OMICS A Journal of Integrative Biology
Volume 20, Number 4, 2016

DOI: 10.1089/omi.2015.0191

Antibiotic Resistome: Improving Detection and Quantification Accuracy for Comparative Metagenomics

Ali H. A. Elbehery,[1] Ramy K. Aziz,[2] and Rania Siam[1,3]

Abstract

The unprecedented rise of life-threatening antibiotic resistance (AR), combined with the unparalleled advances in DNA sequencing of genomes and metagenomes, has pushed the need for *in silico* detection of the resistance potential of clinical and environmental metagenomic samples through the quantification of AR genes (i.e., genes conferring antibiotic resistance). Therefore, determining an optimal methodology to quantitatively and accurately assess AR genes in a given environment is pivotal. Here, we optimized and improved existing AR detection methodologies from metagenomic datasets to properly consider AR-generating mutations in antibiotic target genes. Through comparative metagenomic analysis of previously published AR gene abundance in three publicly available metagenomes, we illustrate how mutation-generated resistance genes are either falsely assigned or neglected, which alters the detection and quantitation of the antibiotic resistome. In addition, we inspected factors influencing the outcome of AR gene quantification using metagenome simulation experiments, and identified that genome size, AR gene length, total number of metagenomics reads and selected sequencing platforms had pronounced effects on the level of detected AR. In conclusion, our proposed improvements in the current methodologies for accurate AR detection and resistome assessment show reliable results when tested on real and simulated metagenomic datasets.

Introduction

The advent of high-throughput sequencing technologies, initially dubbed next-generation sequencing (NGS), had a great impact on biomedical and environmental sciences (Voelkerding et al., 2009). Reduced cost, rapid sequencing, and enhanced accuracy are only a few advantages (Shendure and Ji, 2008); yet the biggest impact of those technologies is opening the gates for thousands of genomic and metagenomic studies to accelerate biological discovery and improve knowledge about our biosphere and our own bodies.

One of the major health problems that DNA sequencing can help solving is the rapid emergence of antibiotic-resistant (AR) pathogens, which are expected to cause more deaths than cancer in 2050 (O'Neill, 2014). To address this problem, efforts have been targeted at sequencing the genomes of AR bacterial strains and human-associated metagenomes to track the dynamics of AR gene transfer. However, attention has lately been directed to tracking antibiotic resistance in the environment, because environmental microorganisms are believed to act as reservoirs for AR genes that can be transferred to pathogenic organisms (Martinez, 2008). AR genes have been discovered in a plethora of environments, including pristine ones (Bhullar et al., 2012; Brown and Balkwill, 2009; Toth et al., 2010), which could explain the continuous emergence of novel AR genes. Thus, further studies of environmental AR genes may give better insight into their potential health risks, as well as their ecological impacts on microbial population dynamics.

Several studies used metagenomics to quantitatively and qualitatively assess AR in various environments (Bengtsson-Palme et al., 2014; Chao et al., 2013; Chen et al., 2013; Kristiansson et al., 2011; Ma et al., 2014; Wang et al., 2013; Yang et al., 2013; Zhang et al., 2011). However, these studies used different NGS platforms, different methodologies for metagenome data analysis, different normalization procedures, and different databases to study AR. For example, BLASTX is the most popular method for mapping sequence reads to AR genes (Chao et al., 2013; Chen et al., 2013; Zhang et al., 2011), but one study (Bengtsson-Palme et al., 2014) used Vmatch sequence analysis software (http://www.vmatch.de/). Some

[1]Graduate Program of Biotechnology and [3]Biology Department, The American University in Cairo, Cairo, Egypt.
[2]Department of Microbiology and Immunology, Faculty of Pharmacy, Cairo University, Cairo, Egypt.

studies used the Antibiotic Resistance Gene database (ARDB) (Chen et al., 2013; Liu and Pop, 2009; Wang et al., 2013; Zhang et al., 2011), while others used the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013) and Resqu antibiotic resistance database (http://www.1928diagnostics.com/resdb/) (Bengtsson-Palme et al., 2014; Ma et al., 2014).

This variability of methods is typical of genomic/bioinformatics studies, but makes the comparison of different studies challenging, and highlights the pressing need for optimization and standardization of such analysis pipelines. In addition, most of the published studies either overlooked or falsely predicted AR conferred by mutations (e.g., AR generated by mutations in *rpoB, gyrA, gyrB, parC, parE*, etc.).

Mutation-generated AR is one of the major mechanisms of resistance. These mutations usually modify the antibiotic target in a way that makes it irresponsive to the antibiotic. In other instances, they modify antibiotic transporters or alter enzymes that activate antibiotic prodrugs (Martínez and Baquero, 2014). Mutations could mediate resistance to several classes of antibiotics (e.g., rifamycins, fluoroquinolones, oxazolidinones and fusidanes). This mechanism is clinically important because it is the principal resistance mechanism in certain microorganisms, such as *Mycobacterium tuberculosis* and *Helicobacter pylori*. In addition, resistance to certain antibiotic classes (e.g., fluoroquinolones and oxazolidinones) is almost exclusively produced via such mutations (Woodford and Ellington, 2007). The spread of mutation-mediated resistance is well evidenced for rifampin (Ferrándiz et al., 2005) and fluoroquinolones (Balsalobre et al., 2003; Ferrándiz et al., 2000; Pletz et al., 2006) in clinical bacterial isolates. Therefore, it is essential to detect such mutations accurately, especially in metagenomic studies in which they are often ignored or falsely assigned.

Because of the limitations of current analysis pipelines, we set out to determine the major factors affecting accurate quantification of AR gene abundance in metagenomes and to optimize the current methodologies to avoid these major confounding factors. In this context, we propose an amendment that accounts for genes whose mutation leads to antibiotic resistance. In addition, we evaluated the influence of this modification using previously published metagenomic datasets.

Materials and Methods

Methodology used for AR estimation within an environment (resistome analysis)

First, BLASTX (Altschul et al., 1990) was used to align metagenomic reads to AR polypeptides from the Comprehensive Antibiotic Resistance Database (CARD) (McArthur et al., 2013). Reads with matches passing a threshold of 90% identity over at least 25 amino acids (Kristiansson et al., 2011)

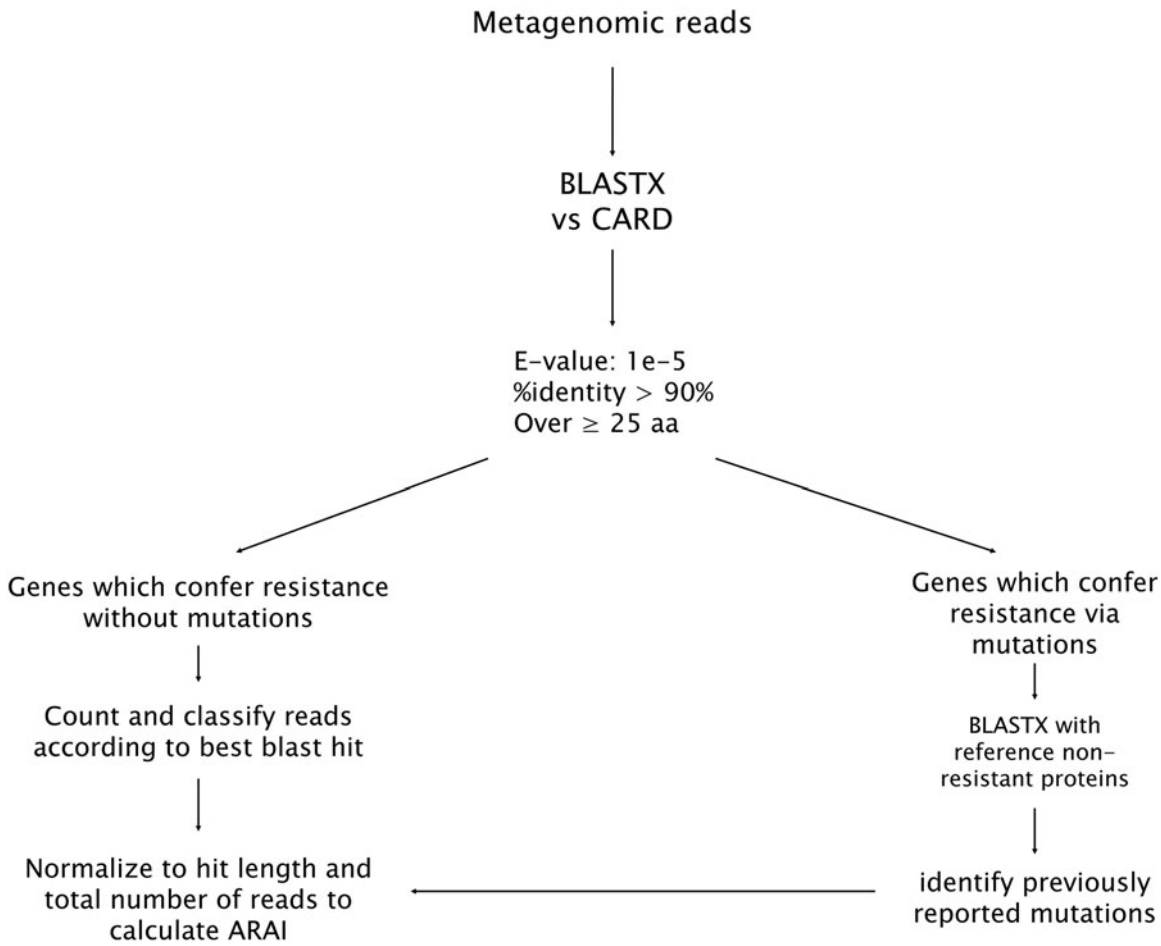

FIG. 1. Flowchart of proposed antibiotic resistance gene detection pipeline. CARD, Comprehensive antibiotic resistance database; ARAI, Antibiotic resistance abundance index.

were assigned the function of their best BLASTX hit and then binned into appropriate AR gene classes. If a read was assigned to an antibiotic target gene, which could possibly have mutation(s), it was further characterized by alignment to the respective gene (Fig. 1). Only reads with previously reported nonsynonymous mutations (Supplementary Table S1; supplementary material is available online at ftp.liebertpub.com/omi) were retained. Such filtering was performed by a custom Perl script (available through: https://github.com/aelbehery/mutation-detection).

Assessing the improved resistome analysis methodology

We analyzed three different metagenomes (Supplementary Table S2) using our proposed improved methodology. Metagenomes were downloaded from Sequence Read Archive (SRA). To allow direct comparison of results, precise quality control of the data was performed as described in the respective publications (Bengtsson-Palme et al., 2014; Ma et al., 2014; Zhang et al., 2011). As previously described, Pearf was used for quality filtering of the Swedish lake metagenome (Pearf options ''-q 28 -f 0.25 -t 0.05 -l 30''). Reads with ≥10% ambiguous bases and/or ≥50% of bases with a Phred score lower than 20 were eliminated in the fresh water fish pond sediment metagenome. In the case of the plasmid metagenome, we discarded reads with a number of ambiguous bases ≥3 or those contaminated by adapters. For the analysis of the Swedish lake metagenome, we used a threshold of 95% identity over at least 20 amino acids similar to the stringency threshold used by Bengtsson-Palme and coworkers (Bengtsson-Palme et al., 2014) to neutralize this factor. In case of the fresh water fish pond sediment metagenome, we repeated the reported analysis (Ma et al., 2014) using an updated version of CARD—the same version that was used in the assessment of our method as well.

Metagenomic simulation experiments

Effect of genome size. Six bacterial chromosomes with different lengths ranging from 0.58 Mbp to more than 10 Mbp (Supplementary Table S3) were used to inspect the influence of genome size on the detected number of AR reads. Each chromosome was manually spiked with an *ampC* beta lactamase gene (genome accession: NC_002516, position: 4594029–4595222 bp, length: 1194 bp). This was followed by *in silico* metagenomic read generation for each chromosome separately. The *in silico* metagenomes were generated with MetaSim (Richter et al., 2008) with the following parameters: error model: user-defined empirical model, made according to software manual instructions to simulate 100 bp reads produced by Illumina sequencing technology, read length: 100 bp, number of reads: 100,000, mate pairs: false. Generated reads were analyzed as described above to determine the number of AR reads for each chromosome.

Effect of AR gene length. Six genes of varying lengths, ranging from 308–3108 bp (Supplementary Table S4), were used to investigate the influence of AR gene length on the number of detected AR reads. These genes were manually inserted, evenly spaced, in the genome of *Mycoplasma genitalium*. Subsequently, 100,000 simulated Illumina reads, each with a length = 100 bp, were generated by MetaSim and analyzed, as described above, to determine the number of AR reads.

Effect of number of reads. To study the effect of generated number of reads on AR gene quantification, we used the same parameters described in assessing AR gene length to produce varying numbers of reads (50K, 100K, 150K, 200K, and 250K). Likewise, the reads were analyzed as described above.

Effect of read length. To study the effect of read length on AR gene detection and quantification, we generated simulated metagenomes with different read lengths (80, 100, 150, and 200 bp) while fixing the number of reads to 100K. The 80 bp Illumina error model was downloaded from the MetaSim website (http://ab.inf.uni-tuebingen.de/software/metasim/errormodel-80bp.mconf). Error models for simulating Illumina reads of 100, 150, 200, and 250 bp length were created according to the instructions in the MetaSim manual. Again, generated reads were analyzed as explained above.

Effect of sequencing platform. The same simulation experiments used for studying the effect of AR length were repeated, but with MetaSim's built-in models for Roche-454 and Sanger sequencing platforms. The total number of reads for each platform was 100,000. Average read length was selected to be 400 bp for 454, and 1000 bp for Sanger. Generated reads were analyzed as above.

All simulation experiments were performed in triplicates.

Results

In this study, we provide an improved methodology for accurate quantification of antibiotic resistance in metagenomes (resistome analysis) based on identifying a caveat in the current AR detection pipelines. In addition, we use the improved methodology to systematically assess the influence of five potential confounding factors (Table 1) that are likely to skew the number of retrieved AR reads.

Description of the antibiotic resistome analysis methodology

The workflow presented here effectively addresses pitfalls in studying AR generated by target gene mutations in metagenomes. The method depends on CARD, a comprehensive AR database, which includes AR genes, AR gene SNPs, and antibiotic target genes (cf., ARDB). CARD BLAST hits either align to acquired AR genes or to antibiotic target genes. The latter are carefully checked for the presence of previously reported AR-producing mutations as indicated in Materials and Methods.

We evaluated this workflow by analyzing three different metagenomes, each previously analyzed by a different AR detection pipeline (Tables 2 and 3). The first was a Swedish lake metagenome (Bengtsson-Palme et al., 2014) previously reported to have 10 AR reads. Our method detected 814 AR hits, 20 of which were nonsynonymous mutations in target genes (Table 2). The second metagenome was a freshwater fish pond sediment (Ma et al., 2014). This metagenome was previously shown to contain 51.8% mutation-generated AR genes (*gyrA*, *gyrB*, *parC*, *parE, and rpoB*). These genes are antibiotic

Table 1. Factors Affecting the Number of Retrieved Antibiotic Resistance Reads Using Metagenomic Approaches

Factor	*Description*	*Effect*
Genome size	Length of the genome in which the AR gene is found.	The larger the genome size, the lower the chance of an AR gene in this genome to be represented in the metagenome.
AR gene length	Length of the AR gene from which AR metagenomic reads are generated.	The longer the AR gene, the higher the chance it will generate more AR metagenomic reads.
Metagenome size	The total number of reads of a metagenome	The larger the metagenome size the larger the number of retrieved AR reads.
Read length	The average length of metagenomic reads. It varies for different platforms	Read length showed no significant effect on the number of retrieved AR reads. Nevertheless, reads need to be longer than the minimum alignment length set in the AR detection pipeline (at least 75 bp)
Platform	Sequencing platform used for generation of the metagenomic reads.	Each platform produces significantly different results due to differences in sequencing errors.

target genes that, otherwise, play essential functions in bacteria. When we used the updated version of CARD without mutation screening, the ratio of target genes, which could possibly have mutations was more or less the same (49.7%). In contrast, after these target genes were screened for non-synonymous mutations, only 3.3% of them contained resistance-conferring mutations. The third metagenome was a plasmid metagenome from a sewage treatment plant, in which Zhang and colleagues detected 699 AR reads (Zhang et al., 2011), while we detected 1833 reads, 61 of which were target gene mutations.

Assessing the impact of intrinsic genome characteristics on AR quantification results

We examined the impact of intrinsic genome characteristics, such as genome size and AR gene length, on AR detection in metagenomes, using a set of *in silico* simulations. To examine the effect of genome size on the number of retrieved AR reads, we chose six sequenced chromosomes with varying genome sizes. The same AR gene was used to spike each of the six chromosomes, and then these spiked chromosomes were *in silico*-fragmented to generate simulated metagenomes, which we analyzed using our method. The analysis demonstrated that the larger the genome size, the lower the number of retrieved AR reads. The relation fits an exponential decay curve with a regression coefficient $R^2 = 0.9945$ (Fig. 2).

We similarly conducted an *in silico* experiment to examine the effect of AR gene length on the number of retrieved AR reads within a metagenomic sample. In this case, one genome was spiked with six different AR genes with varying lengths to generate one simulated metagenome. The number of detected AR reads was clearly linearly correlated with AR gene length (Pearson correlation coefficient $r = 0.9985$; $R^2 = 0.9931$, Fig. 3). Longer AR genes have higher chances of being detected.

Table 2. Comparing the Pipeline to Previously Used Methods

	Study results		*Proposed pipeline results*	
Study	*Total AR reads*	*Target gene mutations*	*Total AR reads*	*Target gene mutations*
(Bengtsson-Palme et al., 2014)	10	0	814	20
(Ma et al., 2014)	9293	4621	4829	157
(Zhang et al., 2011)	699	0	1833	61

The pipeline was applied to the metagenomes after quality control as described by their respective publications.

Assessing the impact of technical differences in metagenome data on AR quantification results

We similarly performed *in silico* simulation experiments to assess the impact of technical differences in metagenome data including variations in metagenome size, read length, and sequencing platforms on the retrieved AR results. To investigate the effect of metagenome size on AR gene retrieval and quantification, we repeated the same experiment conducted to examine the effect of AR gene length, but we performed the experiment on five simulated metagenomes with five different sizes (i.e., total number of reads per metagenome). A linear correlation between AR gene length and the detected number of reads was evident for the five different metagenomes (R^2 of 50K reads = 0.9845, 100K reads = 0.9931, 150K reads = 0.9923, 200K reads = 0.9965, and 250K reads = 0.9955). Thus, the higher the number of reads, the higher the number of detected AR reads and the steeper the curve (Fig. 4a). To verify the linearity of this relation, we plotted the slopes of the curves against the number of reads (Fig. 4b). The relation was linear ($R^2 = 0.997$).

Simulation experiments were repeated for metagenomes with different sequence read lengths. Whereas the relation between AR gene length and the number of detected AR reads remained linear for all read lengths (R^2 values: 80 bp = 0.9889, 100 bp = 0.9931, 150 bp = 0.9954, and 200 bp = 0.9977), sequence read length, *per se*, had no significant effect on the number of detected AR reads. Regression lines nearly

Table 3. Antibiotic Resistance Detection Methods Used in Selected Studies

Database(s) used	*Account for antibiotic target genes with possible mutations*	*Alignment method*	*Stringency threshold*	*Mutation detection*	*Comments*	*Reference(s)*
ARDB	No	BLASTX	90% identity over at least 25 amino acids	None		(Zhang et al., 2011; Chen et al., 2013; Wang et al., 2013)
Clean ARDB	No	BLASTX	90% identity over at least 25 amino acids	None	Clean ARDB consisted of ARDB after removal of redundant sequences	(Yang et al., 2013)
ARDB, CARD and core database of ARDB and CARD	ARDB: No; CARD: Yes	BLASTX	90% identity over at least 25 amino acids	None	Core database of ARDB and CARD was created by aligning sequences in ARDB to sequence in CARD with a cutoff of 1E-6	(Chao et al., 2013)
ARDB + sequences from known quinolone resistance genes consisting of sequences from *qnrA-D*, *qnrS*, *qepA*, *acrA-B*, *norA-C* and *oqxA-B*	No	BLASTX	90% identity over at least 25 amino acids	None		(Kristiansson et al., 2011)
CARD	Yes	BLASTX	90% identity over at least 25 amino acids	None		(Ma et al., 2014)
Resqu antibiotic resistance database	No	Vmatch sequence analysis software	95% identity over at least 20 amino acids	None	Resqu contained 3019 non-redundant horizontally transferred antibiotic resistance genes manually extracted from literature	(Bengtsson-Palme et al., 2014)

AR, antibiotic resistance; ARDB, Antibiotic Resistance Gene Database; CARD, Comprehensive Antibiotic Resistance Database.

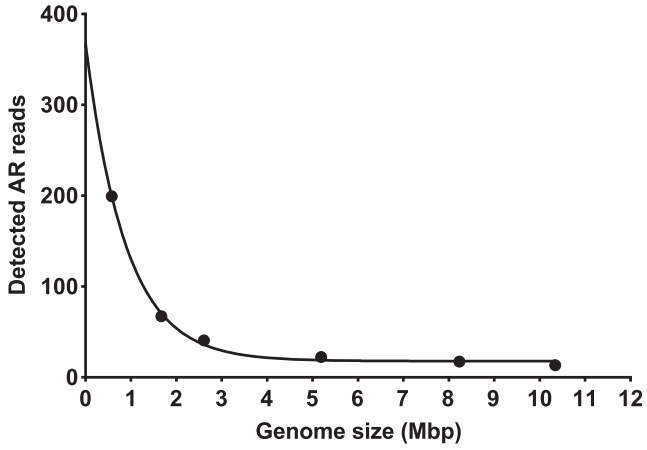

FIG. 2. Metagenomic simulation experiment to test the effect of genome size on the number of detected AR reads. The detected AR reads in each of the six different chromosomes, manually spiked with the same AR gene, are presented. The larger the genome size, the lower the number of retrieved AR reads. $R^2 = 0.9945$.

superimposed (Fig. 5), and analysis of variance (ANOVA) confirmed that the detected AR reads of all read lengths tested were not significantly different ($p = 0.9981$).

The effect of three different sequencing platforms, namely Illumina, Roche-454, and Sanger was studied (Fig. 6). Interestingly, the linearity of the relation between AR gene length and the number of detected AR reads was not affected by platform change (R^2 values for the different platforms were 0.9521, 0.9971, and 0.9773 for 454, Illumina, and Sanger, respectively). However, the number of detected AR reads was highest in case of Sanger, followed by Illumina, followed by 454. Slopes of the curves for the different platforms were significantly different from one another (ANOVA: $p < 0.0001$; Tukey's *post hoc* multiple comparison: 454 vs. Illumina: $p < 0.0001$, 454 vs. Sanger: $p < 0.0001$, Illumina vs. Sanger: $p < 0.0055$).

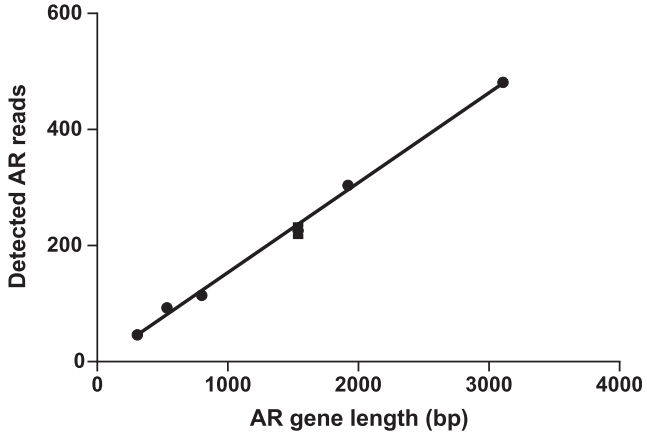

FIG. 3. Metagenomic simulation experiments to test the effects of AR gene length on the number of detected AR reads. Six different AR genes with varying lengths were introduced into one chromosome, which was fragmented using MetaSim. The number of AR reads detected in this simulated metagenome are presented. The longer AR gene, the higher its chance of being detected. Correlation coefficient $r = 0.9985$; $R^2 = 0.9931$.

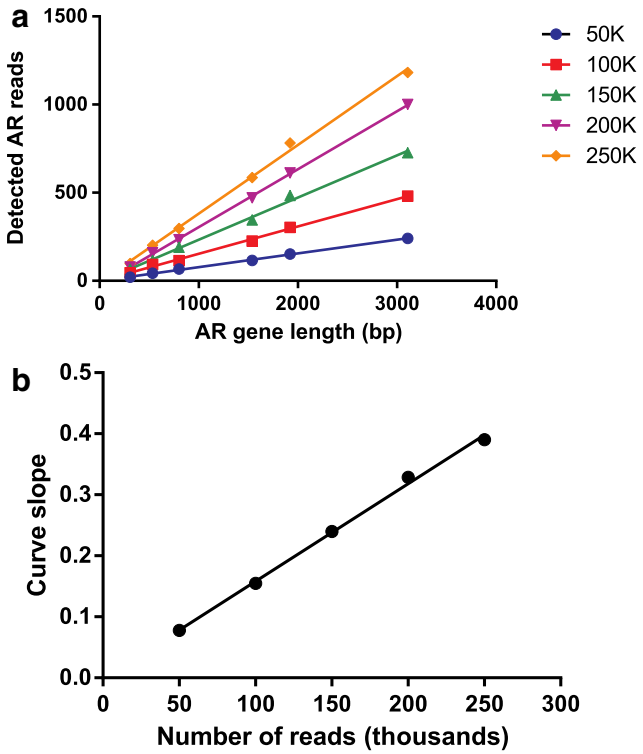

FIG. 4. Metagenomic simulation experiments to test the effects of metagenome size on the number of detected AR read. **(a)** A plot showing the relation between AR gene length and the number of detected AR reads for five different metagenome sizes (50, 100, 150, 200, and 250 thousand reads). **(b)** A plot between the slopes of the 5 curves shown in **(a)** and metagenome size. $R^2 = 0.9970$.

Discussion

In an attempt to improve antibiotic resistome analysis, we tested, optimized, and improved current AR detection methodologies, and we suggest specific modifications to the currently used pipelines (Fig. 1). Additionally, we tested our workflow on various published and simulated metagenomes to evaluate the methodology and to study factors that may affect the number of retrieved AR reads.

Of note, although sequence identity alone does not necessarily imply functional similarity, sequence similarity remains the standard method for function prediction in metagenomics, because high-throughput sequencing technologies produce relatively short reads. Therefore, similar to previous methods, we relied on BLASTX to detect resistance genes. To avoid false predictions based on partial similarities, we used a rather stringent threshold (90% identity over ≥25 amino acids) for selection of positive AR reads, a threshold previously suggested (Kristiansson et al., 2011). This threshold was selected because we mainly wanted to shed light on the abundance of known antibiotic resistance genes or altered target genes. Another target of resistome studies is, of course, gene discovery. In that case, we suggest using less stringent thresholds. When new resistance gene discovery is the purpose, choosing assemblies and filtering them with different methods (including Hidden Markov Models, protein family analysis, and motif analysis) is recommended.

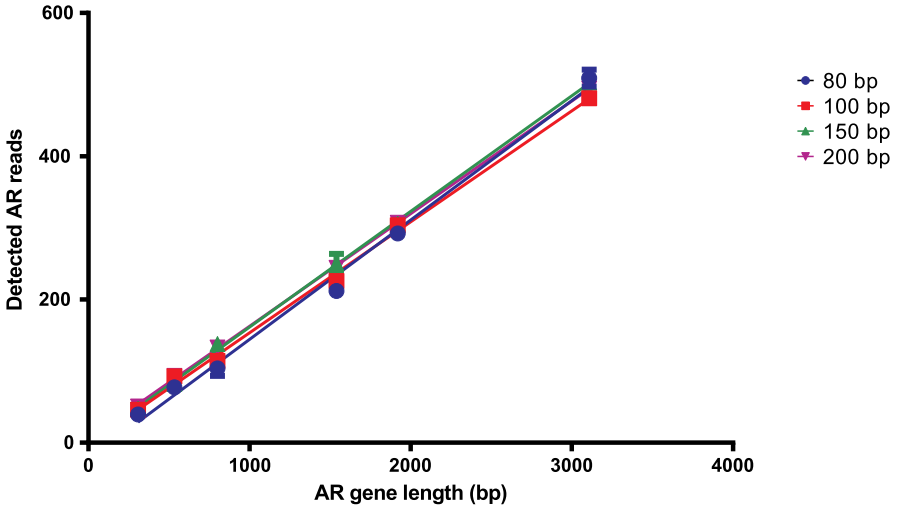

FIG. 5. Metagenomic simulation experiments to test the effect of read length on the detected number of AR reads. Simulation experiments were repeated for four different read lengths. Detected AR reads linearly correlated to AR gene length for all read lengths (R^2 values: 80 bp = 0.9889, 100 bp = 0.9931, 150 bp = 0.9954 and 200 bp = 0.9977). Regression lines nearly superimposed (ANOVA shows no significant difference, $p = 0.9981$).

The major improvement in our methodology is allowing the sensitive but accurate identification of AR conferred by mutations. Several metagenomic studies for AR detection largely relied on the Antibiotic Resistance Genes Database (ARDB) (Liu and Pop, 2009), the first antibiotic resistance database developed (Table 3). However, ARDB does not account for AR resulting from target gene polymorphism. Therefore, studies relying on the ARDB database alone accounted for acquired or horizontally transferred AR and neglected mutational AR. Later, CARD was launched, and it included antibiotic target genes (e.g., *rpoB*, *gyrA* and *gyrB*). However, these are highly conserved housekeeping genes and are therefore present in virtually all bacteria, where they perform essential cellular functions (Gil et al., 2004). Accordingly, the mere presence of such genes is completely uncorrelated with resistance, and assigning these genes for AR prior to mutational scanning would be false. We showed a decrease of detection from ~50% to 3.3% in one such cases (see Results). Therefore the use of BLAST alone does not reveal the mutations that confer AR. This is particularly true when dealing with short reads that do not cover the full range of the AR gene. In such cases, even the most stringent BLAST search, with a 100% identity threshold, can align a short metagenomic read to the region of the antibiotic target gene that does not contain a resistance-generating mutation.

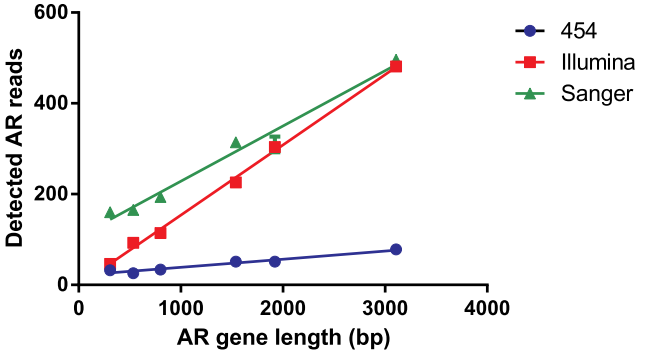

FIG. 6. Effect of three different sequencing platforms on the number of detected AR reads. Simulation experiments were repeated using the error models of three different sequencing platforms (Sanger, Roche-454, and Illumina). Regression lines were significantly different from one another (ANOVA: $p < 0.0001$; Tukey's *post hoc* multiple comparison: 454 vs. Illumina: $p < 0.0001$, 454 vs. Sanger: $p < 0.0001$, Illumina vs. Sanger: $p < 0.0055$).

This limitation may explain the large discrepancy described by Chao and colleagues for results obtained with ARDB versus CARD (Chao et al., 2013). Similarly, this observation most likely accounts for the high rifampin resistance reported by Ma et al. (2014) in an environmental sample. Rifampin resistance is mediated by mutations in *rpoB*, a conserved essential gene, that it is used as a molecular marker (Case et al., 2007). Resqu (http://www.1928diagnostics.com/resdb/), a more recent AR database compared to ARDB (last updated July 3, 2009), has also been used for AR screening of metagenomes (Bengtsson-Palme et al., 2014). Like ARDB, Resqu only includes horizontally transferred antibiotic resistance genes. However, this database is currently not being updated.

Comparing our methods to the former ones showed an improvement in AR detection to cover a wider range of the bacterial resistome. Our proposed workflow detected 2.6 times more AR reads in a plasmid metagenome (Zhang et al., 2011), compared to using ARDB, and detected over 80 times more AR reads in a Swedish lake metagenome (Bengtsson-Palme et al., 2014). This could be explained by (i) the detection of reads pertaining to mutated antibiotic target genes, which were missed in the original studies; (ii) Database difference, since CARD is maintained up to date in contrast to ARDB and is apparently more comprehensive than Resqu.

On the other hand, careful checking of reads for resistance-producing mutations prevented false positive assignment of AR in a metagenome. This specificity was evident as we only detected 3.3% of AR reads in the freshwater fish pond sediment (Ma et al., 2014), in contrast to the original report, suggesting that mutational AR reads constituted ~50% of

detected AR reads (Ma et al., 2014). Of note, Ma and colleagues showed that rifampin resistance comprised more than 35% of total detected resistance, all of which were attributed to *rpoB* gene. On the contrary, careful mutational scanning of *rpoB* reads, for resistance-generating mutations show that it only represents 3.1% of total resistance.

Therefore, compared to existing AR screening methods, our method not only provides a higher coverage of the bacterial resistome since it detects more horizontally transferred AR genes, but also precisely accounts for mutation-generated AR. In other terms, we improved the sensitivity and specificity of AR gene detection and resistome estimation in metagenomes. Our method also takes in consideration the strengths and limitations of different databases, and consequently recommends CARD for AR detection since it includes mutated antibiotic target genes in contrast to ARDB and Resqu databases. Ideally, a well-curated custom database that includes antibiotic target genes, and resistance-generated mutations therein, is equally recommended. Interestingly, a recent article (Xavier et al., 2016)—published while this article was being reviewed—also recommends using CARD for AR annotation, especially for whole genome and metagenomic sequences. This recommendation was based on better annotation not only of gene names, but also at the variant level. Besides, CARD predicted the maximum number of resistance genes in the metagenomic assessment conducted by Xavier and colleagues.

In addition to covering a wider spectrum of resistance genes and mechanisms while screening AR in metagenomes, our method is quantitatively more accurate. Inaccurate estimation of AR gene abundance does not only come from false positive assignments (e.g., unmutated *rpoB* picked up as a resistance gene), but also from miscalculation of gene abundance, usually due to lack of proper normalization.

In this study, we systematically investigated the influence of different intrinsic genome variation and technical differences in metagenomes on AR detection and quantification in metagenomes. The different factors tested were size of bacterial genomes carrying AR genes (assumed to be within a metagenomic sample), AR gene length, metagenome sample size (expressed in number of reads), and metagenomic read length. Additionally, the impact of sequencing platform was investigated through the comparison of three different widely used sequencing platforms. We found that genome size influences the number of detected AR reads. In metagenomics, it is still hard to infer the size of the genome from which an AR gene is derived; nevertheless this finding suggests that plasmid-encoded AR genes are more likely to be detected and represented in metagenomic data than chromosomal AR genes. This genome size effect could falsely increase the relative abundance of plasmid-encoded AR genes.

The second factor investigated was the total number of metagenomic reads. Most studies take this factor into account, and routinely normalize abundance data by dividing the number of detected AR reads by the total number of metagenomic reads. Results are then reported as the percent relative abundance or part per million (ppm) (i.e., read per million reads) (Chao et al., 2013; Chen et al.; 2013, Kristiansson et al.; 2011, Ma et al., 2014; Wang et al., 2013; Yang et al. 2013).

The third factor that we inspected was the length of target AR genes. Bengtsson-Palme et al. (2014) considered this factor and normalized their data by dividing the number of hits for each gene by its gene length; but instead of further normalizing to the total number of reads, they normalized to the number of 16S reads within the same sample (Bengtsson-Palme et al., 2014). Although this is a decent way to normalize for selected genes in a metagenome, especially if the environment is a mixed prokaryotic-eukaryotic ecosystem, it neglects phage contribution to antibiotic resistance (Balcazar, 2014). Future studies should consider double normalization, in which the number of hits is divided by (i) target gene length and (ii) number of reads per metagenomic sample.

In our study, we could also confirm that sequence read length has negligible effect on the number of detected AR reads, as long as the read length is longer than the BLAST alignment cutoff, which should be easily adjusted. Although different sequencing platforms may produce quite different ranges of read lengths, we found that read length is not the major factor behind platform-to-platform discrepancies. Instead, we hypothesized that other factors such as sequencing error rate and nature of error are major players in accurate quantification of AR genes.

Several NGS technologies have been developed; yet Illumnia and Roche-454 have been the most frequently used platforms in recent years (Li et al., 2014), and have been used in a large number of publicly available metagenomes. Although Roche-454 produces longer reads (700 bp versus 300×2 bp in case of Illumina), it has longer run time and lower throughput compared to Illumina (Scholz et al., 2012). With regards to error rates, Roche-454 has the lowest average substitution error rate amongst NGS platforms (Kircher and Kelso, 2010), but has a relatively high indel error rate especially in homopolymer regions (Luo et al., 2012). In contrast, Illumina has high substitution and low indel error rates (Fuellgrabe et al., 2015).

Indeed, we showed that different platforms produce significantly different results for the same set of genes. Sanger sequencing retrieved the highest numbers of correctly detected AR reads, which we believe is a result of the higher accuracy of Sanger output compared to other sequencing platforms (Fuller et al., 2009). However, these results do not take into account Sanger's cloning bias (Liang et al., 2011) nor the low throughput of this method (Fuller et al., 2009). Illumina had higher numbers of detected AR reads than Roche 454, probably because of the lower rate of indel errors in Illumina chromatograms compared to those generated by 454 (Diguistini et al., 2009). Indels greatly affect BLASTX results because they typically introduce frameshifts that may allow an AR read to fail the set cutoffs, thus generating a false negative result.

Based on the factors stated above, we propose an antibiotic resistance abundance index (ARAI) for effective normalization of AR levels, in a similar way to the recently suggested ''phage abundance index'' (Aziz et al., 2015). ARAI takes into account target gene length and total number of reads in a given metagenome, as shown by the equation below.

$$ARAI = \sum \frac{number\ of\ reads\ of\ a\ given\ AR\ gene}{gene\ length \times total\ number\ of\ reads}$$

Conclusions

In conclusion, we improved existing methodologies for screening and quantifying AR genes in any sequence dataset

(typically metagenomic sequence libraries). These modifications take into account the limitations of current methods (e.g., Ma et al., 2014 and Chao et al., 2013). Our method uses the CARD database and carefully considers resistance-producing mutations of target genes in metagenomic reads to identify a wider range of resistance in a given environment, while avoiding false positive results caused by over-recruitment of wild-type antibiotic target genes.

If a study is only interested in acquired or horizontally transferred AR genes, using ARDB, or the more recent Resqu database, is recommended—provided these databases are well maintained and regularly updated. Similarly, CARD website now offers a subset download of the database that excludes genes conferring resistance via mutations, and this would work as well for acquired AR resistome assessment.

We discourage the direct comparison of AR from different platforms and recommend the use of ARAI as a quantitative measure of AR in metagenomes. ARAI relies on double normalization and is thus neither sensitive to AR gene length nor metagenomic sample size. Future efforts should be directed to an efficient algorithm for inter-platform normalization as well as a method to take genome size into account.

Acknowledgments

The authors thank Mr. Mustafa Adel for his assistance in writing the Perl script for mutation detection. This work was funded by an American University in Cairo Faculty (Research) Support Grant to RS. AHAE is funded by a Youssef Jameel PhD Fellowship.

Author Disclosure Statement

The authors declare no competing financial interests.

References

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ. (1990). Basic local alignment search tool. J Mol Biol 215, 403–410.

Aziz RK, Dwivedi B, Akhter S, Breitbart M, and Edwards RA. (2015). Multidimensional metrics for estimating phage abundance, distribution, gene density, and sequence coverage in metagenomes. Frontiers Microbiol 6, 381.

Balcazar JL. (2014). Bacteriophages as vehicles for antibiotic resistance genes in the environment. PLoS Pathogens 10, e1004219.

Balsalobre L, Ferrándiz MJ, Liñares J, Tubau F, and de la Campa AG. (2003). Viridans group streptococci are donors in horizontal transfer of topoisomerase IV genes to Streptococcus pneumoniae. Antimicrob Agents Chemother 47, 2072–2081.

Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, and Larsson DG. (2014). Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Front Microbiol 5, 648.

Bhullar K, Waglechner N, Pawlowski A, et al. (2012). Antibiotic resistance is prevalent in an isolated cave microbiome. PLoS ONE 7, e34953.

Brown MG, and Balkwill DL. (2009). Antibiotic resistance in bacteria isolated from the deep terrestrial subsurface. Microb Ecol 57, 484–493.

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, and Kjelleberg S (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. Appl Environ Microbiol 73, 278–288.

Chao Y, Ma L, Yang Y, et al. (2013). Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. Sci Rep 3, e3550.

Chen B, Yang Y, Liang X, Yu K, Zhang T, and Li X. (2013). Metagenomic profiles of antibiotic resistance genes (ARGs) between human impacted estuary and deep ocean sediments. Environ Sci Technol 47, 12753–12760.

Diguistini S, Liao NY, Platt D, et al. (2009). De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. Genome Biol 10, R94.

Ferrándiz MJ, Ardanuy C, Liñares J, et al. (2005). New mutations and horizontal transfer of rpoB among rifampin-resistant Streptococcus pneumoniae from four Spanish hospitals. Antimicrob Agents Chemother 49, 2237–2245.

Ferrándiz MJ, Fenoll A, Liñares J, and De La Campa AG. (2000). Horizontal transfer of parC and gyrA in fluoroquinolone-resistant clinical isolates of Streptococcus pneumoniae. Antimicrob Agents Chemother 44, 840–847.

Fuellgrabe MW, Herrmann D, Knecht H, et al. (2015). High-throughput, amplicon-based sequencing of the CREBBP gene as a tool to develop a universal platform-independent assay. PLoS One 10, e0129195.

Fuller CW, Middendorf LR, Benner SA, et al. (2009). The challenges of sequencing by synthesis. Nat Biotech 27, 1013–1023.

Gil R, Silva FJ, Peretó J, and Moya A. (2004). Determination of the core of a minimal bacterial gene set. Microbiol Mol Biol Rev 68, 518–537.

Kircher M, and Kelso J. (2010). High-throughput DNA sequencing—Concepts and limitations. BioEssays 32, 524–536.

Kristiansson E, Fick J, Janzon A, et al. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. PLoS One 6, e17038.

Li JZ, Chapman B, Charlebois P, et al. (2014). Comparison of Illumina and 454 Deep sequencing in participants failing raltegravir-based antiretroviral therapy. PLoS One 9, e90485.

Liang B, Luo M, Scott-Herridge J, et al. (2011). A comparison of parallel pyrosequencing and Sanger clone-based sequencing and its impact on the characterization of the genetic diversity of HIV-1. PLoS One 6, e26745.

Liu B, and Pop M. (2009). ARDB—Antibiotic Resistance Genes Database. Nucleic Acids Res 37, D443–D447.

Luo C, Tsementzi D, Kyrpides N, Read T, and Konstantinidis KT. (2012). Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One 7, e30087.

Ma L, Li B, and Zhang T. (2014). Abundant rifampin resistance genes and significant correlations of antibiotic resistance genes and plasmids in various environments revealed by metagenomic analysis. Appl Microbiol Biotechnol 98, 5195–5204.

Martinez JL. (2008). Antibiotics and antibiotic resistance genes in natural environments. Science 321, 365–367.

Martínez JL, and Baquero F. (2014). Emergence and spread of antibiotic resistance: setting a parameter space. Upsala J Med Sci 119, 68–77.

McArthur AG, Waglechner N, Nizam F, et al. (2013). The comprehensive antibiotic resistance database. Antimicrob Agents Chemother 57, 3348–3357.

O'Neill J. (2014). *Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations*, Wellcome Trust, London, UK.

Pletz MWR, McGee L, Van Beneden CA, et al. (2006). Fluoroquinolone resistance in invasive Streptococcus pyogenes isolates due to spontaneous mutation and horizontal gene transfer. Antimicrob Agents Chemother 50, 943–948.

Richter DC, Ott F, Auch AF, Schmid R, and Huson DH. (2008). MetaSim—A sequencing simulator for genomics and metagenomics. PLoS One 3, e3373.

Scholz MB, Lo C-C, and Chain PSG. (2012). Next generation sequencing and bioinformatic bottlenecks: The current state of metagenomic data analysis. Curr Opin Biotechnol 23, 9–15.

Shendure J, and Ji H. (2008). Next-generation DNA sequencing. Nat Biotech 26, 1135–1145.

Toth M, Smith C, Frase H, Mobashery S, and Vakulenko S. (2010). An antibiotic-resistance enzyme from a deep-sea bacterium. J Am Chem Soc 132, 816–823.

Voelkerding KV, Dames SA, and Durtschi JD. (2009). Next-generation sequencing: from basic research to diagnostics. Clin Chem 55, 641–658.

Wang Z, Zhang X-X, Huang K, et al. (2013). Metagenomic profiling of antibiotic resistance genes and mobile genetic elements in a tannery wastewater treatment plant. PLoS One 8, e76079.

Woodford N, and Ellington MJ. (2007). The emergence of antibiotic resistance by mutation. Clin Microbiol Infect 13, 5–18.

Xavier BB, Das AJ, Cochrane G, et al. (2016). Consolidating and exploring antibiotic resistance gene data resources. J Clin Microbiol; In Press.

Yang Y, Li B, Ju F, and Zhang T. (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.

Zhang T, Zhang X-X, and Ye L. (2011). Plasmid metagenome reveals high levels of antibiotic resistance genes and mobile genetic elements in activated sludge. PLoS One 6, e26041.

Address correspondence to:
Dr. Ramy K. Aziz
Department of Microbiology and Immunology
Faculty of Pharmacy
Cairo University
Qasr El-Ainy Street
11562 Cairo
Egypt

E-mail: ramy.aziz@gmail.com

and

Prof. Rania Siam
Department of Biology
The American University in Cairo
AUC Avenue
P.O. Box 74, New Cairo
11835 Cairo
Egypt

E-mail: rsiam@aucegypt.edu

Abbreviations Used

AR = antibiotic resistance
ARAI = antibiotic resistance abundance index
ARDB = Antibiotic Resistance Gene database
CARD = Comprehensive Antibiotic Resistance Database
NGS = next generation sequencing
SRA = Sequence Read Archive