Background reading

Microbiota Analysis in R Easily: mare

Katri Korpela

The aim of this package is to offer non-bioinformaticians easy access to microbiota analysis, while ensuring meaningful results. The package utilizes USEARCH (Edgar 2010, Edgar 2013) for many of the pre-processing steps and several R packages (vegan, sp, gstat, ggplot2, MASS, glmmADMB, beanplot, strinr, seqinr, Biostrings, ShortRead, BiocGenerics) for the statistical analysis and visualisation. USEARCH has been shown to perform many of the pre-processing steps better than or equally well as other programs (Smith et al. 2012, Callahan et al. 2015, Edgar and Flyvbjerg 2015, Flynn et al. 2015), and was therefore selected as the basis for the package.

Aims of mare

The aim of the pipeline development was to streamline the process of 16S sequence analysis, allowing the user to run the whole process in R. The different pre-processing, analysis, and visualisation steps are wrapped into a few simple-to-use functions, that require no previous programming or R experience. The pipeline takes the sequencing reads as they come from the sequencing facility, creates taxonomic tables, visualises the results, and finally identifies organisms significantly associated with variables of interest.

Pre-processing affects the accuracy of the resulting data

The performance of mare and optimization of the process were tested using artificial microbial communities of known composition, as well as real faecal samples ranging from prematurely born infants (Blakstad et al. in preparation) to healthy adults (Rasinkangas et al. in preparation). To validate that the pipeline produces are least as good results as a popular alternative, its performance was compared to that of QIIME (Caporaso et al. 2010). Analysing the effect of supplementation with different lactobacillus strains on the microbiota of adults, we found for the first time a clear shift in the microbiota composition (Rasinkangas et al. in preparation), indicating that the package is sensitive enough to discover treatment effect that other methods have failed to observe. The pipeline also unveiled a clear developmental pattern in the intestinal microbiota of preterm neonates (Blakstad et al. in preparation).

When analyzing the mock communities, it became evident that different pre-processing steps yield different results, and that the optimal solution is to limit the number of pre-processing steps to a minimum (Fig. 1). However, the package is flexible and accommodates various options in the preprocessing and data analysis. Compared to the performance of a commonly used 16S sequence analysis platform, QIIME, all of the different USEARCH-based options available in mare gave more realistic results in terms of estimation of microbial richness and composition (Fig. 1). In mare, the main source of differences in the results was the read length (Fig. 1).

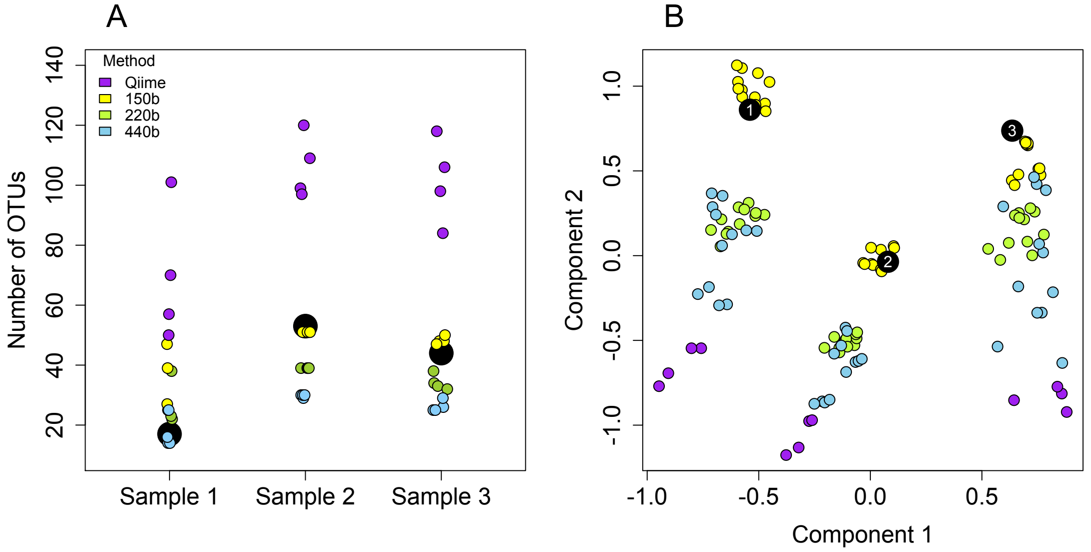

Figure 1. Validation of mare. Four duplicates of the three different mock-communities (two using the 1-step PCR protocol, and two with the 2-step protocol) were analysed using different options in mare (reads truncated to different lengths, quality-filtered or not, and taxonomy annotation based on OTUs or the reads themselves), as well as the standard options in QIIME. Here, only the effect of read length (150, 220, and 440 bases) in mare is shown and compared to the results from QIIME. A) Expected (black symbols) and observed (coloured symbols) species richness (measured as OTU counts). B) Overall microbiota composition visualized as principal coordinates analysis, using Bray-Curtis dissimilarities. The black symbols represent the expected composition and the coloured symbols the observed.

Read length

It is commonly thought that longer reads should be preferred as they have better resolution than short reads, and therefore capture more of the diversity. However, many paired-end merging tools perform poorly (Edgar and Flyvbjerg 2015). Furthermore, the longer the read, the more errors it will contain, which compromises the accuracy of OTU clustering and taxonomic annotation. Merging results in the loss of data, as many reads cannot be merged due to low read quality in the overlapping region. If certain phylogenetic groups are more prevalent in the non-merged reads, the result is phylogenetically biased data. As it is evident that non-merged forward reads, truncated to 150 or 220 bases, reproduce the original composition better than merged reads (Fig. 1), it may be advisable to skip the paired-end read merging altogether, and use only the forward reads.

If one chooses not to use paired-end reads but only the forward reads, the decision to be made is how long the reads should be. The error-rate increases towards the end of the read, so truncating the reads at a certain length may be wise. Read quality usually starts to drop at 150 bases, and therefore I chose to test the performance of the 150b reads. The 150b reads performed clearly better than the 220b reads, and achieved very good correspondence to the expected microbiota. The results indicate that reads longer than 150b are not necessary.

Filtering out errors

The next step in 16S sequence processing is usually quality filtering based on the quality scores. Again, this step may causes biases, depending on how the filtering algorithm (Brannock and Halanych 2015) and on the reliability of the quality scores. It has been noted that most paired-end read merging tools miscalculate the quality scores for the merged reads, which introduces an additional source of error (Edgar and Flyvbjerg 2015). Instead of quality filtering based on the quality scores, the filtering can be done using the abundances of unique reads: reads that occur often in the dataset are likely to be free of errors, while rare reads are more likely to be the result of errors. In Fig. 1, the quality-filtered vs. not-filtered reads all cluster together, depending on read length, indicating that the filtering-by-quality-score is an unnecessary step, as long as filtering-by-prevalence is done, which is the default in mare.

Usually reads occurring only once in the data set are removed. However, testing different cut-offs for read prevalence in the mock communities as well as real faecal samples, it became clear that 1 may be too lenient of a cut-off. To obtain a realistic estimate of the microbial richness, cut-off between 10 and 100, or ca. 0.01% or all reads, was optimal according to the mock community data. In practice, the conservative cut-off of 100 has produced meaningful results and is therefore the default in mare. Some rare taxa may be missed, and when interested in the rare taxa, it may be worth trying different cut-offs, while bearing in mind the increased chance of errors in the data. Rare taxa are problematic to analyse, as their observed abundance and presence or absence is likely to depend strongly on chance, and have little biological relevance.

OTU clustering and taxonomic annotation

After the reads have been processed, they are usually clustered to form operational taxonomic units (OTUs), which is another serious and unnecessary source of errors (Tikhonov et al. 2015). Different OTU clustering methods may produce different results by clustering together reads from unrelated bacteria, or failing to cluster related bacteria (Flynn et al. 2015). The clustering method in USEARCH has been shown to produce good results (Flynn et al. 2015) and can be used to reliably estimate the number of species-level phylotypes in the samples, provided that rare reads are removed. However, to avoid the potential bias in taxonomic annotation caused by OTU clustering, it may be advisable to perform taxonomic annotation directly on the reads themselves, not the OTUs. This does not take a prohibitively long time and eliminates another potential source of error.

Taxonomic annotation and the choice of the reference database can introduce further errors. The taxonomic annotation method of USEARCH produces reliable results (Fig. 1), when used in combination with a reliable database, such as RDP or Silva, which are compatible with mare. When analysing human intestinal samples, it is wise to limit the reference database to taxa known to inhabit the human intestine. In mare, it is easy to create a human gut-associated database. The taxonomic annotation in mare is conducted at several phylogenetic levels, up to the species/strain level, depending on the taxonomic resolution in the reference database. It is also possible to perform taxonomic annotation using BLAST in mare.

Statistical testing

The reliability of the results strongly depends on the accuracy of the statistical tests used for inference of significance. Using analysis of variance or t-tests is not suitable for sequencing data, as these tests assume that the response variable is measured on a continuous scale and normally distributed, which the bacterial read counts are not. The bacterial abundances in sequencing data are measured as counts of reads assigned to a specific taxon: they are not continuous nor are they normally distributed. A further complication arises from the fact that different samples are sequenced at different depths: the counts of reads per taxon are therefore not directly comparable between samples. An often-used but sub-optimal option is to randomly take a certain number of reads from each sample, which results in the unnecessary loss of data. The rarefied read counts are then often log-transformed to obtain normal distributions, which is a dramatic and unnecessary transformation of data. Alternatively, non-parametric test are used, which is also unnecessary, as parametric models exist for count data. A better option is to analyse the counts as they are, without unnecessary transformations, which may bias the results. Read counts can be analysed using generalized linear models, assuming the Poisson or negative binomial distribution. The data are often over-dispersed so negative binomial is a safer general choice. To account for the varying sequencing depth, the number of reads per sample should be included in the model as an offset. No data transformations are required, nor is rarefying, which should be avoided (McMurdie and Holmes 2014). In mare, all of these considerations are automatically implemented in the statistical testing, and the user does not need to define which type of model to use. In mare, the basic statistical inference is conducted utilized the glm.nb function in the R package MASS (Venables and Ripley 2002) for identification of bacterial taxa significantly associated with a given variable of interest, allowing simultaneously the inclusion of several potential confounding variables in the models.

Many of the bacterial taxa are often observed only in a few samples, meaning that their counts may be zero-inflated. It is possible to analyse such data using zero-inflated negative binomial models. In mare, such models are available, utilizing the implementation in the glmmADMB package (Fournier et al. 2012). However, the observed abundance or presence/absence of the rare taxa are likely to be affected strongly by measurement error or chance, and their counts may not be very unreliable or biologically informative. A simple solution is to limit the analysis to taxa, which are observed in a sufficient proportion of samples, eliminating the problem of zero inflation. In mare, both the zero-inflated option and the limitation to common taxa are very easy to specify, without the user needing to construct complicated models.

All statistical tests assume that the samples are independent of each other, unless stated otherwise. Often several samples are taken from the same individuals, from related individuals, or e.g. from mice housed in the same cage. These instances clash against the assumption of independence and should be taken into account by using mixed effects models with the e.g. subject ID or family ID as a random factor. In mare, the user does not need to know which model to use in which occasion, but simply specifies the subject ID in the function, which then utilizes generalized linear mixed effects models, implemented in the glmmADMB package for the inclusion of random factors.

All too often researchers rely solely on p-values for inference of biological significance (Halsey et al. 2015). It is enticing to make interpretations based on statistically significant results, but one should always look at the data critically and determine whether the pattern is sufficiently clear, not caused by outliers with extreme values, and biologically realistic, before drawing conclusions. While mare provides p-values and calculates the multiple testing corrections for them, it also automatically visualises the results and allows for the dampening of outlier effects to enable critical evaluation of the results.

Principal co-ordinates (PCoA) analysis is often used to visualise the overall pattern microbial community composition among the samples. The choice of distance measure used to perform the analysis is not a trivial question (Faith et al. 1987). Common practice is to use Euclidean distances on log-transformed bacterial abundances, i.e. to use Principal Components Analysis (PCA). Theoretically, this is not a sound solution, as it requires continuous, normally distributed data (also in the multivariate space) and assumes that the abundances of the bacterial taxa are linearly associated with the underlying components. The former assumptions are not met by count data, and the latter assumption is also unlikely to be met by any ecological survey data. Theoretically, all species have optimal environmental condition, above and below which their growth and survival is reduced. This implies that organisms are unlikely to exhibit linear relations with environmental variables – PCA is thus not likely produce a realistic impression of the data, and has been shown to perform poorly with ecological data (Faith et al. 1987). Instead, PCoA based on the so-called Bray-Curtis distance measure is considered valid for ecological data and therefore utilized in this package (using the capscale function in the package vegan, (Oksanen et al. 2013)). In addition to the basic PCoA, the package estimates the effect of a given grouping variable on the overall microbiota composition, using the adonis function in package vegan. Furthermore, there is the possibility to visualise the association between the overall microbiota composition and a continuous variable by colouring the background of the PCoA-space using distance-weighted interpolation implemented in the package gstat (Pebesma 2004).

References

Brannock, P. M., and K. M. Halanych. 2015. Meiofaunal community analysis by high-throughput sequencing: Comparison of extraction, quality filtering, and clustering methods. Marine genomics 23:67-75.

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson, and S. P. Holmes. 2015. DADA2: High resolution sample inference from amplicon data. bioRxiv :024034.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Pena, J. K. Goodrich, and J. I. Gordon. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature methods 7:335-336.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature methods 10:996-998.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics (Oxford, England) 26:2460-2461.

Edgar, R. C., and H. Flyvbjerg. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics (Oxford, England) 31:3476-3482.

Faith, D. P., P. R. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57-68.

Flynn, J. M., E. A. Brown, F. J. Chain, H. J. MacIsaac, and M. E. Cristescu. 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. Ecology and evolution 5:2252-2266.

Fournier, D., H. Skaug, J. Ancheta, J. Ianelli, A. Magnusson, M. Maunder, A. Nielsen, and J. Sibert. 2012. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optim. Methods Softw. 27:233-249.

Halsey, L. G., D. Curran-Everett, S. L. Vowler, and G. B. Drummond. 2015. The fickle P value generates irreproducible results. nature methods 12:179-185.

McMurdie, P. J., and S. Holmes. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS computational biology 10:e1003531.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, and H. Wagner. 2013. vegan: Community Ecology Package. R package version 2.0-6. :.

Pebesma, E. J. 2004. Multivariable geostatistics in S: the gstat package. Computers & Geosciences :683-691.

Smith, B. C., T. McAndrew, Z. Chen, A. Harari, D. M. Barris, S. Viswanathan, A. C. Rodriguez, P. Castle, R. Herrero, and M. Schiffman. 2012. The cervical microbiome over 7 years and a comparison of methodologies for its characterization. PloS one 7:e40425.

Tikhonov, M., R. W. Leach, and N. S. Wingreen. 2015. Interpreting 16S metagenomic data without clustering to achieve sub-OTU resolution. The ISME journal 9:68-80.

Venables, W., and B. Ripley. 2002. Modern Applied Statistics with S. Springer, New York.