Bioinformatics, 32(15), 2016, 2346–2351
doi: 10.1093/bioinformatics/btw136
Advance Access Publication Date: 12 March 2016
Original Paper

Data and text mining

ARGs-OAP: online analysis pipeline for antibiotic resistance genes detection from metagenomic data using an integrated structured ARG-database

Ying Yang[1,2,†], Xiaotao Jiang[1,†], Benli Chai[3,†], Liping Ma[1], Bing Li[1], Anni Zhang[1], James R. Cole[3], James M. Tiedje[3,*] and Tong Zhang[1,*]

[1]Environmental Biotechnology Laboratory, Department of Civil Engineering, The University of Hong Kong, Hong Kong, China, [2]School of Marine Sciences, Sun Yat-Sen University, Guangzhou, China and [3]Department of Plant, Soil, and Microbial Sciences, Michigan State University, East Lansing, MI, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.
Associate Editor: Jonathan Wren

Received on November 15, 2015; revised on February 17, 2016; accepted on March 5, 2016

Abstract

Motivation: Environmental dissemination of antibiotic resistance genes (ARGs) has become an increasing concern for public health. Metagenomics approaches can effectively detect broad profiles of ARGs in environmental samples; however, the detection and subsequent classification of ARG-like sequences are time consuming and have been severe obstacles in employing metagenomic methods. We sought to accelerate quantification of ARGs in metagenomic data from environmental samples.
Results: A Structured ARG reference database (SARG) was constructed by integrating ARDB and CARD, the two most commonly used databases. SARG was curated to remove redundant sequences and optimized to facilitate query sequence identification by similarity. A database with a hierarchical structure (type-subtype-reference sequence) was then constructed to facilitate classification (assigning ARG-like sequence to type, subtype and reference sequence) of sequences identified through similarity search. Utilizing SARG and a previously proposed hybrid functional gene annotation pipeline, we developed an online pipeline called ARGs-OAP for fast annotation and classification of ARG-like sequences from metagenomic data. We also evaluated and proposed a set of criteria important for efficiently conducting metagenomic analysis of ARGs using ARGs-OAP.
Availability and Implementation: Perl script for ARGs-OAP can be downloaded from https://github.com/biofuture/Ublastx_stageone. ARGs-OAP can be accessed through http://smile.hku.hk/SARGs.
Contact: zhangt@hku.hk or tiedjej@msu.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The prevalence of antibiotic resistance genes (ARGs) and the escalating threats they pose to public health have attracted worldwide concern (Laxminarayan *et al.*, 2013). A growing number of organizations and governments have enhanced their monitoring of antibiotic resistant bacteria and ARGs. The World Health Organization (2014) initiated a global surveillance of antibiotic resistance in 114 countries, with their first surveillance report released in April 2014; the Public Health Agency of Canada also provided a 4-year surveillance report on antibiotic resistant organisms in Canada (2014); the U.K. Parliament

released a note titled 'Antibiotic resistance in the environment' (2013) and the U.S. White House issued an executive order on national strategy (2014) for combating antibiotic resistance.

Along with concerns about antibiotic resistance in clinical settings, there has been increased interest in ARGs in the environment. ARGs have been detected in soil (Cytryn, 2013), natural waters (Pruden *et al.*, 2012) and sediments (Chen *et al.*, 2013), where they are potentially transferable from host bacteria to pathogens by horizontal gene transfer (Davies, 1994). ARGs are now regarded as emerging pollutants (Pruden *et al.*, 2006) and their dissemination in the environment has attracted much media attention.

Recently, metagenomic approaches have been applied to the investigation of ARGs in environmental samples. Nesme *et al.* (2014) studied the occurrence and abundance of ARGs in 71 environmental samples. Their results revealed the diversity and abundance of ARGs in different environments and suggested these genes were not randomly distributed. A study by Li *et al.* (2015) showed the influence of anthropogenic activities on the distribution of ARGs in 50 samples from 10 different environments. Through network analysis, they proposed *tet*M and genes coding aminoglycoside resistance as indicators that can be used to quantify other co-occurring ARGs. A more comprehensive investigation of the diversity and abundance of ARGs in various environments would provide insight into the distribution pattern of ARGs, the influence of human activity and the effectiveness of intervention or stewardship programs.

Any large-scale environmental analysis for ARGs by metagenomic sequencing of environmental samples requires an appropriate analysis method as well as a reference database. Several databases and analysis pipelines have been constructed specifically for ARG analysis, such as the Antibiotic Resistance Genes Database (ARDB) (Liu and Pop, 2009), the Comprehensive Antibiotic Resistance Database (CARD) (McArthur *et al.*, 2013), ResFinder (Zankari *et al.*, 2012), Antibiotic Resistance Gene Online (ARGO) (Scaria *et al.*, 2005) and ARG-ANNOT (Gupta *et al.*, 2013). ARGO focuses on vancomycin and beta-lactam resistance genes (Scaria *et al.*, 2005) while ARG-ANNOT was designed to detect ARGs in bacterial genomes instead of environmental samples. ResFinder offers ARG detection functions but needs longer query reads. For a sequence to be detected as an ARG in ResFinder, it must cover at least two-fifths of the length of the matching ARG in the database with no less than 50% identity (Zankari *et al.*, 2012). ARDB and CARD together provide most of the publicly available ARG sequences. They are the two most commonly used reference databases in the investigation of ARGs in environmental samples (Gibson *et al.*, 2015; Kristiansson *et al.*, 2011; Ma *et al.*, 2014; Yang *et al.*, 2013). Nevertheless, ARDB and CARD provide only limited online analysis capabilities for metagenomic data. Moreover, neither database can provide detailed ARG profiles for environmental samples, i.e. classifications of the identified ARGs-like sequences and abundance information for each detected ARG type/subtype, which makes it difficult for users to interpret the search results from the huge metagenomic data.

Here we describe our efforts in developing a Structured ARG database (SARG) by integrating sequences from ARDB and CARD in a hierarchical structure (type-subtype-reference sequence). Using SARG, ARG-like sequences obtained from similarity searches can be automatically assigned into different types and subtypes, avoiding tedious manual classification. Furthermore, an online analysis pipeline (ARGs-OAP) is available using SARG. This pipeline allows users to determine ARG profiles from large numbers of environmental samples in a single run. The pipeline aims to exploit the power of metagenomic analysis for better understanding the distribution and dissemination of ARGs in different environments.

2 Methods

2.1 Optimizing the integrated database from ARDB and CARD

The procedure for the optimization and construction of the integrated ARG database is summarized in Figure 1. Protein sequences were downloaded from CARD (http://arpcard.mcmaster.ca/, version on 15 April 2014) and ARDB (http://ardb.cbcb.umd.edu/index.html, version 1.1). The downloaded CARD database contained 2513 sequences while the ARDB database contained 7828 sequences. Only a small portion of sequences were shared by the two databases (586 sequences from ARDB and CARD with redundant sequences removed). Details of the identification of shared sequences process are in SI. The ARDB and CARD databases were first examined to remove non-ARG sequences (Supplementary Table S1). The two databases were integrated and redundant sequences were identified with a custom Perl script (SI). Redundant sequences were defined as sequences with 100% identity for the complete protein sequences. Each redundant group was manually inspected and the sequence with the most representative description was retained. By removing the redundant sequences, the size of the integrated database was reduced by 57% (from 10 337 sequences to 4401 sequences). The slimmed database was further examined to remove those ARG sequences related to single nucleotide polymorphisms (SNP), according to the updated CARD database. Moreover, sequences with description of 'hypothetical protein' or 'unnamed protein' were removed to ensure accurate and informative annotation of ARG-like sequences.

2.2 Construction and validation of SARG

Unlike the analysis of microbial communities or metabolic pathways, post-alignment analyses of ARG-like sequences from metagenomic data usually require significant error-prone manual work, mainly due to the need to sort ARG-like sequences into different types/subtypes in order to measure subtype abundance. Resistance 'Types' represent the class of antibiotics to which ARGs confer resistance, such as Type 'tetracycline resistance genes' or Type 'sulfonamide resistance genes'. Resistance 'Subtypes' represent

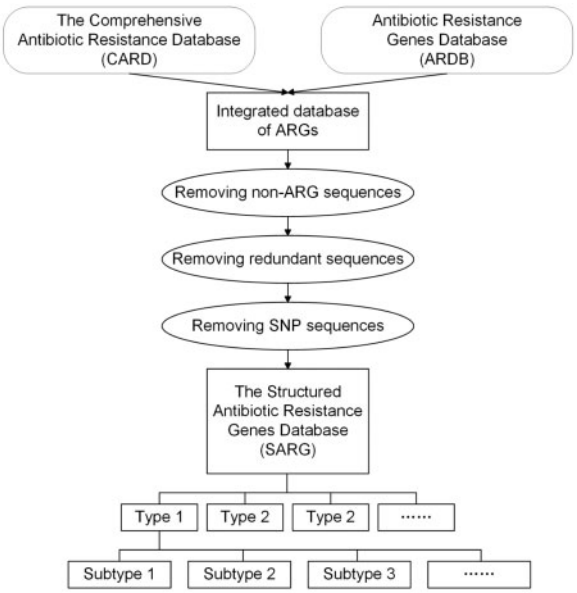

Fig. 1. Flowchart of constructing the integrated structured database of ARGs from ARDB and CARD

individual kinds of ARGs, such as subtype '*tet*A' of 'tetracycline resistance genes' or subtype '*sul*1' of 'sulfonamide resistance genes'.

The idea for a structured database had been proposed in previous studies (Ma *et al.*, 2014; Yang *et al.*, 2013) to facilitate classification of ARG-like sequences, with a hierarchical structure based on the concept of 'Types' and 'Subtypes'. The preliminary classification of different ARG types/subtypes in SARG was done based on sequence description through both keyword search and manual examination. In the present study, we adapted the classification of ARG 'Types' and 'Subtypes' from the previous structured ARDB and CARD. We first used the names of ARG subtypes as keyword to search the annotation of ARG sequences from the integrated database. The matched sequences were extracted and filed under the subtype's name in fasta format. We then checked these files one by one manually to ensure the correct classification of ARG sequences. The structure of the integrated database was also renewed from the updated information of ARG types and subtypes through literature search. The classification was further validated by sequence similarity. Potential problematic sequences (potential misplaced sequences) were identified through both direct BLASTP alignment and distance calculation. These problematic sequences were aligned with NCBI-NR database and the final classification of the problematic sequences was done according to the validated hits from NCBI-NR annotation. Our process of validation is further explained in SI.

2.3 Evaluation of similarity search method in ARG detection

The similarity search method outlined above was evaluated using the ARG sequences and non-ARG sequences from Swiss-Prot as test sequences (SI). The sequences from Swiss-Prot were enumerated into overlapping substrings (kmer) of different lengths, including 33 amino acids (aa), 50 aa, 67 aa and 100 aa, to mimic reads from Illumina high throughput sequencing (100, 150, 200 and 300 bp). These simulated sequences were annotated using ARGs-OAP. Ratios of false-positive, false-negative and cross-talking results were calculated at different E-value cutoffs, sequence identity and hit length. Using a specific cutoff, a read from the ARG set was counted as 'true positive (TP)' if it was correctly annotated to ARG (type or subtype) or 'false negative (FN)' if it was annotated as non-ARG sequence; a read from non-ARG set was defined as 'false positive (FP)' if it was annotated as ARG and 'true negative (TN)' if it does not match any sequence in the database at the specific cutoff. In cases where ARG reads were mistakenly annotated as a different ARG subtype, these reads were labeled as 'cross talking' (CT).

The Matthews Correlation Coefficient (MCC) (Matthews, 1975) was used to measure the effectiveness of the method on all tested values for *E*-value, identity and hit length. Two other criteria, sensitivity and precision, were also used to assess the performance of the method. Details for these measurements are summarized in SI.

2.4 Development of ARGs-OAP

ARGs-OAP was developed to facilitate the analysis of ARG sequences from metagenomic data, utilizing SARG and the Galaxy web server (Goecks *et al.*, 2010) with an intuitive user interface.

There are two parts of ARGs-OAP (Fig. 2): a pre-screening of potential ARG sequences using the user's local computers to reduce the size of sequence files for uploading followed by online annotation/classification of ARG sequences using a web-based platform. Before uploading metagenomic sequences to the online Galaxy based analysis platform (http://smile.hku.hk/SARGs), it is recommended to perform a local pre-screening for ARG-like and 16S rRNA gene sequences using UBLAST through the supplied Perl script (the script package can be downloaded from https://github.com/biofuture/Ublastx_stageone). This fast pre-screening was designed to remove irrelevant sequences which usually account for >99.3% of total sequences (Yang *et al.*, 2014), thus significantly reducing the size of file for uploading and accelerating the BLASTX analysis on the web-based platform. The potential ARG sequences uploaded to the web-based platform are aligned against SARG using BLASTX and classified according to the SARG hierarchy. The abundance of ARGs in the metagenomic data can be normalized to the ARG reference sequence length (nucleotide) and the number of 16S rRNA genes using the following equation from our previous study (Eq. 1) (Li *et al.*, 2015):

$$\text{Abundance} = \sum_1^n \frac{N_{\text{ARG-like sequence}} \times L_{\text{reads}} / L_{\text{ARG reference sequence}}}{N_{\text{16S sequence}} \times L_{\text{reads}} / L_{\text{16S sequence}}} \quad (1)$$

where $N_{\text{ARG-like sequence}}$ is the number of the ARG-like sequences annotated to one specific ARG reference sequence; L_{reads} represents the length of the reads; $L_{\text{ARG reference sequence}}$ is the nucleotide sequence length of the correspondingly specific ARG reference sequence; $N_{\text{16S sequence}}$ is the number of the 16S rRNA gene sequences; $L_{\text{16S sequence}}$ is the full length of 16S rRNA gene, i.e. 1432 bp in the present study; n is the number of mapped ARG reference sequences belonging to the ARG type or subtype. This normalized result avoids bias due to different prokaryotic DNA portions in environmental samples and is comparable to qPCR results normalized against the number of 16S rRNA genes.

The abundance of ARGs in the metagenomic data, which has been widely used in previous studies, could also be normalized against the ARG reference sequences length and the cell numbers from the metagenomic datasets using the web-based platform (Eq. 2). Cell numbers are estimated by retrieving microbial community identified by 16S rRNA gene hypervariable region from metagenomics data by USEARCH, calculating the average 16S rRNA gene copy number by copy number information according to CopyRighter database (Angly *et al.*, 2014) with abundance of each taxon as the weight, and finally dividing the total number of 16S rRNA gene sequences (converted to full sequence) by 16S rRNA gene average copy number to get the cell number (Eq. 3). Cell number is the estimated number of prokaryotic cells of the metagenomics dataset. Reads with hit length coverage of the database

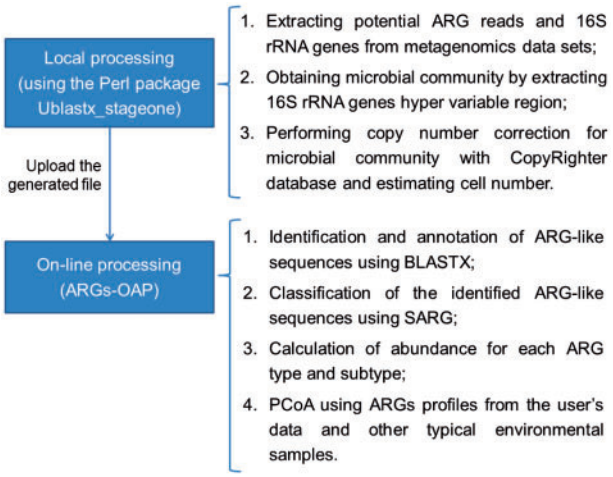

Fig. 2. Analysis flowchart and the functions of the Perl package for local process and ARGs-OAP

hypervariable region lower than 50% are excluded. The searching criteria of USEARCH are -query_cov 0.2 -target_cov 0.5 -id 0.9. The equations used in this normalization process are:

$$\text{Abundance} = \sum_1^n \frac{N_{\text{ARG-like sequence}} \times L_{\text{reads}} / L_{\text{ARG reference sequence}}}{\text{cell number}} \quad (2)$$

$$\text{Cell number} = \frac{\sum_1^n N_{\text{16S sequence}} \times L_{\text{reads}} / L_{\text{16S sequence}}}{\sum_{i=1}^m M_i \times a_i / A} \quad (3)$$

In Eq. 3 for cell number calculation, m represents the total taxa detected from metagenomics dataset according to extracted hyper variable region information; a_i is the number of aligned hypervariable sequences of taxon i in the metagenomics datas set; A is total number of aligned hypervariable sequences of all the m taxa and M_i represents the copy number of taxon i from CopyRighter database.

Details of cell number estimation from metagenomics data can be referred to UBLAST wiki document on ARGs-OAP (https://github.com/biofuture/Ublastx_stageone/wiki/Transform-ARGs-abundance-against-cell-number). We also evaluated the effectiveness of cell number estimation by metagenomics simulation datasets (SI).

3 Results

3.1 Overview of the database

After classification and validation of all the sequences, the SARG comprises of 23 ARG types, 1227 ARG subtypes and 4246 reference sequences (Fig. 3; Supplementary Table S2). The number of reference sequences in each ARG type and subtype are listed in Supplementary Tables S2 and S3.

Over 72% of the 1227 ARG subtypes belong to the beta-lactam resistance type (887 subtypes), although in total, the beta-lactam resistance subtypes account for only about 35% (1497) sequences in SARG. Following the beta-lactam resistance type, multidrug (935 sequences) and aminoglycoside (275 sequences) resistance types are additional major types in this database (Fig. 3).

3.2 Evaluation of ARGs-OAP for ARG annotation using simulated datasets

The effects of database completeness, cutoffs used for BLASTX (i.e. *E*-value, identity and hit length) and sequence length were investigated for their influences on the annotation of ARG sequences. To evaluate the influence of the database completeness on sequence annotation, two simulated datasets were used. Simulated Dataset 1 contained ARG sequences which were included in the integrated ARG database as well as some non-ARG sequences. Simulated Dataset 2 contained sequences from Simulated Dataset 1 and some ARG sequences in Swiss-Prot which were not included in the integrated ARG database. For Simulated Dataset 1, SARG was a complete reference database. While for Simulated Dataset 2, SARG was incomplete. Results showed that if the test data contained new ARGs sequences (Simulated Dataset 2), MCC values dropped significantly when the identity cutoff was set higher than 60% (Fig. 4a and b). Sensitivity was also reduced dramatically at this cutoff level (Figs 4d and 4e). However, the incompleteness of the database had little influence on the annotation precision (Fig. 4g and h).

The effects of *E*-value and identity on these three evaluation indices are also shown in Figure 4. The MCC value and precision increased with decreasing *E*-value but sensitivity did not change much. On the other hand, identity showed more impact compared to *E*-value. Under the cutoff usually applied in metagenomic analysis using short reads in previous ARGs studies (*E*-value of 1e-5 and identity of 90%) as shown by the blue arrow in Figure 4, MCC value and sensitivity were low for ARG annotation, indicating the false-negative rate was high and many ARG-like sequences were missed. To reveal a more comprehensive profile of ARGs using this similarity search method, the recommended cutoff is 60% for identity and 1e-7 for *E*-value, as indicated by the red arrow, based on MCC results as shown using Simulated Dataset 2.

The effect of hit length was investigated using the *E*-value 1e-7 and identity 60% cutoffs. Results showed that change in hit length

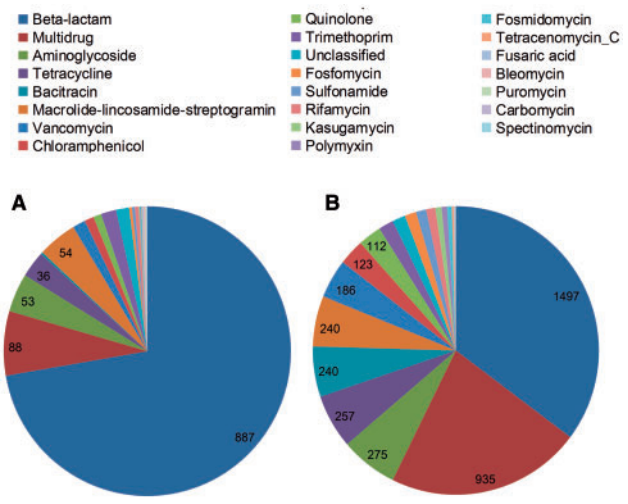

Fig. 3. Number of subtypes (**A**) and reference sequences (**B**) in the 23 ARG types in the integrated structured database of ARGs (Color version of this figure is available at *Bioinformatics* online.)

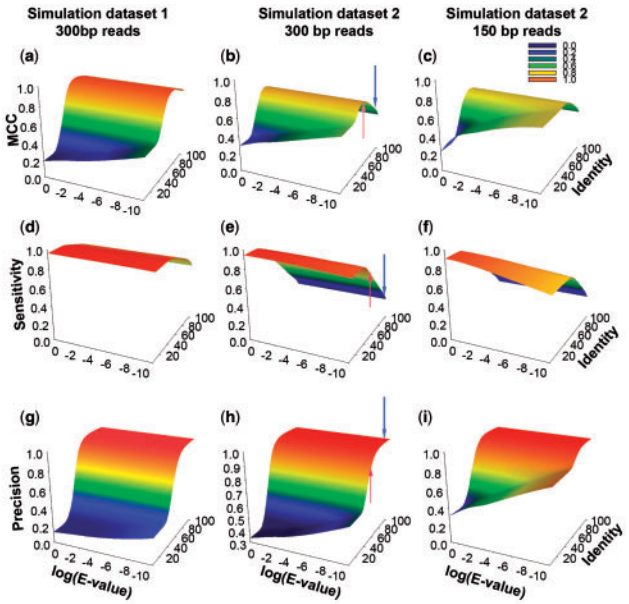

Fig. 4. Evaluation of similarity searching method to annotate ARGs. Mesh plots show MCC, sensitivity and precision state space under different conditions, e.g. different simulated datasets, different sequencing length and alignment criteria. Blue arrows showed value of identity as 90% and red arrows pointed to value identity as 60%. (**a**), (**d**) and (**g**) were MCC, sensitivity and precision state space for different *E*-values and identity tested by Simulated Dataset 1 of 300 bp metagenomics data simulation; b), e) and h) were results for Simulated Dataset 2 of 300 bp; (**c**), (**f**) and (**j**) were results for 150 bp Simulated Dataset 2. Comparison of (a), (d) and (g) with (**b**), (**e**) and (**h**) reflect the incompleteness of database (Color version of this figure is available at *Bioinformatics* online.)

had little influence on sensitivity and MCC value for hit length shorter than 85% of sequence length. But sensitivity and MCC value dropped drastically when hit length increased from 85 to 100% (Supplementary Fig. S1), indicating a more stringent hit length would miss more ARGs-like sequences.

Since the current sequencing platforms can generate reads with different lengths, the effect of sequence length was also evaluated. Figure 3b, 3c and Supplementary Figure S2 show that longer read lengths resulted in higher MCC and sensitivity. In addition, the effect of identity on annotation between ARG type and subtype is shown in Supplementary Figure S3. Generally, the MCC value, sensitivity and precision were higher at ARG type level than at subtype level (Supplementary Tables S4–S6). We assessed cross-talk ratios with different criteria using Simulated Dataset 2. The number of reads which were wrongly assigned to another subtype/type was counted and the fraction of these reads in the total dataset was calculated as the cross-talk ratio. The cross-talk ratio was $1.2 \pm 0.15\%$ at the subtype level and $0.19 \pm 0.21\%$ at the type level for Simulated Dataset 2 of 300 bp reads. The effect of identity, *E*-value and read length were evaluated (Supplementary Figs S4–S7). Longer read length, higher identity and lower *E*-value lowered the cross-talk ratios. Overall, the cross-talk ratios were very low with none exceeding 2 and 3% on type and subtype levels respectively for all simulated datasets.

In summary, this similarity search method achieved satisfactory results using proper hit cutoffs. Not surprisingly, longer query sequences improved annotation accuracy while the completeness of the reference database has a significant impact on annotation result. Of the hit cutoff variables applied during similarity search, hit length had little impact on annotation result when it was below 85% of the sequence length, while sensitivity and MCC decreased as hit length increased over 85%. An *E*-value of 1e-7 and identity of 60% were found to be the most suitable criteria considering MCC value, sensitivity and precision for ARG annotation using the current version of SARG. Cross-talk between different types was only a few at the recommended identity and *E*-value.

3.3 ARGs-OAP for annotation of ARGs in metagenomics datasets

As previously mentioned, the first step is pre-screening of potential ARGs sequences from metagenomic datasets using UBLAST on the user's local computer, followed by a second step of sequence annotation and classification using the online analysis platform after uploading identified potential ARGs sequences (Yang *et al.*, 2014). This pipeline supports multiple sample analysis, and generates a general table of ARG abundances (expressed as 'copies of ARG per copy of 16S rRNA gene' and 'copies of ARG per prokaryote's cell') which are normalized to the length of the corresponding ARG reference sequence as well as to the number of 16S rRNA genes and cells in the metagenomic data for better comparison to other studies (Li *et al.*, 2015). Analysis results are summarized in files which can be downloaded from the web-based platform, including (1) ARGs abundance of all uploaded samples normalized to 16S rRNA gene number at type, subtype level with other selected reference metagenomic datasets; (2) ARGs abundance normalized to cell numbers tables like (1); (3) and (4) PCoA figures for user uploaded samples and the reference datasets at subtype level. The reference metagenomic datasets adopted from Li *et al.*, (2015) include different environmental samples with various ARG abundances, which could be used to make a general comparison of ARG abundances between the user's datasets and other previous published metagenomic datasets. ARGs-OAP also provides a general Principal Coordinates Analysis (PCoA) using the uploaded datasets of user's samples and reference datasets. PCoA figures could be downloaded in PDF format. Examples of the output results are shown in Supplementary Table S7–S9 and Figure S9.

3.4 Time requirements

Time consumed in the first step of ARGs detection from metagenomic data using ARGs-OAP was evaluated using three kinds of samples with data size of 10 million reads (100 bp) each, including metagenomic data from chicken faeces, activated sludge and sediment, which cover the common range of ARG abundances. The time required for pre-screening of potential ARGs and 16S rRNA genes ranged from 105 to 124 min (Supplementary Table S10) for processing the 10 million reads in each dataset using the 64 bit version of UBLAST and 1 thread on a local computer (Lenovo ThinkStation-D20: CPU 2.40 GHz × 8 cores; Memory 96 GB). A file containing about 307 thousand potential ARG-like reads in total was generated from the three datasets through our supplied script, in which the number of sequences were significantly reduced from 30 million reads in the original datasets.

4 Discussion

The time required for the similarity search and the post-alignment analysis has become the limiting factor in metagenomic studies as sequencing costs decrease and data sizes grow (Mardis, 2011). The major goal of the present study was to provide an online platform to shorten the time required for the detection of ARGs and to improve the classification and enumeration of ARG-like sequences by constructing a structured database.

First, after removing redundant sequences, the ARDB and CARD databases were integrated to provide more comprehensive detection of ARG-like sequences in metagenomic data, which will increase the accuracy and completeness of ARG profiles from environmental samples with diverse microbial communities. Since the search time is dependent on the size of reference database, removing redundant sequences in the two reference databases was an obvious way to reduce the search time without compromising the completeness of the reference database.

Second, a structured database was established for the integrated database following our previous idea of a structured ARDB (Yang *et al.*, 2013). This structured database made automatic classification of different ARG types and subtypes possible, therefore manual classification is no longer needed.

Third, the online analysis platform, ARGs-OAP, was developed for ARG detection applying our previous proposed hybrid approach using UBLAST and BLASTX for ARG annotation of metagenomic sequence data (Yang *et al.*, 2014). The pre-screening of potential ARG sequences could filter out most (for ARGs, more than 99.3%) irrelevant sequences and significantly shorten the required time for the BLASTX process on the online platform. Moreover, the classification of the identified ARG-like sequences was conducted automatically using SARG, which will greatly reduce the required time for post-alignment processing, and automatic generation of PCoA from ARGs-OAP may help users with a rapid assessment of differences between the ARG profiles in their samples and in other reference environmental samples.

Given the less-biased nature of metagenomic datasets, utilizing the ARGs-OAP, different research groups may compare their datasets in a way that is more consistent than other current approaches,

and that can help reveal the distribution of ARGs in different environments.

ARGs-OAP also supports analysis under different criteria (i.e. *E*-value, hit length and identity). If the user wishes, these criteria can be adjusted according to their specific requirements via a platform interface instead of using the default setting, i.e. *E*-value of 1e-7, hit length of 75% of read length (i.e. 25 aa for reads of 100 bp) and identity of 80%.

Our evaluation indicates that the completeness of the database is critical for the comprehensive annotation of ARGs in metagenomic data by similarity search. It is likely that inclusion of additional novel ARG sequences would improve performance. The current database of SARG was constructed from the two most widely applied databases, i.e. ARDB and CARD (Version on 15 April 2014). It is clear that our integrated structured database does not cover all ARG sequences, although it already contains most ARG sequences available in public database. Thus, as for other gene types, updates with new sequences will be important, as they become available. Moreover, annotation evaluation should be repeated when new ARGs sequences are added in the database to achieve accurate annotation for the ARGs profiles in metagenomic data. We believe this pipeline is a powerful method for more comprehensive ARG profiles assessment from metagenomic data, and hence lead to methods to reduce ARG dissemination to the environment and its dissemination to pathogens.

Acknowledgements

Xiaotao Jiang, Liping Ma and Anni Zhang thank The University of Hong Kong for the postgraduate studentship. Dr. Ying Yang and Dr. Bing Li thank The University of Hong Kong for the postdoc fellowship.

Funding

This work has been supported by the Hong Kong GRF (172099/14) and the Center for Health Impacts of Agriculture (CHIA) at Michigan State University.

Conflict of Interest: none declared.

References

Angly,F.E. *et al*. (2014) CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, **2**, 1–13.

Chen,B. *et al*. (2013) Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. *Environ. Sci. Technol*., **47**, 12753–12760.

Coleman,P. *et al*. (2013) *Antibiotic Resistance in the Environment*. Parliamentary Office of Science and Technology, UK.

Cytryn,E. (2013) The soil resistome: the anthropogenic, the native, and the unknown. *Soil Biol. Biochem*., **63**, 18–23.

Davies,J. (1994) Inactivation of antibiotics and the dissemination of resistance genes. *Science*, **264**, 375–382.

Gibson,M.K. *et al*. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J*., **9**, 207–216.

Goecks,J. *et al*. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome. Biol*., **11**, R86.

Gupta,S.K. *et al*. (2013) ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother*., **58**, 212–220.

Kristiansson,E. *et al*. (2011) Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS ONE*, **6**, e17038.

Laxminarayan,R. *et al*. (2013) Antibiotic resistance-the need for global solutions. *Lancet Infect. Dis*., **13**, 1057–1098.

Li,B. *et al*. (2015) Metagenomic and network analysis reveal wide distribution and co-occurrence of environmental antibiotic resistance genes. *ISME J*., **1–13**, 1751–7362.

Liu,B. and Pop,M. (2009) ARDB-antibiotic resistance genes database. *Nucleic Acids Res*., **37**, D443–D447.

Ma,L. *et al*. (2014) Abundant rifampin resistance genes and significant correlations of antibiotic resistance genes and plasmids in various environments revealed by metagenomic analysis. *Appl. Microbiol. Biotechnol*., **99**, 5195–5204.

Mardis,E.R. (2011) A decade's perspective on DNA sequencing technology. *Nature*, **470**, 198–203.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta (BBA) – Protein Struct*., **405**, 442–451.

McArthur,A.G. *et al*. (2013) The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother*., **57**, 3348–3357.

Nesme,J. *et al*. (2014) Large-scale metagenomic-based study of antibiotic resistance in the environment. *Curr. Biol*., **24**, 1096–1100.

Pruden,A. *et al*. (2012) Correlation between upstream human activities and riverine antibiotic resistance genes. *Environ. Sci. Technol*., **46**, 11541–11549.

Pruden,A. *et al*. (2006) Antibiotic resistance genes as emerging contaminants: studies in Northern Colorado. *Environ. Sci. Technol*., **40**, 7445–7450.

Public Health Agency of Canada. (2014) Antimicrobial resistant organisms (ARO) surveillance report – 2009–2013.

Scaria,J. *et al*. (2005) Antibiotic Resistance Genes Online (ARGO): a database on vancomycin and *β*-lactam resistance genes. *Bioinformation*, **1**, 5–7.

The White House, Washington. (2014) National Strategy for Combating Antibiotic-Resistant Bacteria.

World Health Organization. (2014) Antimicrobial Resistance Global Report on Surveillance.

Yang,Y. *et al*. (2014) Evaluation of a hybrid approach using UBLAST and BLASTX for metagenomic sequences annotation of specific functional genes. *PLoS ONE*, **9**, e110947.

Yang,Y. *et al*. (2013) Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environ. Sci. Technol*., **47**, 10197–10205.

Zankari,E. *et al*. (2012) Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother*., **67**, 2640–2644.