Using metagenomics to investigate human and environmental resistomes

Johan BENGTSSON-PALME[1,2,*], D.G. Joakim LARSSON[1,2], Erik KRISTIANSSON[2,3]

The final typeset version of this paper is available from Journal of Antimicrobial Chemotherapy at the following address:
http://dx.doi.org/10.1093/jac/dkx199

When citing this paper, use the following citation:

Bengtsson-Palme J, Larsson DGJ, Kristiansson E: **Using metagenomics to investigate human and environmental resistomes**. *Journal of Antimicrobial Chemotherapy*, 72, 2690–2703 (2017). doi: 10.1093/jac/dkx199

Using metagenomics to investigate human and environmental resistomes

Johan BENGTSSON-PALME[1,2,*], D.G. Joakim LARSSON[1,2], Erik KRISTIANSSON[2,3]

Affiliations

[1] Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46, Gothenburg, Sweden

[2] Center for Antibiotic Resistance research (CARe) at University of Gothenburg, Box 440, SE-405 30, Gothenburg, Sweden

[3] Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden

* Corresponding author: Johan Bengtsson-Palme, phone: +46 31 342 46 26, e-mail: johan.bengtsson-palme@microbiology.se

Running title: Resistome metagenomics

Abstract

Antibiotic resistance is a global health concern declared by the WHO as one of the largest threats to modern healthcare. In recent years, metagenomic DNA sequencing has started to be applied as a tool to study antibiotic resistance in different environments, including in human microbiota. However, a multitude of methods exists for metagenomic data analysis, and not all methods are suitable for the investigation of resistance genes, particularly if the desired outcome is an assessment of risks to human health. In this review, we outline the current state of methods for sequence handling, mapping to databases of resistance genes, statistical analysis and metagenomic assembly. In addition, we provide an overview of important considerations related to the analysis of resistance genes, and recommend some of the currently used tools and methods that are best equipped to inform research and clinical practice related to antibiotic resistance.

Introduction

Antibiotic resistance is a rapidly growing healthcare problem globally, and has been recognized by the WHO as one of the greatest threats to the fundamental achievements of medicine.[1] A key component for understanding the risks for resistance development is the ability to detect and quantify antibiotic resistance in various settings – the so-called resistome – including in bacterial communities dwelling in and on our bodies. This can, for example, aid in understanding transmission of resistance in the hospital environment. Furthermore, there is substantial evidence that a large portion of the resistance genes circulating among human pathogens today originate from bacteria that thrive in the external environment.[2-4] Contaminated water and food also serve as transmission routes for many bacterial pathogens, particularly fecal bacteria. Moreover, the consequences of antibiotics use in animals have become a great concern for human health as well.[5] Hence a one-health approach including human, animal and environmental aspects of the resistance problem is needed.[6,7] This, in turn, makes it important to also understand risks associated with resistance genes encountered in different external environments, including in animals.[8-10] This paper will overview approaches to study resistomes using high-throughput DNA sequencing and outlines some pitfalls that can influence the evaluation of risks associated with resistance gene findings.

Studying the resistome

Resistance patterns among bacteria have traditionally been studied using culturing on media selecting for resistant colonies. However, when we move away from the most well-studied pathogens, the vast majority of microorganisms cannot be cultivated, at least not by standard methods.[11] This limits the possible scope of this method and thereby veils much of the diversity of species and resistance factors, particularly in environmental communities. For this reason, culture-

independent methods to study resistance genes have been developed, enhanced by rapidly declining costs of DNA sequencing throughout the last decade. A common approach is to randomly fragment the total DNA of a complete community and sequence it by high-throughput sequencing, a procedure referred to as shotgun metagenomics.[12] The resulting DNA fragments can be analyzed using similarity searches to sequence databases, or assembled into longer stretches of DNA, allowing for the reconstruction of complete genes from the relatively short read fragments. However, shotgun metagenomics still requires that the obtained genes, or close variants of them, are present in a reference database to enable assignment of them to a (predicted) resistance phenotype. That said, since sequence data can be stored and re-used later, shotgun metagenomics allows for retrospective analysis of resistance genes identified after the initial study has been completed.[13,14] Shotgun metagenomics has been applied to quantify the abundances of many resistance genes in parallel, for example in environments subjected to pharmaceutical pollution,[15,16] sewage treatment plants,[17-19] sea water,[20] tap water,[21] and the human gut.[13,22] However, in terms of measuring specific gene abundances, metagenomics is less sensitive (i.e. has higher detection limit) than quantitative real-time PCR (qPCR), particularly when only a few million reads are generated per sample. In this respect, Illumina sequencing was a major step forward compared to pyrosequencing, simply due to the lower costs associated with each read. Limited sequencing depth affects the sensitivity to estimate both the abundances and diversity of resistance genes in the sample, which will be discussed in a later section of this paper.

Thousands of antibiotic resistance gene variants are known. A major advantage of shotgun metagenomics compared to qPCR is the ability to investigate all of these variants – including variants not detected by the PCR primers – in a single experiment. Moreover, using the same data, it is also possible to detect changes in taxonomic composition and other functional genes, for example those involved in horizontal gene transfer. This can provide clues about whether the

resistance genes detected have potential to move between bacterial cells or not. Furthermore, through metagenomic assembly it is sometimes possible to uncover co-resistance patterns, or even completely novel resistance plasmids.[15,19]

Obtaining sequence data from microbial communities

As a first step of any metagenomics analysis, DNA must be extracted from the community. This is usually done using standard DNA extraction kits. However, as most microbial communities comprise a large diversity of different bacteria and also may contain contaminants of different kinds, this process is not always straightforward. It is important to understand that extraction protocols can bias gene frequencies, as not all bacterial species are affected equally by the reagents used. Bias has been shown to result from differences between DNA extraction kits,[23,24] storage of samples,[24,25] DNA amplification kits,[26] as well as due to biological variation of, for example, GC-content.[27] All these factors contribute noise to the samples already before the sequencing takes place.[28] However, different sequencing techniques also produce different results, partially because of differences in sequence length for each fragment, but also due to the different methodologies used to determine the nucleotides.[29] Before any other analyses are performed, it is advisable to filter the sequence data with respect to sequencing adapters and low-quality reads.[30] Since paired-end sequencing is becoming the norm, filtering software that can consider both reads in a pair simultaneously is desirable. This can be done using a variety of software, for example Trim Galore!, [31] Trimmomatic,[32] Sickle,[33] or Prinseq.[34]

Detecting and quantifying resistance genes in metagenomes

Gaining insights into the resistance gene content of a microbial community from sequence data requires the ability to detect resistance genes among sequence fragments derived from a multitude

111 of different genes. This is achieved through similarity searches, employing the principle that genes
112 sharing homology often perform similar functions. This principle is at the heart of bioinformatic
113 methods, but depending on the questions asked, its usefulness differs. Often, changes of only a
114 few amino acid residues in a protein can alter its substrate preferences,[35,36] binding sites[37,38] or
115 overall functions.[39,40] Therefore, the validity of the assumption that a read matching to a protein in
116 a reference database comes from a gene encoding a protein with the same function is dependent
117 on how similar the read is to the reference sequence.[41] This means that the choice of method for
118 assigning function to metagenomic reads depends on which stringency one aims for. In the case
119 of mobilized genes, their sequences show limited variation once they have appeared on mobile
120 genetic elements (MGEs).[42] Because of the inherent dependency on sequence similarity, selecting
121 an appropriate sequence identity cutoff for calling a matching read a resistance gene becomes
122 crucial.[8] At the same time, reads come with a certain degree of sequencing errors, and there might
123 be slight differences between resistance genes that do have the same function. Therefore, one
124 wants to allow to a certain degree of mismatches between the read and the reference sequence –
125 the question is: how large can this difference be if stringency is to be maintained? The answer to
126 that question depends on how similar resistance genes known to carry out the same function are.
127 However, the percent identity of functionally verified resistance genes within the same group varies
128 substantially (Figure 1). In Resqu[43] – a database containing only resistance genes with
129 experimentally verified function, the average sequence identity between sequences associated with
130 the same gene name and function differs between 68% and completely identical (median 97.9%),
131 while the median lowest identity between two sequences with the same gene name is 95.3%, but
132 can be as low as 52.8% (the *vanSG* vancomycin resistance gene). However, applying a universal
133 cutoff of 50% sequence identity would produce an immense number of false positive hits. Using
134 the IMP beta-lactamase as an example, performing a BLAST search[44] against the NCBI protein

database[45] with the IMP sequences as queries yields more than 200 matches at a 50% identity cutoff (requiring 30 matching amino acids, corresponding to the length of a typical Illumina read). These matches include several major facilitator superfamily (MFS) transporters and sulfurtransferases, indicating that this cutoff would not be feasible.

Indeed, there is no foolproof approach to make sure that a read comes from a functional resistance gene. Even if 100% identical to a resistance gene, the read only represents a part of the gene sequence, and the gene the read is derived from may, for example, be truncated and thus non-functional. However, as seen in the example with IMP, it is important that the cutoffs are not set too low to retain stringency. Thus, requiring sequence identity of 80-95% is probably warranted. Furthermore, the larger the datasets grow, the more computing resources will be required to process them. Read mapping of short read data from high-throughput sequencing allowing for a large number of mismatches is typically computationally much more expensive than searching for high-identity matches. Thus, the choice of cutoff value becomes a tradeoff between speed, sensitivity and stringency. For example, employing a cutoff of two amino acid mismatches per read will correspond to a percent identity of 90-94%, depending on the read length. Many software tools exist to efficiently map reads to protein reference databases, employing different computational approaches, including Vmatch,[46] Usearch[47] and Diamond.[48]

Databases for resistance genes

The choice of reference databases also has important implications for the quality of the information derived. Since annotation based on bioinformatic analysis of sequence similarity never will be more accurate than that of the reference sequences, selecting a reference database with high-quality annotation is crucial.[41] Simply put, if the database only contains resistance genes against beta-lactams, you will likely grossly underestimate the number of resistance genes present. On the other

158 hand, if the database contains genes incorrectly predicted to have resistance functions, the
159 abundance and diversity of resistance genes in the sample will be overestimated. A number of
160 databases containing antibiotic resistance gene information exist. An often used resource,
161 particularly in the early papers using metagenomics to investigate antibiotic resistance, is the
162 Antibiotic Resistance Genes Database (ARDB), established in 2008.[49] However, a few problems
163 exist with ARDB. Most prominently, its last update was in July 2009, meaning that any resistance
164 gene discovered after that date is not included in the database (this includes e.g. the clinically very
165 important carbapenemase NDM-1[50] and the newly discovered *mcr-1*[51]). In addition, ARDB does
166 not make any difference between resistance genes with a confirmed resistance function and those
167 predicted to confer resistance based on homology. Thus, the database may very well contain
168 sequences that in fact are not functional resistance genes. ARDB has subsequently been structured
169 by resistance types and had some obviously erroneous sequences removed,[17] and this version of
170 the database remains in use.[52] However, the basic problems of the database being outdated and
171 that the majority of sequences do not have their functionality demonstrated prevail also in this
172 version. The developers of ARDB instead recommend the use of the Comprehensive Antibiotic
173 Resistance Database (CARD).[53] This database is still in active curation and is possibly the most
174 comprehensive resource for antibiotic resistance gene information available. However, although
175 CARD is based on thorough curation, it does not clearly separate experimentally verified and
176 predicted entries. Furthermore, it is unclear if the genes in the database have been found on MGEs
177 or only have been detected on chromosomes. That said, the use of a single reference sequence for
178 every resistance gene in CARD increases the likelihood that each sequence has been confirmed to
179 confer resistance in at least some species. Similar problems also haunt the ARG-ANNOT
180 database,[54] although to a much larger extent. The ARG-ANNOT database employs what they refer
181 to as "relaxed search criteria" to identify resistance genes, which in reality means that the database

182 contains a multitude of sequences with poor annotation information, and that many entries are
183 unlikely to be functional resistance genes. This limits the value of ARG-ANNOT for identifying
184 true resistance genes. A more stringent approach to this has been taken by the ResFinder[55] and
185 Resqu[43] databases. Both these databases only contain sequences of acquired antibiotic resistance
186 genes present on MGEs. However, a drawback associated with Resqu is that it has not been
187 updated since 2013, while ResFinder remains actively curated.

188 ## How the database content affects results

189 Depending on the database used, reported resistance gene abundances may differ, despite that the
190 same bioinformatics protocols are applied. For example, ARDB, CARD and Resqu report radically
191 different numbers of resistance genes in the human gut and sediment from a Swedish lake (Figure
192 2; data from Bengtsson-Palme *et al.* 2014 and 2015[16,56]). Resqu consistently reports the lowest
193 numbers, likely since it only contains resistance genes with a verified resistance function that have
194 been shown to be present on MGEs and thus excludes many generic efflux pumps that may confer
195 low-level antibiotic resistance. From a risk perspective, mobile resistance genes are probably the
196 most relevant to detect and quantify.[8,10] Furthermore, many multidrug efflux pumps are relatively
197 well conserved between variants with and without capacity to export antibiotics.[8] Using the full
198 CARD database (2015 version) consistently reports resistance gene counts two to three times
199 higher than ARDB. In a newer version of CARD,[53] chromosomal genes where point mutations
200 provide resistance have been removed, and this version generates roughly the same results as
201 ARDB (although not for the lake sediments). Genes containing such point mutations may indeed
202 provide resistance, but are rarely transferrable between bacteria and are – importantly – very similar
203 to the susceptible variants of the target genes. The latter means that even reads stemming from
204 susceptible ("wild-type") bacteria in a metagenome would map to these "resistance genes",

particularly if, e.g., a 90% identity threshold were used. Diluting the database with such genes means that the total resistance gene content will undoubtedly be overestimated, as many of these target genes are ubiquitously occurring essential genes, highly conserved between bacterial species. For example, the *rpoB* gene (the target gene of rifampicin; mutated variants are present in the full CARD database) is present in a single copy in most bacterial species[57] and has thus been proposed as a possible per-genome normalization gene for metagenomics.[58] The presence of around one such "resistance gene" per 16S rRNA in the Swedish lake sediment, as reported when using the full 2015 version of CARD (Figure 2) therefore seems reasonable. However, the vast majority of the reads associated with these "resistance genes" actually derive from antibiotic-sensitive variants of essential target genes.

It is important to realize that this is not a problem related to the CARD database *per se*. The database website clearly states that target genes are present among its sequences, and also provides a separate dataset with the target genes removed for use in metagenomic studies. Recently, CARD was also updated to fully separate target sequences and functional resistance genes in different files.[53] Still, if care is not taken in examining the content of the database used, this may lead to partially misleading conclusions, which may explain the surprising results of some studies.[59]

A similar problem is the use of general annotation pipelines, such as the commonly used MG-RAST,[60] that are not curated with regards to antibiotic resistance. The use of MG-RAST to annotate resistance genes has led to some peculiar reports suggesting that almost one in 25 genes found in human feces would confer antibiotic resistance.[61] The non-stringent identity cutoffs used by default in MG-RAST are likely to be a major cause of these results. Similar use of low identity thresholds in other studies has led to unexpectedly high estimates of resistance gene abundances in other human feces samples.[62] This emphasizes the importance of accounting for technical factors

that could explain unexpected results in metagenomic studies. Overall, there is a clear need for improved stringency with regards to database usage and parameter choices in metagenomics studies aiming to quantify resistance gene abundances.

Unsolved statistical problems for metagenomics

Once gene counts have been established, the next aim is usually to identify differences in resistance gene abundances between samples. Although this sounds straightforward, a number of technical obstacles remain.[28] The most fundamental problem affecting the statistics of metagenomic data is that the data is high dimensional in the sense that there are generally many more observed genes than biological replicates. Furthermore, the variation between samples in the same group can be fairly large, meaning that higher numbers of replicates are required to detect statistically significant differences.[63] However, because sequencing is relatively expensive, a tradeoff exists between obtaining sufficient sequencing depth for quantification of genes in each individual sample and the number of replicate samples sequenced. Finally, since metagenomics generates counts, the resulting data is discrete, and many existing statistical tests assume continuous, normally distributed data. The last few years have seen tremendous development of statistical methods for metagenomic analysis,[64] somewhat reminiscent of the early method advances in microarray analysis.[65] However, many of those methods provide a descriptive picture of the studied community rather than highlighting statistically significant differences.[66] Interestingly, it took about ten years of microarray usage for statistical methods to "catch up" and become standardized,[67] and it is reasonable to assume that shotgun metagenomics faces a similar development towards robust standardization within the next few years.

Normalization of data to make samples comparable

Another problem with metagenomic sequence data is that the generated libraries may be of vastly different size, which influences the number of counts from different samples. Furthermore, the composition of the samples may be different, and technical factors can bias the sample processing. To make libraries from different samples comparable, normalization is applied. However, depending on the research question, different means of normalization can be appropriate. If one is merely interested in compensating for the different size of the sequence libraries, simply dividing each count by the total number of reads of each library generating, for example, a count-per-million value may be sufficient. However, when investigating antibiotic resistance it is often more relevant to determine the counts relative to the bacterial fraction of the sample (trying to exclude contributions from e.g. eukaryotes and viruses). For this purpose, a bacterial marker gene is often used for normalization, most commonly the SSU 16S rRNA, yielding gene counts per 16S rRNA. However, although the rRNA genes are well studied and often applied for normalization purposes, they can occur in multiple copies within the same genome,[68,69] and thus other, single-copy, bacterial marker genes have been suggested for normalization,[70,71] such as the ribosomal protein *rpoB* gene.[57,58] That said, since these normalization methods have not yet gained traction, and because of the legacy of qPCR studies, the 16S rRNA remains the most common normalization gene for studies of bacterial communities. One can imagine other relevant normalization strategies, such as comparing each gene count to the total content of resistance genes. Importantly, the choice of normalization method should be based upon the questions asked, and how these questions are best answered. It is also important to consider whether there are variations between samples that will *not* be compensated for under the normalization method chosen. Such variation may for example be the result of differing 16S rRNA copy numbers, or that not all variants of the marker gene of choice are detected by the methods used, which is a common problem, particularly when read lengths are short.[72] There are also completely different approaches to normalization used in

RNAseq, based on minimizing the overall fold-change between experiments, thereby attempting to reduce technical noise.[73] Similar thoughts have been carried over into recent metagenomic analysis packages,[74] although the task of identifying a subset of genes that can be assumed to be stable between samples is not as straightforward in data from communities comprised of mixtures of species.

An additional factor that also may influence gene abundance estimates based on reads mapped to a reference database, is the length of the reference genes. If this is not compensated for, longer genes may recruit more reads simply by chance. This effect is not relevant to compensate for if one only compares data between samples, but if the abundance levels between genes are compared, taking gene length into account becomes necessary. This type of normalization makes sense, but whether or not it is meaningful to compensate for it in real situations is debated.[75,76] Some authors have suggested that compensating for gene lengths may even be detrimental to differential analyses of RNAseq data,[77] although if the same argument is valid also for metagenomic data is unclear.

Data transformation approaches

Currently, the statistics for handling metagenomic count data are centered on three fundamentally different approaches: standard tests on transformed counts, tests assuming distributions that account for the features of count data, and non-parametric tests. Data transformations are often used to change the distribution of the data so that it better fits the normal assumptions of standard tests, such as t-tests and ANOVA. For count data, the variance is always dependent on the mean, and proper data transformations remove this relationship. Such variance-stabilizing transforms include the square-root transform and various logarithm transforms. Note that logarithm transforms "penalize" very large values harder than the square-root transform, and thus analysis of logarithm-transformed data is less influenced by the most abundant genes. Transformation

297 methods allow the use of standard microarray analysis tools on count data, as implemented in e.g.
298 the Voom package, which estimates and weights the mean-variance relationships of each
299 observation and subsequently analyze the transformed counts using Limma.[78,79] One problem that
300 becomes apparent when applying a logarithm transform to metagenomic count data is the large
301 number of zeros present. Zeros lead to two problems. The first is practical – zeros cannot be
302 logarithm transformed, and the second is that a zero can either represent that a gene is not present
303 at all, or that it is so rare that the sequencing depth was not sufficient to detect it. The
304 transformation problem can be solved by adding a pseudocount to all observations in the dataset,
305 usually simply a count of one. However, the pseudocounts may influence effect sizes (and thus
306 statistical significances), particularly when overall gene counts are low, which have led some authors
307 to advise against the use of transformation methods for count data in those cases.[80] The latter
308 problem associated with zeros is harder to deal with, and is particularly troublesome when
309 estimating the richness and diversity of taxa or genes, a problem we will return to later. Efforts to
310 handle zero-inflation have been made in, for example, the metagenomeSeq package, which uses a
311 zero-inflated Gaussian model to correct for undersampling-related bias.[81]

312 *Non-parametric and count-adapted tests*

313 As an alternative to data transformation, statistical tests that do not make as specific assumptions
314 on the distribution of the data can be used. These are referred to as non-parametric tests,[82] and
315 include e.g. tests based on the ranks of the observation rather than their actual values. These
316 methods – for better and worse – do not depend on distributional assumptions and are therefore
317 more robust to outliers in the data. Other non-parametric tests include permutation tests that
318 resample the data instead of assuming that it follows any particular distribution.[83-85] Finally, there
319 are also statistical tests designed to better handle count data, usually based on assumptions of

Poisson or negative binomial distributed data, such as ShotgunFunctionalizeR,[86] which allows fitting of generalized linear models to metagenomic count data. Such models are also implemented in the RNAseq analysis packages edgeR[87] and DESeq,[88] which couple the variance and mean either naïvely (edgeR) or by determining the optimal coupling for each individual gene (DESeq). Both these tools are developed for RNAseq data, and although this technique generates similar count data, their assumptions may not be entirely valid for metagenomic analysis. A recent evaluation of different statistical approaches to identify significantly differing genes between metagenomes concluded that the number of replicates, the effect sizes and the gene abundances greatly affected the outcomes of each method, and that no single method is suitable for all metagenomic datasets and questions.[64] That said, the methods based on Poisson or negative binomial distributions used for RNAseq overall performed better, particularly with small group sizes, with DESeq and overdispersed Poisson linear models coming out on top. Surprisingly, ordinary square-root transformed t-tests performed relatively robustly also at small group sizes. However, the evaluation also showed that several methods (non-transformed t-tests, Fisher's exact test and the binomial test) perform poorly and should be avoided. Furthermore, non-parametric methods also perform subpar and should in most cases be replaced by methods based on transformation or appropriate modeling of counts.

Correction for multiple testing

Regardless of which method that is used to determine which genes that are significantly enriched in a group of samples, one p-value will be obtained for each gene tested. This means that with a large reference database, hundreds or thousands of tests will be performed. Since the p-value represents the probability of obtaining a particular result by chance, under the null hypothesis given certain model assumptions,[89] performing multiple tests will increase the probability of obtaining

false positive observations tremendously.[90] Therefore, large experiments with many measurements, such as using metagenomics to detect resistance genes, require some form of correction for multiple testing. One way of doing this is to simply multiply each p-value with the number of tests performed (i.e. the number of genes investigated), referred to as the Bonferroni correction.[91,92] However, in many explorative studies the Bonferroni correction is regarded to be too conservative, and therefore more relaxed approaches, such as the Benjamini-Hochberg false discovery rate, are commonly used in large-scale experiments to control the number of false positive observations.[93]

Measuring abundance and diversity of resistance genes

Not only the abundance of resistance genes in certain settings may be of importance for determining risks, but also the diversity of such genes found. However, it is debated how to best establish the diversity of resistance genes, for example whether or not the relative abundances of different genes should be taken into account. Similar difficulties with estimating species richness in different communities have haunted ecology for more than half a century.[94] A plethora of diversity indices designed for community ecology exist and are currently in use, each with its own advantages and shortcomings. The most basic such measurement would be to simply count the number of different resistance gene types encountered, establishing what is called the richness of the sample. This, however, is not without problems.[95] First of all, the richness is intimately connected with sampling effort (in the metagenomics case the size of the sequencing library). One could account for this by normalizing the abundances of each gene in all samples to the size of each sample, thereby making them comparable, and then only count entries with a normalized abundance corresponding to finding at least one copy of the gene in the smallest sample. However, while this reduces the dependency on library size, it instead introduces a bias towards the most abundant entities. For this reason, rarefaction methods, in which the number of different resistance

gene types encountered are plotted against the sampling effort required to detect them, have instead been suggested to deal with this problem in community ecology.[96,97]

Furthermore, the studied sample of resistance genes only comprises a subset of the total resistance gene types likely present in a community. Thus, the true richness of the sample is unknown, and information on the abundances associated with lowly abundant genes is either poorly estimated or lacking. This means that it might be informative to account for the unseen resistance genes in some way. Measures for extrapolating richness could be borrowed from ecology, for example the Chao1[98] and ACE[99] estimators. In addition, resampling methods have been suggested to estimate the true richness of samples.[100] However, these estimators have been shown to fluctuate substantially with changing sample size.[101] As ecologists and statisticians still struggle with the problem of estimating the number of rare species in a community, we can conclude that accounting for those is hard, and that for the time being we are probably best off comparing the richness of detected resistance genes in different samples and hope that those numbers reflect the true richness reasonably well. In addition, the methods for finding resistance genes using shotgun metagenomics only allow detection of known genes present in a reference database. The yet undiscovered resistance genes, of which there seem to be a multitude both in the environment and in the human microbiome,[2,102-107] and which avoid detection regardless of being abundant or rare, are incredibly hard to account for using richness estimators. Once again, one could assume that a large diversity of known resistance genes implies a broad range of unknown resistance factors as well, but to which degree this is true remains unknown.

What are the benefits of assembling metagenomes?

Depending on where an antibiotic resistance gene is located, its ability to confer resistance, as well as its potency to spread to other bacteria, varies considerably.[8,10,108] A central limitation of using

389 short-read metagenomic data to study antibiotic resistance is thus that it is not possible to associate
390 a read mapped to an identified resistance gene to a specific species or strain with certainty,
391 hampering the evaluation of risks associated with resistance gene findings. In addition, different
392 promoter regions may enhance or reduce the expression of a gene, and interactions with other
393 gene products may influence the resistance function of the gene. Furthermore, a gene that is
394 present on a plasmid or other mobile genetic element is vastly more likely to spread between
395 bacteria than one firmly located on the bacterial chromosome.[8,109] Also, the compatibility of a
396 mobile resistance gene with its host influences whether the gene encodes an efficient resistance
397 mechanism in that specific context. Finally, genes mobilized by integrases or transposases may have
398 modified 3' and/or 5' ends, which may also alter their expression in the new context. The latter is
399 thought to have contributed to the efficiency of the NDM-1 carbapenemase gene in a variety of
400 hosts.[110,111] Because of the complex interplay between the host, its resistance genes and their genetic
401 environment, it is important to consider the genetic context around resistance genes, as well as the
402 taxonomy of their carriers. To fully understand the genetic context of resistance genes, functional
403 selection of resistant strains or resistance plasmids followed by analysis of their complete sequences
404 is in principle required.[112-116] This is, however, a rather labor-intensive approach, and it is also
405 restricted to isolates that can be cultured and/or plasmids that can be captured by cultivable
406 bacteria. Another approach to gain insights into the contexts of resistance genes is through the use
407 of metagenomic shotgun sequencing followed by computational assembly of the reads.[16,52] While
408 this method is limited to resistance regions abundant in the sample, due to the requirement of large
409 sequencing depth, it circumvents the need for cultivation and phenotypic resistance selection.

The current state of assemblers for metagenomic sequence data

Early metagenomics projects, which generated longer and fewer reads, generally utilized the same assemblers as genome projects, such as the Celera assembler,[117] Newbler[118] or MIRA[119]. The assemblers used on long-read data are most often based on the overlap-layout-consensus algorithm,[120] which works well on smaller data sets, but quickly becomes vastly time and memory consuming, as its complexity scales roughly quadratic with the number of reads due to the all-to-all comparisons of reads required.[121,122] For the massive amount of short-reads generated by e.g. the Illumina platform, such algorithms are unsuitable because of the dramatically increased complexity. The first widely used assemblers for short-read data – e.g. SSAKE[123] – solved this by greedy approaches, which are less computationally expensive, but produce sub-optimal solutions to the assembly problem.[122] Instead, methods that reduce the complexity of the assembly graph by converting it into a de Brujin graph[124,125] quickly gained traction and remain the most used assembly methods for Illumina data. The de Brujin graph is less complex to build and traverse than the overlap-layout-consensus graph, making the assembly problem easier to solve.[126] This has resulted in a plethora of assembly algorithms based on de Brujin graphs, of which some popular examples are Velvet,[127] ABySS[128] and SOAPdenovo.[129] With increasing popularity of metagenomics, specialized software for metagenomic *de novo* assembly has also been developed. These programs are often modified versions of genomic assemblers, such as Meta-Velvet,[130] Meta-IDBA,[131] metaSPAdes[132] and Ray Meta.[133] Although these adaptions in theory can improve assembly quality, the discernible difference between assemblies produced by e.g. Velvet and Meta-Velvet is minute,[134] which is also consistent with our own observations (Bengtsson-Palme J., unpublished data). Benchmarking of different assemblers on data where the true result is known has shown that the N50 metric, which is often used to assess assembly quality, is generally useless since an assembler that merges too many reads together will get high N50 values (generally interpreted as good), but

434 does so at the cost of generating chimeric contigs.[135,136] This problem may be relatively minor for
435 single genome assembly, since the possibilities for manual inspection and correction are larger.
436 However, for metagenomic samples where many species are mixed, assessing which contigs that
437 may be chimeric is almost impossible, which makes the numbers of errors a central metric in
438 selecting an assembler software. In this context, it is worrying to note that particularly
439 SOAPdenovo, but also Velvet, produce relatively high number of errors compared to other
440 assemblers,[135] such as ABySS and ALLPATHS-LG.[137] However, ALLPATHS-LG requires a very
441 specific set of sequence libraries to operate, making it unsuitable as a general-purpose assembly
442 tool. Furthermore, other comparisons indicate that ABySS and Velvet perform similarly (and
443 produce comparatively few errors) on short-read data from bacterial genomes.[138]

444 Aside of avoiding assembly errors, another important consideration as metagenomic datasets
445 continue to grow is the issue of scalability. An efficient assembler must not only be able to deliver
446 mostly correct contigs, but must also do so within a reasonable timeframe and within attainable
447 memory limits. Even though metagenomic assembly generally is carried out on large computer
448 clusters with hundreds of gigabytes of RAM, assembly of some metagenomic datasets is still not
449 feasible with current methods.[139,140] This leads to compromises between the most accurate and most
450 efficient assembly algorithms. One key parameter of large-scale assembly is that the software
451 should be scalable across multiple processor cores and nodes (individual machines) in a computer
452 cluster. Two assemblers have struck a reasonable balance between accuracy and scalability for
453 metagenomic assembly: ABySS and Ray. Both are highly scalable, while still producing results
454 comparable to those of Velvet.[56,138] However, for really large metagenomes neither of these
455 assemblers are sufficiently memory efficient, which has spurred the development of alternative
456 assembly strategies. For example, reads can be binned based on k-mer content prior to assembly,
457 reducing the need to assemble all the reads at once.[141] Furthermore, reads from low-coverage

regions can be filtered out prior to assembly,[142,143] or reads from high coverage regions can be set aside, a strategy referred to as digital normalization.[140] Finally, merging of sub-samples of reads assembled individually has been proposed as a possible, albeit sub-optimal, assembly strategy.[144] A completely different approach to metagenomic assembly is to target only regions of interest in the metagenome, which also reduces the complexity of assembly. Such approaches have been implemented in assemblers such as TriMetAss,[16] and the SAT-Assembler.[145]

Assembly of genes existing in multiple genomic contexts

The greatest obstacle to enable assessment of the context of mobile resistance genes identified in metagenomic data is the nature of the resistance genes themselves. We are often interested in investigating whether a resistance gene is present on a MGE or not, as this property is strongly related to the relative risk associated with the gene.[8,10] However, resistance genes present on MGEs are often better conserved between species (since they can be transferred directly) than chromosomal resistance genes. In addition, if they are mobilized in integrative elements they can exist in multiple similar, but not identical, genetic contexts.[16,146,147] This presents a problem for assembly software working with short reads. Many times, there can be multiple possible branches out from a highly conserved part of a resistance gene or resistance gene cassette (Figure S1a). Almost all assembler software handle this by splitting the contigs at the branching points, although some use coverage information or other external data (such as read-pair information) to avoid unnecessary splits and handle splits more intelligently. Regardless, the result is a fragmented assembly that does not contain much information about the genetic context of the resistance gene of interest. In the example presented in Figure S1a, no contextual information is retrieved for resistance gene A, since it ends up on a single contig without any flanking regions. This not only obscures the information about whether a resistance gene is transferrable between bacteria, but

also severely limits our ability to detect resistance genes that are co-localized. In addition, closely related resistance genes are often not identical across their entire length, but rather contain identical regions. In those cases, the individual resistance genes may also be split up on multiple shorter contigs, further complicating the assembly (Figure S1b).

The problems related to multiple contexts usually get worse the more common a resistance gene is, since common resistance genes are more likely to be detected in multiple contexts. In addition to these examples where true biological variation causes assembly problems, sequencing errors may also break the assembly up in a similar fashion as in Figure S1b, although assemblers are generally better at handling such problems than true biological variation. Similarly to resistance genes existing in multiple contexts, integrases and transposases are prone to the same types of problems, and break assemblies up in an analogous way, resulting in contigs containing, e.g., one or two resistance genes and a (sometimes partial) ISCR or integrase sequence.

Clinical resistome analysis using metagenomics

A variety of studies have investigated the abundance and diversity of resistance genes in the human microbiome, revealing overall trends related to body compartments,[14] antibiotics usage,[13,148] early development in infants,[149] and travel.[56] These studies have together contributed a baseline knowledge of how the human resistome is composed and how it varies across different countries. As a broad-encompassing research tool to characterize the overall resistance gene composition of the human microbiota, metagenomic sequencing has proven to provide reliable and reproducible results. However, implementation of metagenomic approaches for clinical purposes is not without problems. First of all, for most sample types from humans except feces, the vast majority of the reads will be derived from the human genome, unless some depletion strategy for human material is employed. Furthermore, even in feces it has been shown to be hard to detect clinically important

pathogens and resistance genes that could be isolated through selective culturing.[56] That said, with appropriate purification protocols, it is possible to reliably detect resistant pathogens in e.g. urine samples using metagenomic sequencing.[150,151] The use of sequencing technology for this purpose may not yet be sufficiently fast and reliable for clinical diagnostics, but is likely to mature in the very near future.[152,153] It is at present unclear if the benefits of shotgun metagenomics justify the costs of implementing it as a clinical diagnostic tool,[5] particularly as PCR and culturing-based approaches remain vastly more sensitive.[56,154] However, metagenomic approaches could be used in epidemiology to track transmission, although this would at present be a costly practice. However, sequence data can be re-investigated when novel resistance factors are discovered,[155] which enables probing of if a new resistance gene is already widely spread in the human microbiome.

The influence of environmental fecal contamination

Detecting relatively larger numbers of antibiotic resistance genes in a metagenome than expected in the studied environment is often interpreted as a product of selection for antibiotic resistance. However, this is not necessarily the case. In the environment, the abundance of resistance genes often is tied to the relative proportion of fecal bacteria (Figure 3; data from Pal *et al.*[14]). This makes it difficult to infer whether an enrichment of resistance genes in a particular sample is due to selection for the resistance factor, or merely the by-product of contamination with feces. Thus the detection of resistance gene enrichments in certain sample types will not tell much about selection unless placed into a taxonomic context, or if the levels detected are substantially above those in human feces, which would also indicate selection for resistance. Because of the relationship between resistance genes and fecal pollution, it becomes important to estimate the proportion of bacteria derived from feces in different environments. Since metagenomics enables detection of a wide diversity of taxa, it has been proposed to use the bacteria present in the human gut

microbiome genome catalog[156] as reference for tracking human feces contamination in the environment.[157] Still, this approach will only provide an upper bound for the human-associated bacterial content, as many of the species present in that genome catalog can exist also in the gut microbiome of other species, or in the external environment. Finding appropriate fecal markers remains an unsolved problem for using metagenomics in environmental resistance gene research, and a perfect solution may not even exist.

Conclusions

As should be evident from this overview, a multitude of approaches exist for resistance gene quantification and investigation in metagenomes. While the choice of methods should ultimately be made with respect to the questions asked and the samples investigated, some methods are clearly better suited for resistance gene studies than others. A suggested workflow for resistance gene analysis with the currently best-suited tools is given in Figure 4 We would like particularly to emphasize the importance of choosing appropriate normalization strategies, and sufficiently stringent sequence identity cutoffs to avoid over-classification of resistance genes. Furthermore, the choice of database is also of utter importance to avoid misleading conclusions. Finally, appropriate statistical methodologies for metagenomic analysis is just starting to emerge[64] and we would like to encourage the reader to stay updated on those to make the most possible use of their metagenomic sequencing data. Nevertheless, the need for proper replication of samples will not disappear by the introduction of more sophisticated statistical methods. Although still costly, metagenomic sequencing is on the verge of finding clinical use in specific diagnostics situations, such as in rapid characterization of urine and blood samples. Most likely, progress in sequencing technology will facilitate this development by driving prices down further, but will also yield longer reads and reads with lower error rates. This would be beneficial to get substantial insights into the

genetic contexts of resistance genes, which is fundamental to differentiate risks associated with resistance gene findings in different cohorts and environments.[8,158]

Funding

This work was supported by the Swedish Research Council for Environment, Agriculture and Spatial Planning (FORMAS; grant numbers 2012-86 to DGJL and 2016-786 to JBP); the Swedish Research Council (VR); the Life Science Area of Advance at Chalmers University of Technology; MISTRA (grant number 2004-147); the Wallenberg Foundation; the Adlerbertska Research Foundation; and the Center for Antibiotic Resistance Research (CARe) at University of Gothenburg.

Transparency declarations

None to declare.

References

1. WHO. *Antimicrobial resistance: global report on surveillance 2014*. Geneva, Switzerland: World Health Organization; 2014. http://www.who.int/drugresistance/documents/surveillancereport/en/

2. Allen HK, Moe LA, Rodbumrer J, *et al.* Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J* 2009; **3**: 243–51.

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

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

5. 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.

6. World Health Organization. *Global Action Plan on Antimicrobial Resistance*. Geneva, Switzerland: WHO; 2015. http://www.who.int/antimicrobial-resistance/publications/global-action-plan/en/

7. Collignon P. The importance of a One Health approach to preventing the development and

spread of antibiotic resistance. *Curr Top Microbiol Immunol* 2013; **366**: 19–36.

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

9. Berendonk TU, Manaia CM, Merlin C, *et al.* Tackling antibiotic resistance: the environmental framework. *Nat Rev Microbiol* 2015; **13**: 310–7.

10. Bengtsson-Palme J, Larsson DGJ. Antibiotic resistance genes in the environment: prioritizing risks. *Nat Rev Microbiol* 2015; **13**: 396.

11. Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* 1995; **59**: 143–69.

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

13. Forslund K, Sunagawa S, Kultima JR, *et al.* Country-specific antibiotic use practices impact the human gut resistome. *Genome Res* 2013; **23**: 1163–9.

14. Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ. The structure and diversity of human, animal and environmental resistomes. *Microbiome* 2016; **4**: 54.

15. Kristiansson E, Fick J, Janzon A, *et al.* Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS ONE* 2011; **6**: e17038.

16. Bengtsson-Palme J, Boulund F, Fick J, *et al.* Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. *Front Microbiol* 2014; **5**: 648.

17. Yang Y, Li B, Ju F, *et al.* Exploring variation of antibiotic resistance genes in activated sludge over a four-year period through a metagenomic approach. *Environ Sci Technol* 2013; **47**: 10197–205.

18. Yang Y, Li B, Zou S, *et al.* Fate of antibiotic resistance genes in sewage treatment plant revealed by metagenomic approach. *Water Res* 2014; **62**: 97–106.

19. Bengtsson-Palme J, Hammarén R, Pal C, *et al.* Elucidating selection processes for antibiotic resistance in sewage treatment plants using metagenomics. *Sci Total Environ* 2016; **572**: 697–712.

20. Port JA, Wallace JC, Griffith WC, *et al.* Metagenomic profiling of microbial composition and antibiotic resistance determinants in Puget Sound. *PLoS ONE* 2012; **7**: e48000.

21. Shi P, Jia S, Zhang X-X, *et al.* Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. *Water Res* 2013; **47**: 111–20.

22. Hu Y, Yang X, Qin J, *et al.* Metagenome-wide analysis of antibiotic resistance genes in a large cohort of human gut microbiota. *Nat Commun* 2013; **4**: 2151.

609 23. Knauth S, Schmidt H, Tippkötter R. Comparison of commercial kits for the extraction of
610 DNA from paddy soils. *Lett Appl Microbiol* 2013; **56**: 222–8.

611 24. McCarthy A, Chiang E, Schmidt ML, *et al.* RNA preservation agents and nucleic acid
612 extraction method bias perceived bacterial community composition. *PLoS ONE* 2015; **10**:
613 e0121659.

614 25. Choo JM, Leong LE, Rogers GB. Sample storage conditions significantly influence faecal
615 microbiome profiles. *Sci Rep* 2015; **5**: 16350.

616 26. Pinard R, de Winter A, Sarkis GJ, *et al.* Assessment of whole genome amplification-induced
617 bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics* 2006;
618 **7**: 216.

619 27. Dohm JC, Lottaz C, Borodina T, *et al.* Substantial biases in ultra-short read data sets from
620 high-throughput DNA sequencing. *Nucleic Acids Res* 2008; **36**: e105.

621 28. Jonsson V, Österlund T, Nerman O, *et al.* Variability in Metagenomic Count Data and Its
622 Influence on the Identification of Differentially Abundant Genes. *J Comput Biol* 2016:
623 cmb.2016.0180.

624 29. Glenn TC. Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 2011; **11**: 759–69.

625 30. O'Rawe JA, Ferson S, Lyon GJ. Accounting for uncertainty in DNA sequencing data. *Trends*
626 *Genet* 2015; **31**: 61–6.

627 31. Babraham Bioinformatics. Trim Galore! 2012. Available at:
628 http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

629 32. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina Sequence
630 Data. *Bioinformatics* 2014; **30**: btu170–2120.

631 33. Joshi NA, Fass JN. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ
632 files. Available at: https://github.com/najoshi/sickle

633 34. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets.
634 *Bioinformatics* 2011; **27**: 863–4.

635 35. Smooker PM, Whisstock JC, Irving JA, *et al.* A single amino acid substitution affects substrate
636 specificity in cysteine proteinases from Fasciola hepatica. *Protein Sci* 2000; **9**: 2567–72.

637 36. Johnson ET, Ryu S, Yi H, *et al.* Alteration of a single amino acid changes the substrate
638 specificity of dihydroflavonol 4-reductase. *Plant J* 2001; **25**: 325–33.

639 37. Glaser L, Stevens J, Zamarin D, *et al.* A single amino acid substitution in 1918 influenza virus
640 hemagglutinin changes receptor binding specificity. *J Virol* 2005; **79**: 11533–6.

641 38. Dabrazhynetskaya A, Brendler T, Ji X, *et al.* Switching protein-DNA recognition specificity by
642 single-amino-acid substitutions in the P1 par family of plasmid partition elements. *J Bacteriol* 2009;
643 **191**: 1126–31.

644 39. Atkinson HJ, Babbitt PC. An atlas of the thioredoxin fold class reveals the complexity of
645 function-enabling adaptations. *PLoS Comput Biol* 2009; **5**: e1000541.

646 40. Bianchi L, Díez-Sampedro A. A single amino acid change converts the sugar sensor SGLT3
647 into a sugar transporter. *PLoS ONE* 2010; **5**: e10241.

648 41. Bengtsson-Palme J, Boulund F, Edström R, *et al.* Strategies to improve usability and preserve
649 accuracy in biological sequence databases. *Proteomics* 2016; **16**: 2454–60.

650 42. Pal C, Bengtsson-Palme J, Rensing C, *et al.* BacMet: antibacterial biocide and metal resistance
651 genes database. *Nucleic Acids Res* 2014; **42**: D737–43.

652 43. 1928 Diagnostics. Resqu: A database of mobile antibiotic resistance genes. Available at:
653 http://1928diagnostics.com/resdb/

654 44. Altschul SF, Madden TL, Schäffer AA, *et al.* Gapped BLAST and PSI-BLAST: a new
655 generation of protein database search programs. *Nucleic Acids Res* 1997; **25**: 3389–402.

656 45. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology
657 Information. *Nucleic Acids Res* 2015.

658 46. Kurtz S. The Vmatch large scale sequence analysis software. Available at: http://vmatch.de/

659 47. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 2010;
660 **26**: 2460–1.

661 48. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat*
662 *Methods* 2015; **12**: 59–60.

663 49. Liu B, Pop M. ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res* 2009; **37**:
664 D443–7.

665 50. Nordmann P, Dortet L, Poirel L. Carbapenem resistance in Enterobacteriaceae: here is the
666 storm! *Trends Mol Med* 2012; **18**: 263–72.

667 51. Liu Y-Y, Wang Y, Walsh TR, *et al.* Emergence of plasmid-mediated colistin resistance
668 mechanism MCR-1 in animals and human beings in China: a microbiological and molecular
669 biological study. *Lancet Infect Dis* 2016; **16**: 161–8.

670 52. Ma L, Xia Y, Li B, *et al.* Metagenomic Assembly Reveals Hosts of Antibiotic Resistance
671 Genes and the Shared Resistome in Pig, Chicken, and Human Feces. *Environ Sci Technol* 2016; **50**:
672 420–7.

673 53. Jia B, Raphenya AR, Alcock B, *et al.* CARD 2017: expansion and model-centric curation of
674 the comprehensive antibiotic resistance database. *Nucleic Acids Res* 2016: gkw1004.

675 54. Gupta SK, Padmanabhan BR, Diene SM, *et al.* ARG-ANNOT, a new bioinformatic tool to
676 discover antibiotic resistance genes in bacterial genomes. *Antimicrob Agents Chemother* 2014; **58**:
677 212–20.

678 55. Zankari E, Hasman H, Cosentino S, *et al.* Identification of acquired antimicrobial resistance
679 genes. *J Antimicrob Chemother* 2012; **67**: 2640–4.

680 56. Bengtsson-Palme J, Angelin M, Huss M, *et al.* The Human Gut Microbiome as a Transporter
681 of Antibiotic Resistance Genes between Continents. *Antimicrob Agents Chemother* 2015; **59**: 6551–
682 60.

683 57. Dahllöf I, Baillie H, Kjelleberg S. rpoB-based microbial community analysis avoids limitations
684 inherent in 16S rRNA gene intraspecies heterogeneity. *Appl Environ Microbiol* 2000; **66**: 3376–80.

685 58. Bengtsson-Palme J, Alm Rosenblad M, Molin M, *et al.* Metagenomics reveals that
686 detoxification systems are underrepresented in marine bacterial communities. *BMC Genomics*
687 2014; **15**: 749.

688 59. Ma L, Li B, Zhang T. Abundant rifampin resistance genes and significant correlations of
689 antibiotic resistance genes and plasmids in various environments revealed by metagenomic
690 analysis. *Appl Microbiol Biotechnol* 2014; **98**: 5195–204.

691 60. Meyer F, Paarmann D, Dsouza M, *et al.* The metagenomics RAST server - a public resource
692 for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* 2008;
693 **9**: 386.

694 61. Durso LM, Miller DN, Wienhold BJ. Distribution and Quantification of Antibiotic Resistant
695 Genes and Bacteria across Agricultural and Non-Agricultural Metagenomes. *PLoS ONE* 2012; **7**:
696 e48325.

697 62. Nesme J, Cécillon S, Delmont TO, *et al.* Large-scale metagenomic-based study of antibiotic
698 resistance in the environment. *Curr Biol* 2014; **24**: 1096–100.

699 63. Knight R, Jansson J, Field D, *et al.* Unlocking the potential of metagenomics through
700 replicated experimental design. *Nat Biotechnol* 2012; **30**: 513–20.

701 64. Jonsson V, Österlund T, Nerman O, *et al.* Statistical evaluation of methods for identification
702 of differentially abundant genes in comparative metagenomics. *BMC Genomics* 2016; **17**: 78.

703 65. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating
704 differentially expressed gene lists from microarray data. *BMC Bioinformatics* 2006; **7**: 359.

705 66. Dinsdale EA, Edwards RA, Bailey BA, *et al.* Multivariate analysis of functional metagenomes.
706 *Front Genet* 2013; **4**: 41.

707 67. Bumgarner R. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc*
708 *Mol Biol* 2013; **Chapter 22**: Unit22.1.

709 68. Klappenbach JA, Dunbar JM, Schmidt TM. rRNA operon copy number reflects ecological
710 strategies of bacteria. *Appl Environ Microbiol* 2000; **66**: 1328–33.

711 69. Větrovský T, Baldrian P. The variability of the 16S rRNA gene in bacterial genomes and its
712 consequences for bacterial community analyses. *PLoS ONE* 2013; **8**: e57923.

713 70. Sunagawa S, Mende DR, Zeller G, *et al.* Metagenomic species profiling using universal
714 phylogenetic marker genes. *Nat Methods* 2013; **10**: 1196–9.

715 71. Manor O, Borenstein E. MUSiCC: a marker genes based framework for metagenomic
716 normalization and accurate profiling of gene abundances in the microbiome. *Genome Biol* 2015;
717 **16**: 53.

718 72. Bengtsson-Palme J, Hartmann M, Eriksson KM, *et al.* Metaxa2: Improved identification and
719 taxonomic classification of small and large subunit rRNA in metagenomic data. *Mol Ecol Resour*
720 2015; **15**: 1403–14.

721 73. Robinson MD, Oshlack A. A scaling normalization method for differential expression
722 analysis of RNA-seq data. *Genome Biol* 2010; **11**: R25.

723 74. Sohn MB, Du R, An L. A robust approach for identifying differentially abundant features in
724 metagenomic samples. *Bioinformatics* 2015; **31**: 2269–75.

725 75. Rapaport F, Khanin R, Liang Y, *et al.* Comprehensive evaluation of differential gene
726 expression analysis methods for RNA-seq data. *Genome Biol* 2013; **14**: R95.

727 76. Dillies M-A, Rau A, Aubert J, *et al.* A comprehensive evaluation of normalization methods for
728 Illumina high-throughput RNA sequencing data analysis. *Brief Bioinformatics* 2013; **14**: 671–83.

729 77. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems
730 biology. *Biol Direct* 2009; **4**: 14.

731 78. Smyth GK. Linear models and empirical bayes methods for assessing differential expression
732 in microarray experiments. *Stat Appl Genet Mol Biol* 2004; **3**: Article3.

733 79. Law CW, Chen Y, Shi W, *et al.* Voom: precision weights unlock linear model analysis tools for
734 RNA-seq read counts. *Genome Biol* 2014; **15**: R29.

735 80. O'Hara RB, Kotze DJ. Do not log-transform count data. *Methods Ecol Evol* 2010; **1**: 118–22.

736 81. Paulson JN, Stine OC, Bravo HC, *et al.* Differential abundance analysis for microbial marker-
737 gene surveys. *Nat Methods* 2013; **10**: 1200–2.

738 82. Schlenker E. Tips and Tricks for Successful Application of Statistical Methods to Biological
739 Data. *Methods Mol Biol* 2016; **1366**: 271–85.

740 83. Rodriguez-Brito B, Rohwer F, Edwards RA. An application of statistics to comparative
741 metagenomics. *BMC Bioinformatics* 2006; **7**: 162.

742 84. White JR, Nagarajan N, Pop M. Statistical methods for detecting differentially abundant
743 features in clinical metagenomic samples. *PLoS Comput Biol* 2009; **5**: e1000352.

744 85. Bengtsson-Palme J, Thorell K, Wurzbacher C, *et al.* Metaxa2 Diversity Tools: Easing
745 microbial community analysis with Metaxa2. *Ecol Inform* 2016; **33**: 45–50.

746 86. Kristiansson E, Hugenholtz P, Dalevi D. ShotgunFunctionalizeR: an R-package for

functional comparison of metagenomes. *Bioinformatics* 2009; **25**: 2737–8.

87. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010; **26**: 139–40.

88. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010; **11**: R106.

89. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos Mag Series 5* 1900; **50**: 157–75.

90. Noble WS. How does multiple testing correction work? *Nat Biotechnol* 2009; **27**: 1135–7.

91. Dunn OJ. Estimation of the medians for dependent variables. *Ann Math Stat* 1959; **30**: 192–7.

92. Dunn OJ. Multiple comparisons among means. *JASA* 1961; **56**: 52–64.

93. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B (Methodological)* 1995; **57**: 289–300.

94. Magurran AE. *Measuring Biological Diversity*. Oxford, UK: Blackwell Science Ltd; 2004.

95. Lundin D, Severin I, Logue JB, *et al*. Which sequencing depth is sufficient to describe patterns in bacterial a- and b-diversity? *Environ Microbiol Rep* 2012.

96. Hurlbert SH. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 1971; **52**: 577–86.

97. Hughes JB, Hellmann JJ. The application of rarefaction techniques to molecular inventories of microbial diversity. *Meth Enzymol* 2005; **397**: 292–308.

98. Chao A. Nonparametric Estimation of the Number of Classes in a Population. *Scand J Stat* 1984; **11**: 265–70.

99. Chao A, Lee S-M. Estimating the Number of Classes via Sample Coverage. *JASA* 1992; **87**: 210–7.

100. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. *Philos Trans R Soc Lond, B, Biol Sci* 1994; **345**: 101–18.

101. Hughes JB, Hellmann JJ, Ricketts TH, *et al*. Counting the uncountable: statistical approaches to estimating microbial diversity. *Appl Environ Microbiol* 2001; **67**: 4399–406.

102. Riesenfeld CS, Goodman RM, Handelsman J. Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. *Environ Microbiol* 2004; **6**: 981–9.

103. Sommer MOA, Dantas G, Church GM. Functional characterization of the antibiotic resistance reservoir in the human microflora. *Science* 2009; **325**: 1128–31.

104. Lang KS, Anderson JM, Schwarz S, *et al.* Novel florfenicol and chloramphenicol resistance gene discovered in Alaskan soil by using functional metagenomics. *Appl Environ Microbiol* 2010; **76**: 5321–6.

105. Torres-Cortés G, Millán V, Ramírez-Saad HC, *et al.* Characterization of novel antibiotic resistance genes identified by functional metagenomics on soil samples. *Environ Microbiol* 2011; **13**: 1101–14.

106. Wichmann F, Udikovic-Kolic N, Andrew S, *et al.* Diverse antibiotic resistance genes in dairy cow manure. *MBio* 2014; **5**: e01017.

107. Munck C, Albertsen M, Telke A, *et al.* Limited dissemination of the wastewater treatment plant core resistome. *Nat Commun* 2015; **6**: 8452.

108. Dantas G, Sommer MO. Context matters - the complex interplay between resistome genotypes and resistance phenotypes. *Curr Opin Microbiol* 2012; **15**: 577–82.

109. Martinez JL. Bottlenecks in the transferability of antibiotic resistance from natural ecosystems to human bacterial pathogens. *Front Microbiol* 2011; **2**: 265.

110. Dortet L, Nordmann P, Poirel L. Association of the emerging carbapenemase NDM-1 with a bleomycin resistance protein in Enterobacteriaceae and Acinetobacter baumannii. *Antimicrob Agents Chemother* 2012; **56**: 1693–7.

111. Toleman MA, Spencer J, Jones L, *et al.* blaNDM-1 is a chimera likely constructed in Acinetobacter baumannii. *Antimicrob Agents Chemother* 2012; **56**: 2773–6.

112. Johnning A, Moore ERB, Svensson-Stadler L, *et al.* Acquired genetic mechanisms of a multiresistant bacterium isolated from a treatment plant receiving wastewater from antibiotic production. *Appl Environ Microbiol* 2013; **79**: 7256–63.

113. Casali N, Nikolayevskyy V, Balabanova Y, *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat Genet* 2014; **46**: 279–86.

114. Salipante SJ, Roach DJ, Kitzman JO, *et al.* Large-scale genomic sequencing of extraintestinal pathogenic Escherichia coli strains. *Genome Res* 2015; **25**: 119–28.

115. Holt KE, Wertheim H, Zadoks RN, *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc Natl Acad Sci USA* 2015; **112**: E3574–81.

116. Flach C-F, Johnning A, Nilsson I, *et al.* Isolation of novel IncA/C and IncN fluoroquinolone resistance plasmids from an antibiotic-polluted lake. *J Antimicrob Chemother* 2015; **70**: 2709–17.

117. Myers EW, Sutton GG, Delcher AL, *et al.* A whole-genome assembly of Drosophila. *Science* 2000; **287**: 2196–204.

118. Margulies M, Egholm M, Altman WE, *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 2005; **437**: 376–80.

119. Chevreux B, Wetter T, Suhai S. Genome sequence assembly using trace signals and additional sequence information. *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* 1999; **99**: 45–56.

120. Staden R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res* 1979; **6**: 2601–10.

121. Pop M. Genome assembly reborn: recent computational challenges. *Brief Bioinformatics* 2009; **10**: 354–66.

122. Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. *Genomics* 2010; **95**: 315–27.

123. Warren RL, Sutton GG, Jones SJM, *et al.* Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* 2007; **23**: 500–1.

124. Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. *J Comput Biol* 1995; **2**: 291–306.

125. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA* 2001; **98**: 9748–53.

126. Li Z, Chen Y, Mu D, *et al.* Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct Genomics* 2012; **11**: 25–37.

127. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 2008; **18**: 821–9.

128. Simpson JT, Wong K, Jackman SD, *et al.* ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009; **19**: 1117–23.

129. Li R, Zhu H, Ruan J, *et al.* De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res* 2010; **20**: 265–72.

130. Namiki T, Hachiya T, Tanaka H, *et al.* MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res* 2012; **40**: e155.

131. Peng Y, Leung HCM, Yiu SM, *et al.* Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics* 2011; **27**: i94–101.

132. Nurk S, Meleshko D, Korobeynikov A, *et al.* metaSPAdes: a new versatile metagenomic assembler. *Genome Res* 2017. doi: 10.1101/gr.213959.116

133. Boisvert S, Raymond F, Godzaridis E, *et al.* Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biol* 2012; **13**: R122.

134. Vázquez-Castellanos JF, García-López R, Pérez-Brocal V, *et al.* Comparison of different assembly and annotation tools on analysis of simulated viral metagenomic communities in the gut. *BMC Genomics* 2014; **15**: 37.

135. Salzberg SL, Phillippy AM, Zimin A, *et al.* GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res* 2012; **22**: 557–67.

136. Magoc T, Pabinger S, Canzar S, *et al.* GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics* 2013; **29**: 1718–25.

137. Butler J, Maccallum I, Kleber M, *et al.* ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res* 2008; **18**: 810–20.

138. Narzisi G, Mishra B. Comparing de novo genome assembly: the long and short of it. *PLoS ONE* 2011; **6**: e19175.

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

140. Howe AC, Jansson JK, Malfatti SA, *et al.* Tackling soil diversity with the assembly of large, complex metagenomes. *Proc Natl Acad Sci USA* 2014; **111**: 4904–9.

141. Pell J, Hintze A, Canino-Koning R, *et al.* Scaling metagenome sequence assembly with probabilistic de Bruijn graphs. *Proc Natl Acad Sci USA* 2012; **109**: 13272–7.

142. Hess M, Sczyrba A, Egan R, *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* 2011; **331**: 463–7.

143. Mackelprang R, Waldrop MP, Deangelis KM, *et al.* Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. *Nature* 2011; **480**: 368–71.

144. Scholz M, Lo C-C, Chain PSG. Improved Assemblies Using a Source-Agnostic Pipeline for MetaGenomic Assembly by Merging (MeGAMerge) of Contigs. *Sci Rep* 2014; **4**: 6480.

145. Zhang Y, Sun Y, Cole JR. A Scalable and Accurate Targeted Gene Assembly Tool (SAT-Assembler) for Next-Generation Sequencing Data. *PLoS Comput Biol* 2014; **10**: e1003737.

146. Frost LS, Leplae R, Summers AO, *et al.* Mobile genetic elements: the agents of open source evolution. *Nat Rev Microbiol* 2005; **3**: 722–32.

147. Norman A, Hansen LH, Sørensen SJ. Conjugative plasmids: vessels of the communal gene pool. *Phil Trans R Soc B: Biol Sci* 2009; **364**: 2275–89.

148. Raymond F, Ouameur AA, Déraspe M, *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J* 2016; **10**: 707–20.

149. Pehrsson EC, Tsukayama P, Patel S, *et al.* Interconnected microbiomes and resistomes in low-income human habitats. *Nature* 2016; **533**: 212–6.

150. Hasman H, Saputra D, Sicheritz-Ponten T, *et al.* Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *J Clin Microbiol* 2014; **52**: 139–46.

151. Schmidt K, Mwaigwisya S, Crossman LC, *et al.* Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. *J Antimicrob Chemother* 2017; **72**: 104–14.

152. Cao MD, Ganesamoorthy D, Elliott AG, *et al.* Streaming algorithms for identification of pathogens and antibiotic resistance potential from real-time MinION(TM) sequencing. *Gigascience* 2016; **5**: 32.

153. Grumaz S, Stevens P, Grumaz C, *et al.* Next-generation sequencing diagnostics of bacteremia in septic patients. *Genome Med* 2016; **8**: 73.

154. Munk P, Andersen VD, de Knegt L, *et al.* A sampling and metagenomic sequencing-based methodology for monitoring antimicrobial resistance in swine herds. *J Antimicrob Chemother* 2017; **72**: 385–92.

155. Hu Y, Liu F, Lin IYC, *et al.* Dissemination of the mcr-1 colistin resistance gene. *Lancet Infect Dis* 2016; **16**: 146–7.

156. Human Microbiome Jumpstart Reference Strains Consortium, Nelson KE, Weinstock GM, *et al.* A catalog of reference genomes from the human microbiome. *Science* 2010; **328**: 994–9.

157. Lee JE, Lee S, Sung J, *et al.* Analysis of human and animal fecal microbiota for microbial source tracking. *ISME J* 2011; **5**: 362–5.

158. Ashbolt NJ, Amézquita 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.

Figure legends

Figure 1. Sequence identity between variants assigned to the same resistance gene group in the Resqu database. Sequences were aligned using MAFFT and pairwise identities were calculated as the number of identical amino acids in corresponding positions, discarding gaps in one or both of the sequences. The x-axis represents the numbers of sequences corresponding to each group of resistance genes (gene name). The x-axis is log-transformed for viewing purposes.

Figure 2. Differences in total resistance abundance reported by the same bioinformatic method using four different reference databases: ARDB, the full 2015 version of the CARD database, the metagenomics-adapted version of CARD, and Resqu.

Figure 3. Relationship between the abundances of human-associated bacteria (classified as being present in the Human Microbiome Project genome catalog) and antibiotic resistance genes in the 864 metagenomes investigated by Pal *et al.*[14]

Figure 4. A suggested workflow for resistance gene analysis in metagenomes. Specific recommended tools and databases are indicated by white boxes, while conceptual approaches are given in black boxes. Methodological steps are marked in grey boxes.

Figure S1. Identical resistance genes may exist in (a) multiple genetic contexts or have certain regions that are identical between variants even if they encode slightly different proteins (b). This presents assembly software with serious problems, as the reads that originated from which context cannot be identified (center). Almost all assemblers solve this by splitting the contigs at the ambiguous positions, resulting in a fragmented assembly (bottom). Notice how the repetition of resistance gene A in (a) cause a loop in the assembly graph, resulting in two short contigs containing

924 no genes. In (b), most resistance regions are not assigned to any context, and no full-length variant
925 of the resistance gene could be assembled.

926