MBMG
Metabarcoding & Metagenomics

Metabarcoding and Metagenomics 3: 59–67
DOI 10.3897/mbmg.3.36236

Software Description

Mumame: a software tool for quantifying gene-specific point-mutations in shotgun metagenomic data

Shruthi Magesh[1,2], **Viktor Jonsson**[3], **Johan Bengtsson-Palme**[1,4,5]

1 *Wisconsin Institute for Discovery, University of Wisconsin-Madison, 330 North Orchard Street, Madison WI 53715, USA*
2 *Department of Biotechnology, School of Bioengineering, SRM Institute of Science and Technology, Kattankulathur 603203, Tamil Nadu, India*
3 *Chalmers Computational Systems Biology Infrastructure, Chalmers University of Technology, SE-412 96, Gothenburg, Sweden*
4 *Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Guldhedsgatan 10, SE-413 46, Gothenburg, Sweden*
5 *Centre for Antibiotic Resistance research (CARe) at University of Gothenburg, Gothenburg, Sweden*

Corresponding author: *Johan Bengtsson-Palme* (johan.bengtsson-palme@microbiology.se)

Academic editor: *Florian Leese* | **Received** 15 May 2019 | **Accepted** 6 August 2019 | **Published** 4 September 2019

Abstract

Metagenomics has emerged as a central technique for studying the structure and function of microbial communities. Often the functional analysis is restricted to classification into broad functional categories. However, important phenotypic differences, such as resistance to antibiotics, are often the result of just one or a few point mutations in otherwise identical sequences. Bioinformatic methods for metagenomic analysis have generally been poor at accounting for this fact, resulting in a somewhat limited picture of important aspects of microbial communities. Here, we address this problem by providing a software tool called Mumame, which can distinguish between wildtype and mutated sequences in shotgun metagenomic data and quantify their relative abundances. We demonstrate the utility of the tool by quantifying antibiotic resistance mutations in several publicly available metagenomic data sets. We also identified that sequencing depth is a key factor to detect rare mutations. Therefore, much larger numbers of sequences may be required for reliable detection of mutations than for most other applications of shotgun metagenomics. Mumame is freely available online (http://microbiology.se/software/mumame).

Key Words

Antibiotic resistance, bioinformatic tools, metagenomics, mutation frequencies, mutation detection, statistical methods

Introduction

The revolution in sequencing capacity has created an unprecedented ability to glimpse into the functionality of microbial communities, using large-scale shotgun metagenomic techniques (Quince et al. 2017). This has yielded important insights into broad functional patterns of microbial consortia (Yooseph et al. 2007; Human Microbiome Project Consortium 2012; Sunagawa et al. 2015). However, while overall pathway abundances inferred from metagenomic data can tell us much about the general functions of communities and how they change with, for example, environmental gradients (Bengtsson-Palme 2018; Bahram et al. 2018), there are many important functional differences that are hidden in the subtleties of these communities (Österlund et al. 2017). For example, many antibiotic resistance phenotypes are the results of single point mutations rather than acquisition of novel pathways or genes (Johnning et al. 2013). This complicates the studies of selection pressures in environmental communities as analysis of such mutations is generally limited to a narrow range of species (Johnning et al. 2015a, 2015b; Kraupner et al. 2018).

Because of the immense increase in available sequence data, it would be desirable to study these mutations from shotgun metagenomic libraries, much as other traits have

PENSOFT.

been studied at a large scale (Pal et al. 2016). However, attempts to quantify point mutations in metagenomic sequencing data often go wrong because the methods do not distinguish sufficiently well between mutated and wildtype variants of the same gene. For example, a sequenced read may map to a region identical in the mutated and wildtype variant of a gene, causing problems for quantifying their relative proportions (Bengtsson-Palme et al. 2017). In addition, because the sought-after mutations are generally rare in most types of samples, and metagenomic studies are often under-sampled in terms of replicates (Jonsson et al. 2017), commonly applied statistical methods may not be sufficiently sensitive to reliably detect differences between samples (Jonsson et al. 2016).

In this study, we provide a partial remedy to these problems through the introduction of a software tool, Mumame (Mutation Mapping in Metagenomes), that can quantify and distinguish between wildtype and mutated gene variants in metagenomic data, and through suggesting a statistical framework for handling the output data of the software. In contrast to available tools for investigating nucleotide variants, including StrainPhlAn (Truong et al. 2017), ConStrains (Luo et al. 2015), and SeekDeep (Hathaway et al. 2018), Mumame is not aiming to find strain-level differences in taxonomic composition, thus enabling it to operate at much lower sequencing depths as complete coverage of the targeted genomes is not necessary for the analysis. Furthermore, while tools such as MIDAS (Nayfach et al. 2016) and metaSNV (Costea et al. 2017) allow detection of nucleotide variant differences between population at a large scale, they are reliant on collections of high-quality reference genomes to which they map reads. This allows them to provide more information than Mumame does in situations where the studied community can be expected to be well represented by the reference databases. In contrast, Mumame operates on the protein level (or optionally the nucleotide level when relevant), allowing detection of specific, functionally relevant point mutations even in evolutionary distant homologs to the targeted genes. This enables the application of our method to a larger body of metagenomes, with less bias towards already well-characterized species.

Finally, we demonstrate the ability of Mumame to detect relevant differences between environmental sample types, estimate the sequencing depths required for the method to perform reliably through simulations, and exemplify the utility of the software on detecting resistance mutations in publicly available metagenomes. The Mumame software package is open-source and freely available (http://microbiology.se/software/mumame or https://github.com/bengtssonpalme/mumame).

Methods

Software implementation

Mumame is implemented in Perl and consists of two commands: mumame, which performs read alignment to a database of mutations, and mumame_build which builds the database for the former command. The mumame_build command takes a FASTA sequence file and a list of mutations (CSV format) as input. For each entry in the mutation list, it finds the corresponding sequence(s) in the FASTA file, either by sequence identifier or by CARD ARO accessions (Jia et al. 2017). It then excerpts a number of residues upstream and downstream of the mutation position (by default 20 residues for proteins and 55 for nucleotide sequences) and creates one wildtype version and one mutated version of the sequence excerpt with unique sequence IDs. For cases where multiple mutations can occur close to each other on the same sequence, the software attempts to create all possible combinations of mutations (if memory permits; in some situations this is not possible because the number of combinations increases exponentially). The software tool also generates a mapping file between sequence IDs in the database and mutation information from the list.

The main mumame command takes any number of input files containing DNA sequence reads in FASTA or FASTQ format and aligns those against the Mumame database using Usearch (Edgar 2010). For this read alignment, the software runs Usearch in usearch_global mode with target coverage set to 0.55 (by default; any value ≥0.51 should be feasible for target coverage). The output is then matched to the wildtype or mutation information in the Mumame database, and data is collected for each input file and combined into a single output table. The software uses a best-hit strategy, with hits sorted by identity to the reference sequence. As every sequence in the database is present in two versions, one variant with and one without the point mutations, the top hit will always discriminate between the two variants. In addition, as only the immediate region around the mutation site is included in the database, there cannot be any spurious hits adding to wildtype or mutated variants, providing Mumame with the maximum possible degree of precision. The choice of Usearch for read alignment was made because it constitutes the most versatile sequence search software in that it is, unlike most other sequence search tools for large-scale data, able to align nucleotide sequences to both protein and nucleotide sequence databases. However, the software design of Mumame is fairly flexible, allowing implementation of other software for the sequence search process with relatively small effort.

The main output of Mumame is a file with the suffix ".table.txt". This file contains the reads from each library aligned to the mutation database, with mutation counts in the first set of columns and wildtype counts in the second set of columns. The last line of this file contains the total number of reads in each library, which can be used, e.g., for normalization purposes. The software also saves the output from the Usearch run and, optionally, the read alignments to the database. The output table generated by Mumame can be analyzed using the R script (R Core Team 2016) supplied with the Mumame package. The script reads the read counts for all mutation positions detected, both for wildtype and mutated sequences. The script also takes into account the total library sizes, either explicitly as normalization factors in the count model or implicitly in the proportional model as total library size cancels out (see below). The script assesses if there are significantly different proportions of mutations between different sample groups through a generalized linear model. Alternatively, an overdispersed Poisson generalized linear model (GLM) accounting for the discrete nature of the data and the differences in sequencing depth can be used (Jonsson et al. 2016; Bengtsson-Palme et al. 2017). These two tests were selected after investigating the performance of the two GLM tests, the Student's *t*-test on the mutation proportions and the Chi-Square test on the total counts in a simulation study. We simulated a set of different numbers of replicates, effect sizes, sequencing depths and average gene abundances (Suppl. material 1: Table S1), 100 times for each combination of conditions and assumed that counts were Poisson distributed. The simulations showed that the two GLM models overall perform better in terms of detecting significances, particularly when the sequencing depths and effect sizes are small, while generating the expected proportions of false positive detections. The Poisson model is preferable when the number of counts for a targeted gene is low in all sample groups. However, this model performs poorly when estimating effect sizes with small numbers of counts. This is due to situations where one group has zero counts for all replicates. In practice, this means that even if a result is deemed significant by the model, the estimated effect size should only be considered an indication of the directionality of the effect when counts are all zeros in one group.

The Mumame software is freely available (http://microbiology.se/software/mumame or https://github.com/bengtssonpalme/mumame) and can also be installed via Conda, using the command "conda install -c bengtssonpalme mumame".

Quantification of mutations in metagenomes

To quantify the abundances of fluoroquinolone resistance mutations in the *gyrA* and *parC* genes (Johnning et al. 2015b), we downloaded the CARD database on 2018–05–24 (Jia et al. 2017). We extracted all mutation information regarding the *gyrA* and *parC* genes from the "snps.txt" file and created a new file with that information. We then created a new Mumame database using mumame_build with default options. That database was used to align all the reads from the samples generated by Kraupner et al. (2018; data provided by courtesy of Stefan Ebmeyer and Joakim Larsson) to the database using Mumame in the Usearch mode (Edgar 2010) and the following options "-d gyrA_parC -c 0.95". This study exposed aquatic bacterial communities in an aquarium system to different concentrations of ciprofloxacin (0, 0.1, 1 and 10 µg/L) in triplicates. The communities were then subjected to both amplicon sequencing of the target genes for ciprofloxacin (56,394–290,391 reads per sample after preprocessing) and shotgun metagenomics (109–190 million reads per sample). We analyzed both the shotgun metagenomics data as well as the amplicon sequences derived specifically from Enterobacteriaceae *gyrA* and *parC* genes. Prior to this sequence similarity search raw reads were quality filtered using Trim Galore! (Babraham Bioinformatics 2012) with the settings "-e 0.1 -q 28 -O 1". We then used the R script provided with the Mumame software to compare the numbers of matches to mutated and wildtype sequences in the database. The same database and method combination was used to quantify fluoroquinolone resistance mutations in sequence data from an Indian lake exposed to ciprofloxacin pollution (Bengtsson-Palme et al. 2014; ENA project PRJEB6102; https://www.ebi.ac.uk/ena/data/view/PRJEB6102), as well as in an Indian river upstream and downstream of a wastewater treatment plant processing pharmaceutical waste (Kristiansson et al. 2011; Pal et al. 2016; MG-RAST project 18323; https://www.mg-rast.org/linkin.cgi?project=mgp18323). These samples were preprocessed in the same way as in the Indian lake study (Bengtsson-Palme et al. 2014). These samples were taken as part of a survey of environments close to a pharmaceutical production waste treatment facility and consist of two samples of river sediment from upstream of the treatment plant, one by the outlet, three taken downstream, and one from a nearby lake. These were subjected to shotgun metagenomic sequencing, generating 14.5–37.3 million reads per sample.

Finally, we investigated data from the experiment by Lundström et al. (2016; ENA project PRJEB11402; https://www.ebi.ac.uk/ena/data/view/PRJEB11402), in which aquatic bacterial communities were investigated in an aquarium system exposed to different concentrations of tetracycline (0, 0.1, 1, 10, 100 and 1000 µg/L). The samples were sequenced using shotgun metagenomics to a depth of 122–273 million reads per sample. The experiment used two replicates per concentration, meaning that we can only test for statistically significant trends in the data rather than differences between specific concentrations. To quantify resistance mutations in the sequence data, we created a Mumame database for tetracycline resistance mutations in the 23S rRNA gene targeted by tetracycline. We extracted the mutational information related to tetracycline from the CARD "snps.txt" file and then built the database using mumame_build with the additional option "-n". We then aligned all reads from the Lundström et al. (2016) data to the Mumame database using the options "-d Tet -c 0.95 -n". Reads were quality filtered and statistical differences were assessed as above.

Software evaluation

To assess the limitations of the method in terms of sequencing depth, the samples from the highest and lowest ciprofloxacin concentrations generated by Kraupner et al. (2018; 10 µg/L and 0 µg/L, respectively) were downsampled to 1, 5, 10, 20, 30, 40, and 50 million reads. There-after, the reads from the downsampled libraries were aligned to the fluoroquinolone resistance mutation database using Mumame as above. Statistical differences were assessed at all simulated sequencing depths and average effect sizes calculated for the significantly altered genes.

Results

Mumame can quantify point mutation frequencies in metagenomic data

As a proof-of-concept that our method to identify point mutations in metagenomic sequence data is functional, we used Mumame to quantify the mutations in amplicon data from the *gyrA* and *parC* genes. These genes are targets of fluoroquinolone antibiotics, and often acquire resistance mutations attaining high levels of resistance. We quantified such mutations in an amplicon data set specifically targeting these two genes in *Escherichia coli*. This data set derives from an exposure study with increasing ciprofloxacin concentrations, and enrichments of mutations in the classical fluoroquinolone resistance determining positions S83 and D87 (*gyrA*) and S80 and E84 (*parC*) have previously been verified using other bioinformatic methods (Kraupner et al. 2018). This data set, therefore, serves as an ideal positive control for our novel method. We found that Mumame was able to identify the difference between the highest concentration (10 µg/L) and the lower ones reported in the original study (Fig. 1). However, Mumame only reported an average frequency of mutations of around 11–12% for *gyrA* mutations (Fig. 1A), while the original paper found frequencies of 60–85% (S83) and 30–40% (D87). The A67 position was not quantified in the original paper. The exact reason for the discrepancies is unknown, but it is likely caused by a taxonomic filtration step that selects for *E. coli* reads used in the Kraupner et al. (2018) study, while Mumame does not perform prior filtering. The decision to exclude filtering was made in order to mimic a situation with true metagenomic data where several target species may co-exist. For *parC*, Mumame only quantified the S80 position (Fig. 1C), because the E84 mutations were not included in the version of the CARD database used for this study. For position S80, Mumame identified around 35% mutated sequences at the highest concentration of ciprofloxacin, while the original study reported around 50%. When a similar *E. coli* filtering step was introduced in our analysis, the proportions were much closer to those in the original paper (data not shown).

We next evaluated the performance of Mumame on the real shotgun data that was generated from the same samples as the amplicon libraries. Ideally, this analysis should generate virtually the same result as the amplicon analysis. Indeed, we found similar results for the A67 and S83 *gyrA* mutations (Fig. 1B). For the D87 mutation, the frequencies were much lower than for the other two mutations, albeit still significantly larger than at the lower concentrations ($p < 0.01$). For the *parC* gene, the shotgun metagenomic analysis had large variability within the sample groups, which prevented any statistically significant results (Fig. 1D). This is surprising, given that the total number of reads detected for both the mutant and wildtype variants were higher for the *parC* gene than for *gyrA*. Thus, the large degree of variability could potentially be due to features of the *parC* gene. For example, it is possible that this gene has higher similarity to closely related species, leading to that species replacement could influence the results. Taken together, these results indicate the high noise levels present for individual gene variants even in deeply sequenced shotgun metagenomes from controlled exposure studies.

The limits to quantification

Noting the much more instable levels of mutations in the shotgun metagenomes, we next investigated the effects of sequencing depth on the ability of our method to detect significantly altered mutation frequencies. For this analysis, we used downsampled data from the shotgun metagenomic library of the ciprofloxacin exposure study (Fig. 2). As expected, we found that the number of significantly altered mutation frequencies detected increased with larger sequencing depth (Fig. 2A). In addition, the average effect size of the significant mutations became gradually lower with larger sequence depth, also in accordance with expectations (Fig. 2B). Importantly, the average effect size of detectable mutation frequency differences seems to decrease linearly with sequencing depth. This means that we can calculate an expected detection limit for the method given the characteristics of the data and experimental setup. At 10 million reads, we expect that the proportion of reads with mutation must be 30–40% higher in the exposed sample in order for it to be detected as significant. This proportion decreases, on average, to 10% for 50 million reads (Fig. 2B). These numbers also depend on other factors, such as the number of replicates per treatment, but nevertheless they can be used as an approximation to aid the design of metagenomic studies or to interpret non-significant results derived from Mumame analyses.

Tetracycline-exposed Escherichia coli populations do not harbor higher abundances of resistance mutations

After validating the method and testing the limit of detection, we used Mumame to quantify resistance mutations in a similar controlled aquarium setup under exposure to the antibiotic tetracycline (Lundström et al. 2016). In this study, no amplicon sequencing of the target gene for tetracycline (23S rRNA) was performed, and thus there was no *a priori* true result to which we could compare. While Mumame was able to successfully detect tetracycline resistance mutations in the data, we somewhat surprisingly found no enrichment of such mutations in this data (Fig. 3). Notably, this result was obtained despite a very high sequencing depth (on average 181,595,072 paired-end se-

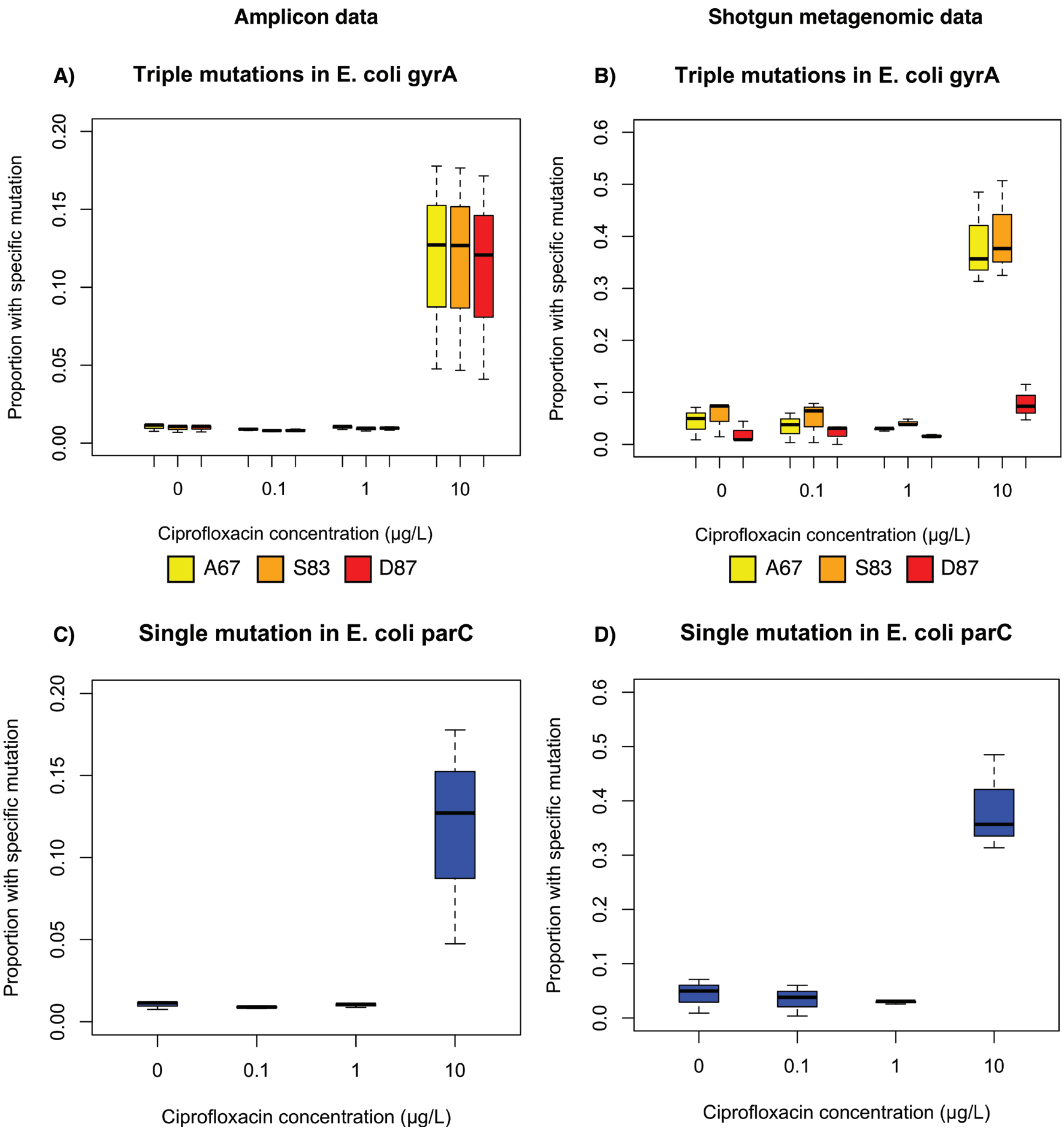

Figure 1. Fluoroquinolone resistance mutations in ciprofloxacin-exposed bacterial communities. Total mutation frequencies quantified using Mumame for three known mutations conferring resistance to fluoroquinolone in the *E. coli gyrA* gene based on amplicon sequencing (**A**) and shotgun metagenomic data (**B**) from the same samples. Corresponding data for the S80 mutation in *parC* is shown in (**C**) for amplicon data, and (**D**) for shotgun data.

quences per library). Obtaining a negative result at this sequencing depth suggests that there actually is no enrichment of known *E. coli* resistance mutations in the samples.

Fluoroquinolone resistance mutations in ciprofloxacin-polluted sediments

As a final investigation of the performance of the method, we also let Mumame quantify the fluoroquinolone resistance mutations in river and lake sediments polluted by antibiotic manufacturing waste, primarily ciprofloxacin (Kristiansson et al. 2011; Bengtsson-Palme et al. 2014; Pal et al. 2016). These libraries are fairly old and were not as deeply sequenced as the other data sets we investigated. While the experimental setup of these studies in terms of number of replicate samples per geographical location does not allow for proper statistical testing, we found overall more fluoroquinolone resistance mutations downstream of the pollution source, at least for the *E. coli gyrA* and *parC* genes (Fig. 4). We also detected a few

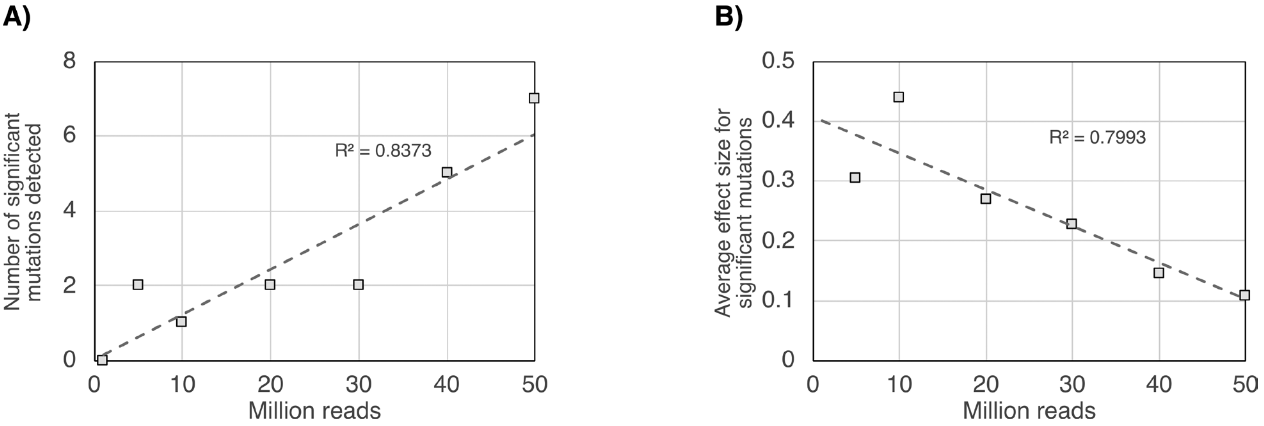

Figure 2. Influence of sequencing depth on detected mutations and their effect sizes. Relationship between the number of investigated reads and number of mutations with significantly altered frequencies (**A**) and the average effect size for those mutations (**B**); as assessed using Mumame on shotgun metagenomic data from a ciprofloxacin exposure experiment.

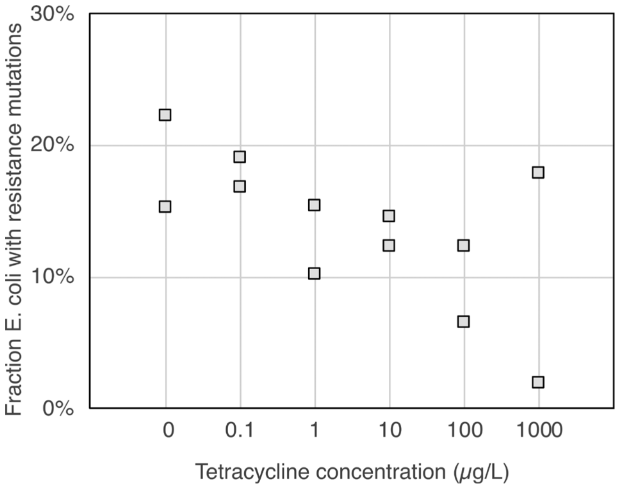

Figure 3. Resistance mutations in tetracycline-exposed bacterial communities. Frequencies of *E. coli* tetracycline resistance mutations at exposure to different concentrations of tetracycline, based on shotgun metagenomic data.

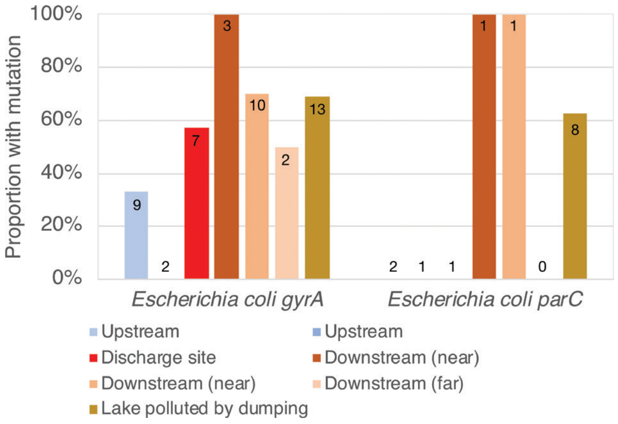

Figure 4. Resistance mutations in antibiotic-polluted sediments. Relative frequency of *gyrA* and *parC* sequences with resistance mutations in samples taken downstream, at, or upstream of the pharmaceutical production wastewater treatment plant, as well as in a lake polluted by dumping of pharmaceutical production waste. The numbers at the top of the bars show the total number of *gyrA*/*parC* sequences (wildtype or mutated) identified in each sample.

such mutations in other species, but the counts of those were low and the results largely non-informative due to the small number of detections per mutation (Suppl. material 2: Fig. S1). This serves as an example of that mutations can be detected also in shallowly sequenced metagenomic data, but that without proper experimental design, interpretation of the results is difficult or even impossible.

Discussion

Metagenomics often becomes restricted to investigate gross compositional changes to the taxonomy and functional genes of microbial communities. Unfortunately, this obscures important variation between individual sequence variants that may have large impacts on phenotypes (Österlund et al. 2017; Bengtsson-Palme 2018). One example of such point mutations inducing strong phenotypic changes is resistance mutations in the target genes of antibiotics (Kraupner et al. 2018). However, including mutated sequence variants in the antibiotic resistance gene databases is complicated and can lead to gross misinterpretations of the data (see, for example, Ma et al. 2014). Still, understanding relevant variation between sequences and linking that to phenotypes is somewhat of a holy grail of metagenomics. This study has made clear that we are not yet at that point in terms of bioinformatic methods and sequencing depths required to draw firm conclusions. That said, we show in this work that using shotgun metagenomic data to identify significant and relevant differences in resistance mutation frequencies between sample groups is possible, given a sufficiently large sequencing effort. However, the quantitative estimates still seem to be highly variable, even at very large sequencing depths.

The results of the Mumame evaluation also provides a few other important clues on potential pitfalls with inferring mutation frequencies from shotgun metagenomic data. An important such aspect is the disparity between mutation frequencies described by amplicon sequencing and shotgun data. Particularly, the ability to relatively consistently identify the A67 and S83 mutations in *parC*, while the D87 mutation was seemingly less frequent in the shotgun data is somewhat troubling if the goal is to quantify the actual abundances of such mutations. At the same time, the statistical significance of those differences could still be detected. For the A67 and S83 mutations, only 5 million reads were required for a significant effect to be detected, while for the D87 mutations a sequencing depth of 50 million reads was required. This is not necessarily a shortcoming of the Mumame software, but may just as well be due to the much noisier nature of the relatively few counts from metagenomic sequence data compared to the large number of reads corresponding to the same genes deriving from amplicon data (Jonsson et al. 2017).

Another important potential problem highlighted by our evaluation is the need to produce very large sequence data sets to be able to identify and quantify mutated (and wildtype) sequences with any certainty. As a rule of thumb, the targeted regions represent less than 0.005% of the bacterial genome, and each bacterial strain may correspond to only a fraction of a percent of the reads in the shotgun sequence data (depending on its abundance). This means that to identify a single read from a resistance region in the data, one would need to sequence, on average, more than five million reads. To get a reasonably confident measure of reads stemming from wildtype strains versus strains with mutations, approximately 10 reads from each group would be needed per sample (or, say, 20 reads in total). That would, as a rough estimate, correspond to a hundred million reads per sample. This is, unfortunately, far more sequences than what is typically generated per sample by shotgun metagenomic sequencing projects. Naturally, these numbers would depend on the proportions of the targeted microorganisms as well as their genome sizes, but ultimately this still presents the largest limitation to mutation studies based on metagenomic sequence data, regardless of how sophisticated bioinformatics methods that are used. Potentially, this problem could be partially alleviated by analyzing sufficiently large cohorts and performing the statistical analysis for general trends, but even large cohorts would be insufficient for mutations rare enough to pass below the detection limit.

In terms of interpreting the results from the exposure experiments, it is interesting to note the overall clear increase of fluoroquinolone resistance mutations at the highest ciprofloxacin concentration, which nearly perfectly corresponds to increases in mobile *qnr* fluoroquinolone genes in the same samples (Kraupner et al. 2018). This is contrasted by the trend seen in the tetracycline exposure experiments, where tetracycline resistance genes, specifically efflux pumps, were enriched at higher tetracycline concentrations (Lundström et al. 2016), while tetracycline resistance mutation abundances were not significantly altered according to our investigation of the same sequence data. This non-significant result was obtained despite the exceptionally high sequencing depth of those samples. As suggested in the original paper, the apparent decrease of resistance mutations with tetracycline concentration may be due to strain displacement and suggests that in this context these resistance mutations have little influence over the actual ability to tolerate increasing levels of tetracycline.

While we did not have data from an experimental setup suitable to address differences between sediments exposed to different degrees of fluoroquinolone pollution, the quantification of resistance mutations seems to provide an important piece of information to explain the results of previous studies of resistance gene abundances in these river samples (Kristiansson et al. 2011). In the original paper, the abundance of mobile fluoroquinolone resistance genes (*qnr* genes) was shown to be enriched in the low-level polluted upstream samples, compared to the highly polluted downstream samples. Importantly, the *qnr* genes only provide resistance to relatively low levels of fluoroquinolones (Hooper and Jacoby 2015), and Kristiansson et al. (2011) hypothesized that chromosomal mutations in the target genes are probably necessary to survive the selection pressure from antibiotics downstream of the pollution source. Our work suggests that this assumption is likely to be correct. Only a limited number of reads were aligned to these resistance regions and the number of samples unfortunately prevents us from properly assessing a statistical difference between the upstream and downstream samples. Still, the proportions of resistance mutations seem to be systematically higher in the samples downstream of the pollution source, at least for *E. coli*. This indicates that the method we present here can provide important additional information to metagenomic studies of resistance patterns in different environment types, given that a sufficient sequencing depth is achieved.

We have here shown the utility of the Mumame tool for finding resistance mutations in shotgun metagenomic data. In this paper, we have used the CARD database (Jia et al. 2017) as the information source for resistance mutations, but the tool is flexible to use any source of such data. It is also not in any means restricted to the mutations investigated in this paper but is fundamentally agnostic to the input data. It does, however, depend on information on already described mutations and, thus, cannot at present be used for *de novo* identification of mutations. The overall lack of databases comprehensively listing point mutations with important functional implications somewhat limits the use of Mumame in other contexts than detection of antibiotic resistance mutations. We hope that the development of this software can help spur the creation of such resources for a wider variety of biological functions. Mumame can be used in open screening for mutations in any gene present in the database in parallel and can handle different mutations in both RNA and protein coding genes. The ability of Mumame to operate on the protein level enables detection of specific, functionally relevant point mutations even when the target genes only have limited homology. Thus, Mumame can be applied in a larger set of situations and is less biased towards well-characterized taxonomic groups. However, at present this feature comes at the expense of not being able to detect mutations *de novo*; for this task, software such as MIDAS (Nayfach et al. 2016) or metaSNV (Costea et al. 2017) would be better suited. Mumame is flexible and fast and therefore can be implemented as a part of nearly any screening pipeline for antibiotic resistance information in metagenomic data sets.

Conclusion

This paper presents a software tool called Mumame to analyze shotgun metagenomic data for point mutations, such as those conferring antibiotic resistance to bacteria. Mumame can distinguish between wildtype and mutated gene variants in metagenomic data and quantify them, given a sufficient sequencing effort. We also provide a statistical framework for handling the generated count data and account for factors such as differences in sequencing depth. Importantly, our study also reveals the importance of a high sequencing depth, preferably more than 50 million sequenced reads per sample, in order to get reasonably accurate estimates of mutation frequencies, particularly for rare genes or species. The Mumame software package is freely available from http://microbiology.se/software/mumame. We expect Mumame to be a useful addition to metagenomic studies of, for example, antibiotic resistance, and to increase the detail by which metagenomes can be screened for phenotypically important differences.

Acknowledgements

This work was funded by the Swedish Research Council for Environment, Agricultural Sciences, and Spatial Planning (FORMAS; grant 2016-00768).

References

Babraham Bioinformatics (2012) Trim Galore! https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Bahram M, Hildebrand F, Forslund SK, Anderson JL, Soudzilovskaia NA, Bodegom PM, Bengtsson-Palme J, Anslan S, Coelho LP, Harend H, Huerta-Cepas J, Medema MH, Maltz MR, Mundra S, Olsson PA, Pent M, Põlme S, Sunagawa S, Ryberg M, Tedersoo L, Bork P (2018) Structure and function of the global topsoil microbiome. Nature 320: 1039. https://doi.org/10.1038/s41586-018-0386-6

Bengtsson-Palme J (2018) Strategies for taxonomic and functional annotation of metagenomes. In: Nagarajan M (Ed.) Metagenomics: Perspectives, Methods, and Applications. Academic Press, Elsevier (Oxford, UK). https://doi.org/10.1016/B978-0-08-102268-9.00003-3

Bengtsson-Palme J, Boulund F, Fick J, Kristiansson E, Larsson DGJ (2014) Shotgun metagenomics reveals a wide array of antibiotic resistance genes and mobile elements in a polluted lake in India. Frontiers in Microbiology 5: 648. https://doi.org/10.3389/fmicb.2014.00648

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

Costea PI, Munch R, Coelho LP, Paoli L, Sunagawa S, Bork P (2017) metaSNV: a tool for metagenomic strain level analysis. PLoS ONE 12: e0182392. https://doi.org/10.1371/journal.pone.0182392

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Hathaway NJ, Parobek CM, Juliano JJ, Bailey JA (2018) SeekDeep: single-base resolution de novo clustering for amplicon deep sequencing. Nucleic Acids Research 46: e21. https://doi.org/10.1093/nar/gkx1201

Hooper DC, Jacoby GA (2015) Mechanisms of drug resistance: quinolone resistance. Annals of the New York Academy of Sciences 1354: 12–31. https://doi.org/10.1111/nyas.12830

Human Microbiome Project Consortium (2012) Structure, function and diversity of the healthy human microbiome. Nature 486: 207–214. https://doi.org/10.1038/nature11234

Jia B, Raphenya AR, Alcock B, Waglechner N, Guo P, Tsang KK, Lago BA, Dave BM, Pereira S, Sharma AN, Doshi S, Courtot M, Lo R, Williams LE, Frye JG, Elsayegh T, Sardar D, Westman EL, Pawlowski AC, Johnson TA, Brinkman FS, Wright GD, McArthur AG (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. Nucleic Acids Research 45: D566–D573. https://doi.org/10.1093/nar/gkw1004

Johnning A, Kristiansson E, Angelin M, Marathe N, Shouche YS, Johansson A, Larsson DGJ (2015a) Quinolone resistance mutations in the faecal microbiota of Swedish travellers to India. BMC Microbiology 15: 235. https://doi.org/10.1186/s12866-015-0574-6

Johnning A, Kristiansson E, Fick J, Weijdegård B, Larsson DGJ (2015b) Resistance mutations in *gyrA* and *parC* are common in *Escherichia* communities of both fluoroquinolone-polluted and uncontaminated aquatic environments. Frontiers in Microbiology 6:1355. https://doi.org/10.3389/fmicb.2015.01355

Johnning A, Moore ERB, Svensson-Stadler L, Shouche YS, Larsson DGJ, Kristiansson E (2013) Acquired genetic mechanisms of a multiresistant bacterium isolated from a treatment plant receiving wastewater from antibiotic production. Applied and Environmental Microbiology 79: 7256–7263. https://doi.org/10.1128/AEM.02141-13

Jonsson V, Österlund T, Nerman O, Kristiansson E (2017) Variability in metagenomic count data and its influence on the identification of differentially abundant genes. Journal of Computational Biology 24: 311–326. https://doi.org/10.1089/cmb.2016.0180

Jonsson V, Österlund T, Nerman O, Kristiansson E (2016) Statistical evaluation of methods for identification of differentially abundant genes in comparative metagenomics. BMC Genomics 17: 78. https://doi.org/10.1186/s12864-016-2386-y

Kraupner N, Ebmeyer S, Bengtsson-Palme J, Fick J, Kristiansson E, Flach CF, Larsson DGJ (2018) Selective concentration for ciprofloxacin resistance in *Escherichia coli* grown in complex aquatic bacterial biofilms. Environment International 116: 255–268. https://doi.org/10.1016/j.envint.2018.04.029

Kristiansson E, Fick J, Janzon A, Grabic R, Rutgersson C, Weijdegård B, Söderström H, Larsson DGJ (2011) Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. PLoS ONE 6: e17038. https://doi.org/10.1371/journal.pone.0017038

Lundström SV, Östman M, Bengtsson-Palme J, Rutgersson C, Thoudal M, Sircar T, Blanck H, Eriksson KM, Tysklind M, Flach CF, Larsson DGJ (2016) Minimal selective concentrations of tetracycline in complex aquatic bacterial biofilms. Science of the Total Environment 553: 587–595. https://doi.org/10.1016/j.scitotenv.2016.02.103

Luo C, Knight R, Siljander H, Knip M, Xavier RJ, Gevers D (2015) ConStrains identifies microbial strains in metagenomic datasets. Nature Biotechnology 33: 1045–1052. https://doi.org/10.1038/nbt.3319

Ma L, Li B, Zhang T (2014) Abundant rifampin resistance genes and significant correlations of antibiotic resistance genes and plasmids in various environments revealed by metagenomic analysis. Applied Microbiology and Biotechnology 98: 5195–5204. https://doi.org/10.1007/s00253-014-5511-3

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS (2016) An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. Genome Research 26: 1612–1625. https://doi.org/10.1101/gr.201863.115

Österlund T, Jonsson V, Kristiansson E (2017) HirBin: high-resolution identification of differentially abundant functions in metagenomes. BMC Genomics 18: 316. https://doi.org/10.1186/s12864-017-3686-6

Pal C, Bengtsson-Palme J, Kristiansson E, Larsson DGJ (2016) The structure and diversity of human, animal and environmental resistomes. Microbiome 4: 54. https://doi.org/10.1186/s40168-016-0199-5

Quince C, Walker AW, Simpson JT, Loman NJ, Segata N (2017) Shotgun metagenomics, from sampling to analysis. Nature Biotechnology 35:833–844. https://doi.org/10.1038/nbt.3935

R Core Team (2016) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S; Tara Oceans coordinators, Bowler C, de Vargas C, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P (2015) Ocean plankton. Structure and function of the global ocean microbiome. Science 348: 1261359. https://doi.org/10.1126/science.1261359

Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N (2017) Microbial strain-level population structure and genetic diversity from metagenomes. Genome Research 27: 626–638. https://doi.org/10.1101/gr.216242.116

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biology 5: e16. https://doi.org/10.1371/journal.pbio.0050016

Supplementary material 1
Table S1

Authors: Shruthi Magesh, Viktor Jonsson, Johan Bengtsson-Palme

Link: https://doi.org/10.3897/mbmg.3.36236.suppl1

Supplementary material 2
Figure S1

Authors: Shruthi Magesh, Viktor Jonsson, Johan Bengtsson-Palme

Link: https://doi.org/10.3897/mbmg.3.36236.suppl2