundances of reads annotated directly as ARGs, MGEs or pathogens. (**b**) Proposed hazard space derived from estimation of co-occurrence of critical components on assembled contigs. (**c**) Clustering coefficient (silhouette width) for each sample presented in (a) displayed in hazard space. The red dotted line with the red value indicates the average value of all silhouette widths. (**d**) Result of clustering coefficient analysis for the proposed hazard space based on estimation of co-occurrence of critical components on assembled contigs. WWTP- waste water treatment plant.

RESULTS

Characterization and validation of the hazard space

Here, we demonstrated that the proposed hazard space is capable of clearly distinguishing the different environmental sample types (Fig. 4). As expected, samples originating from the same location clustered together with a similar resistome risk ranking, with some minor differences amongst individual samples. To explore how tightly clustered samples of the same type were plotted in risk space, silhouette analysis was conducted (Fig. 4d) (Rousseeuw 1987). As shown in Fig. 4c and d, the average silhouette width of our hazard space surpassed that of the hazard space derived from commonly used abundance measurements, indicating that the proposed hazard space approach yielded higher sensitivity towards characterizing resistome risk for the tested data sets.

The clusters generated from the hazard space demonstrated improved performance in correctly categorizing sample types than the clusters derived from the commonly used normalized read abundance approach (Fig. 5). This suggests that the proposed hazard space can provide higher resolution for distinguishing samples from a range of environments and representing varying levels of resistome risk.

Resistome risk scores obtained from experimental metagenomic data sets

Based on typical concentrations of antibiotics in the types of waters corresponding to the test data sets (up to 45000 ng/L for hospital sewage, medium concentrations ranging from 700 to 6600 ng/L for dairy effluent, and about 600 ng/L for waste water treatment plant effluent) (Brown *et al.* 2006), we expected a high resistome risk for hospital sewage samples, medium resistome risk for dairy lagoon samples and low resistome risk for WWTP effluent (Table 1). We computed resistome risk scores for the selected samples based on distances from the theoretical point representing the highest possible theoretical score in hazard space. All resistome risk scores indicated significant differences from a random distribution of the critical components (Permutation test P-value < 0.0001). We subsequently ranked all samples in descending order of resistome risk score in Table 2. These results were consistent with our expectation a high resistome risk for hospital sewage samples, medium resistome risk for dairy lagoon samples, and low resistome risk for WWTP effluent. One dairy lagoon sample was an exception, ranking lower than one of the WWTP effluent samples.

DISCUSSION

To the authors' knowledge, MetaCompare is the first computational pipeline for ranking resistome risks from metagenomic data sets derived from environmental samples. Here, we build on conceptual framework proposed by Martinez, Coque and Baquero (2015), extending the concept to the resistome level, with some simplifications in the ranking scheme. MetaCompare provides a publically available computational pipeline for ranking resistome risk of a given metagenomic data set derived from a sample of interest based on the development of a metric for co-occurrence of three key critical components annotated on individual assembled contigs, specifically: ARGs, MGEs and human pathogens.

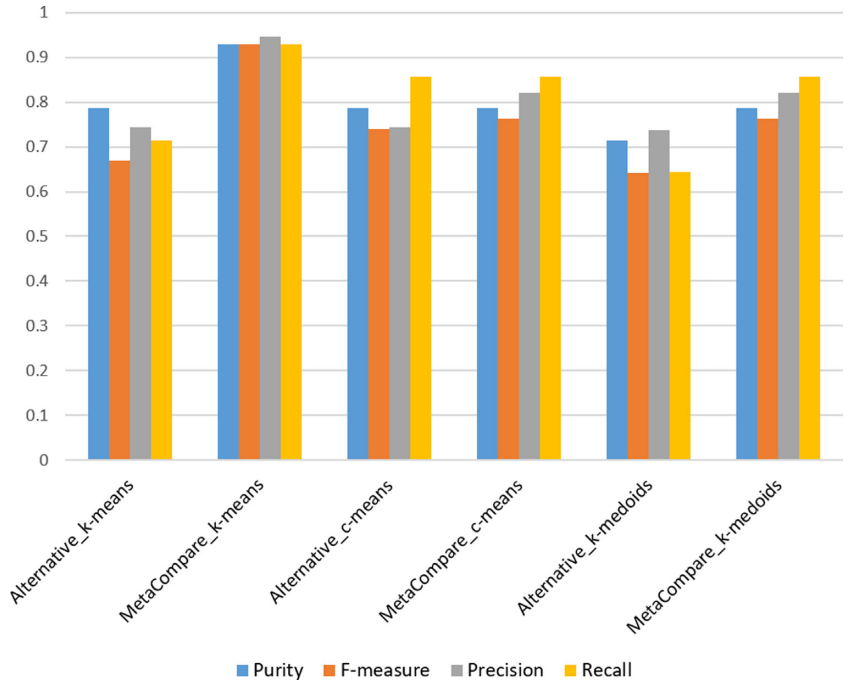

Figure 5. Evaluation of clustering quality. Based on different means of measuring clustering quality: Purity, F-measure, Precision and Recall, as applied to k-means, c-means and k-medoids obtained using 'Alternative' (i.e. sequence read normalized) versus MetaCompare (i.e. co-occurrence of critical components on assembled contigs) approaches to project the samples into 3-D hazard space.

Table 1. Information about the samples from which metagenomics data were obtained.

Water type	Latitude	Longitude	# of samples	Antibiotics detected and average concentrations (μg/L)	Reference
Hospital Sewage	52.174343	0.139346	6	Flucloxacillin (4.9), Vancomycin (671.12), Azithromycin (11.48), Clarithromycin (18.56), Ciprofloxacin (33.76), Moxifloxacin (0.4), Rifampicin (0.26), Sulfamethoxazole (333.08)	Rowe *et al.* (2017)
Dairy Lagoon	52.22259	0.02603	5	Sulfadiazine (0.88)	Rowe *et al.* (2017)
WWTP Effluent	52.234469	0.154614	3	N/A	Rowe *et al.* (2016)
		Total	14		

Table 2. Ranking of risk scores obtained using MetaCompare.

ENA accession[a]	Sample type	Risk score
ERS1019924	Hospital Sewage	43.00
ERS1019927	Hospital Sewage	39.47
ERS1019928	Hospital Sewage	39.24
ERS1019923	Hospital Sewage	36.23
ERS1019925	Hospital Sewage	34.62
ERS1019926	Hospital Sewage	34.47
ERS1019959	Dairy Lagoon	29.02
ERS1019958	Dairy Lagoon	26.84
ERS1020022	Dairy Lagoon	24.82
ERS1019955	Dairy Lagoon	24.20
ERS1020020	WWTP Effluent	22.77
ERS1019956	Dairy Lagoon	22.71
ERS1019947	WWTP Effluent	21.59
ERS1019948	WWTP Effluent	18.42

[a]European Nucleotide Archive sample accession (accessible at https://www.ebi.ac.uk/ena).

MetaCompare enables visual and quantitative comparison of collective resistome risk by projecting the samples into a hazard space and calculating resistome risk scores. We validated our approach using metagenomic data obtained from three distinct aquatic environments that were expected to represent distinct categories of resistome risk. Strikingly, MetaCompare generated a clear and highly resolved clusters according to sample type along with resistome risk rankings in alignment with expectations based on typical antibiotic occurrence and level of water treatment. This kind of robust characterization of resistome risk could be a useful tool for identifying potential hot spots for dissemination of antibiotic resistance and prioritizing mitigation interventions. In particular, the ability to rank risk of dissemination of antimicrobial resistance and potential to spread to human pathogens for various environmental compartments would be highly valuable towards guiding investment in mitigation strategies.

Although this initial version of MetaCompare is promising, it still has several limitations. Many of these stem from limitations of confidence in assembly and in the curation of available databases. First, assembly of metagenomic data may contain some level of error or chimeras (Olson *et al.* 2017). This problem stems from intragenomic repeats of DNA segments within the same genome as well as intergenomic repeats shared among distinct genomes. The Iterative De Bruijn graph *de novo* Assembler for short reads with highly Uneven Depth (IDBA-UD) generates assembled contigs based on de Bruijn graphs and performs analysis of coverages between paths in the graph to improve assembly and avoid chimeras. However, it is not possible to completely eliminate chimeric contigs, which confound annotation of critical components. Secondly, annotation of critical components depends on the principle that genes displaying high sequence similarity generally exert similar functions (Teeling and Glockner 2012). Even when assembled contigs are reliably aligned to reference genes, the contigs may only contain a portion of the reference sequence and thus there is the potential that in reality the portion of the gene identified on the contig originated from a non-functional truncated gene. Thirdly, identification of critical components from assembled contigs relies heavily on the reference databases for ARGs, MGEs and pathogens. Therefore, the MetaCompare annotation profiles depend on the curation quality of the databases. Annotation errors and biases in the databases and/or incompleteness of the curated entries could all adversely affect resistome risk calculations. A closely related issue is that there might be many different databases for the same critical components, such as CARD (Jia *et al.* 2017), DeepARG (Arango-Argoty *et al.* 2017), ARDB (Liu and Pop 2009) and SARG (Yang *et al.* 2016) that can be used to identify ARGs. Applying different combinations of databases for annotating ARGs, MGEs and human pathogens might result in different resistome risk scores. Nonetheless, it is expected that the general trend of resistome risk scores among a given set of samples will be consistent, but this will need to be tested in the future with a larger dataset. Lastly, sequencing depth is a key underlying factor that may limit the ability of MetaCompare to accurately assess the true resistome risks. For complex environments, it is expected that comprehensive identification of critical components and accurate estimation by MetaCompare will benefit from deeper sequencing. With these caveats in mind, we emphasize that MetaCompare is the first computational pipeline to attempt to quantify and prioritize resistome risks among environmental samples. While it is not meant to be comprehensive or to take into account all possible risk factors, it aims to build a framework taking an important step towards beginning to prioritize the resistome risk represented by various environmental compartments. With continued reduction in sequencing cost and availability of new sequencing technologies, deeper sequencing with improved length and accuracy will become more widely available, which will improve the accuracy of the identification of individual critical components and the overall resistome risk rankings determined using the MetaCompare framework.

There are several aspects of MetaCompare that could be further improved in the future. As discussed above, some aspects of risk are either difficult to determine computationally or lacking experimental validation. First, the number of contigs containing critical components are normalized to the total number of contigs. Future improvement can include length normalization to reduce bias in annotation of long genes. Second, the reference database used to identify ARGs was CARD, which may contain: (i) core genomes of some species (e.g. *E. coli*, *P. aeruginosa*, *E. faecium* and *K. oxytoca*) that are associated with intrinsic resistance phenotypes; (ii) mutated transporter genes that hinder antibiotic intak or (iii) mutated target genes of antibiotics that make antibiotics nonfunctioning. These genes could be excluded from resistome risk consideration because they are not likely transferred and dominantly expressed over the wild-type allele in most cases (Martinez, Coque and Baquero 2015). Likewise, chromosomally encoded ubiquitous genes, such as multidrug efflux pumps that encode proteins detoxifying antibiotics should not be considered as critical components, since they do not directly confer resistance to antibiotics that are used in human therapy (Martinez 2009). Furthermore, resistances to certain categories of antibiotics, such as those designated by the World Health Organization as 'critically important,' could be taken into consideration. Such aspects could be incorporated into the resistome risk score, and further improve the precision of the MetaCompare resistome risk ranking. Lastly, in the future, there is need to consider more complex factors influencing resistome risk, including mutation leading to evolution of new resistance determinants, natural selection amplifying concentrations of resistant strains, and context of horizontal gene transfer shuttling ARGs to a pathogen (Ashbolt *et al.* 2013).

FUNDING

This work was funded in part by USDA NIFA AFRI awards #2014-05280 and 2017-68003-26498 and National Science Foundation Partnership in International Research and Education award 1545756.

Conflicts of interest. None declared.

REFERENCES

Allen HK, Donato J, Wang HH *et al.* Call of the wild: antibiotic resistance genes in natural environments. *Nat Rev Microbiol* 2010;**8**:251–9.

Arango-Argoty G, Singh G, Heath LS *et al.* MetaStorm: a public resource for customizable metagenomics annotation. *PLoS One* 2016;**11**:e0162442.

Arango-Argoty GA, Garner E, Pruden A *et al.* DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* 2017, **6**:23.

Ashbolt NJ, Amezquita A, Backhaus T *et al.* Human Health Risk Assessment (HHRA) for environmental development and transfer of antibiotic resistance. *Environ Health Perspect* 2013;**121**:993–1001.

Bengtsson-Palme J, Antibiotic resistance in the food supply chain: Where can sequencing and metagenomics aid risk assessment? *Curr Opin Food Sci*. 2017, **14**:66–71.

Bhullar K, Waglechner N, Pawlowski A *et al*. Antibiotic resistance is prevalent in an isolated cave microbiome. *PLoS One* 2012;**7**:e34953.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;**30**:2114–20.

Boratyn GM, Camacho C, Cooper PS *et al*. BLAST: a more efficient report with usability improvements. *Nucleic Acids Res* 2013;**41**:W29–33.

Breitwieser FP, Lu J, Salzberg SL. A review of methods and databases for metagenomic classification and assembly. *Brief Bioinform* 2017.

Brown KD, Kulis J, Thomson B *et al*. Occurrence of antibiotics in hospital, residential, and dairy effluent, municipal wastewater, and the Rio Grande in New Mexico. *Sci Total Environ* 2006;**366**:772–83.

Clemente JC, Pehrsson EC, Blaser MJ *et al*. The microbiome of uncontacted Amerindians. *Sci Adv* 2015;**1**:e1500183.

D'Costa VM, King CE, Kalan L *et al*. Antibiotic resistance is ancient. *Nature* 2011;**477**:457–61.

Forsberg KJ, Reyes A, Wang B *et al*. The shared antibiotic resistome of soil bacteria and human pathogens. *Science* 2012;**337**:1107–11.

Hyatt D, Chen GL, Locascio PF *et al*. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010;**11**:119.

Jia B, Raphenya AR, Alcock B *et al*. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2017;**45**:D566–73.

Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a classification of mobile genetic elements, update 2010. *Nucleic Acids Res* 2010;**38**:D57–61.

Li B, Yang Y, Ma L *et al*. Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J* 2015;**9**:2490–502.

Liu B, Pop M. ARDB—antibiotic resistance genes database. *Nucleic Acids Res* 2009;**37**:D443–7.

Martinez JL. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc Biol Sci* 2009;**276**:2521–30.

Martinez JL, Coque TM, Baquero F. What is a resistance gene? Ranking risk in resistomes. *Nat Rev Microbiol* 2015;**13**:116–23.

Martínez JL. Antibiotics and antibiotic resistance genes in natural environments. *Science* 2008;**321**:365–7.

Olson ND, Treangen TJ, Hill CM *et al*. Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Brief Bioinform* 2017.

Peng Y, Leung HC, Yiu SM *et al*. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* 2012;**28**:1420–8.

Poirel L, Kampfer P, Nordmann P. Chromosome-encoded Ambler class A beta-lactamase of Kluyvera georgiana, a probable progenitor of a subgroup of CTX-M extended-spectrum beta-lactamases. *Antimicrob Agents Chemother* 2002;**46**:4038–40.

Poirel L, Rodriguez-Martinez JM, Mammeri H *et al*. Origin of plasmid-mediated quinolone resistance determinant QnrA. *Antimicrob Agents Chemother* 2005;**49**:3523–5.

Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987;**20**:53–65.

Rowe W, Verner-Jeffreys DW, Baker-Austin C *et al*. Comparative metagenomics reveals a diverse range of antimicrobial resistance genes in effluents entering a river catchment. *Water Sci Technol* 2016;**73**:1541–9.

Rowe WPM, Baker-Austin C, Verner-Jeffreys DW *et al*. Overexpression of antibiotic resistance genes in hospital effluents over time. *J Antimicrob Chemother* 2017;**72**:1617–23.

Schoen ME, Ashbolt NJ. An in-premise model for Legionella exposure during showering events. *Water Res* 2011;**45**:5826–36.

Teeling H, Glockner FO. Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. *Brief Bioinform* 2012;**13**:728–42.

Truong DT, Franzosa EA, Tickle TL *et al*. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015;**12**:902–3.

U.S. Environmental Protection Agency (U.S. EPA). *Human Health Risk Assessment* 2012, https://www.epa.gov/risk/human-health-risk-assessment.

Wattam AR, Davis JJ, Assaf R *et al*. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. *Nucleic Acids Res* 2017;**45**:D535–42.

Wooley JC, Godzik A, Friedberg I. A primer on metagenomics. *PLoS Comput Biol* 2010;**6**:e1000667.

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

Wright GD. The antibiotic resistome: the nexus of chemical and genetic diversity. *Nat Rev Microbiol* 2007;**5**:175–86.

Yang Y, Jiang X, Chai B *et al*. ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database. *Bioinformatics* 2016;**32**:2346–51.