SOFTWARE TOOL ARTICLE

Streamlined processing and analysis of 16S rRNA amplicon sequencing data with OCMS_16S and OCMSlooksy [version 1; peer review: awaiting peer review]

Sandi Yen, Jethro Johnson, Nicholas E. Ilott

Oxford Centre for Microbiome Studies, Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics, Rheumatology, and Musculoskeletal Sciences, University of Oxford, Oxford, OX3 7FY, UK

Abstract

16S rRNA gene sequencing is a cost-effective method for profiling the bacterial component of a microbiome. Nevertheless, processing and analysis of the resulting sequencing data is often constrained by the availability of dedicated bioinformaticians - creating a bottleneck for biological interpretation. Multiple visualisation and analysis tools now exist for downstream analysis of 16S rRNA data. These tools are designed with biological scientists in mind and therefore consist of a graphical user interface that interacts with taxonomic counts tables to perform tasks such as alpha- and beta-diversity analysis and differential abundance. However, generating the input to these applications still relies on bioinformatics experience, creating a disconnect between data processing and data analysis. We aimed to bridge the gap between data processing and data analysis. To do this we have created two tools - OCMS_16S and OCMSlooksy - that perform data processing and data visualisation/analysis, respectively. OCMS_16S is a cgat-core based pipeline that wraps DADA2 functionality in order to facilitate processing of raw sequence reads into tables of amplicon sequence variant (ASV) counts using a simple command line interface. OCMSlooksy is an RShiny application that takes an OCMS_16S-generated SQLite database as input to facilitate data exploration and analysis. Combining these tools provides a simple, user-friendly workflow to facilitate 16S rRNA gene amplicon sequencing data analysis from raw reads to results.

Keywords

Microbiome, data processing, data exploration, analysis, pipeline
Introduction

16S rRNA gene amplicon sequencing remains a cost-effective and efficient method for profiling the bacterial component of a microbiome. Bioinformatic methods for processing raw sequencing reads are mature and, while they may differ in their conceptual framework, will provide counts of features (either unique sequences or sequence clusters) representing distinct bacterial taxa that can be used for downstream statistical analysis. The processing steps can be considered as a precursor to statistical analysis and visualisation. While these steps can be run on a standard desktop computer, they still require a level of expertise in coding. For example, DADA2\(^1\), which processes sequenced reads into taxonomically annotated amplicon sequence variants (ASVs), requires a familiarity with R in order to run the various steps of the pipeline.

Once processed, feature (i.e. ASV) abundance tables can be used to perform a wide variety of analyses that are common in microbiome research, including alpha- and beta-diversity as well as differential abundance. Multiple analysis and visualisation tools have been developed, aimed at creating a user-friendly interface to enable researchers with limited bioinformatics expertise to interrogate and interpret their processed data. These include RShiny-based applications such as Shiny-phyloseq\(^2\) and MicrobiomeExplorer\(^3\) and web applications such as MicrobiomeAnalyst\(^4\). While these tools provide a comprehensive suite of analysis and visualisation methods, there remains a disconnect between the processing steps and data analysis.

We have developed OCMS\(_{16S}\) and OCMSlooksy in order to narrow the gap between data processing and data analysis/visualisation. OCMS\(_{16S}\) is a cgit-core\(^5\) based workflow that wraps DADA2 functionality in a simple to use command line interface (CLI), similar to the previously developed tool, Dadaist2\(^6\). The main output from OCMS\(_{16S}\) is an SQLite database that serves as input to OCMSlooksy – an R/Shiny application that enables users to perform data analysis and visualisation. A major advantage of using OCMS\(_{16S}\) in combination with OCMSlooksy is the synchronous development and maintenance of the tools. Unlike existing tools, OCMSlooksy enables users to inspect parameters and visualise data across the processing steps implemented in OCMS\(_{16S}\), allowing data analysis and interpretation to be performed in the context of processing performance. In our experience, this is particularly useful when processing and analysis/visualisation steps are performed by different individuals.

Methods

Implementation

**OCMS\(_{16S}\)**

OCMS\(_{16S}\) is a cgit-core\(^5\) based pipeline that is designed to run DADA2 for processing short reads derived from Illumina 16S rRNA gene amplicon sequencing. Ultimately, the output is a set of denoised amplicon sequence variants (ASVs) and their counts. The pipeline consists of multiple R scripts that wrap DADA2 functionality and are executable from the command line. OCMS\(_{16S}\) provides workflow management that enables these scripts to be executed in the required order using a single command line statement. The pipeline can be run on a single unix computer, or on a computer cluster with a Distributed Resource Management Application API (DRMAA) compatible queue management system. Dependencies for OCMS\(_{16S}\) can either be managed through the creation of a conda environment or manual installation. OCMS\(_{16S}\) can be installed using pip or through cloning of the repository (https://github.com/OxfordCMS/OCMS\(_{16S}\)). Archived source code can also be found at Zenodo\(^7\). Details on how to install and run the pipeline as well as in-depth tutorials are available online at readthedocs.

OCMS\(_{16S}\) creates an SQLite database that is designed as input into OCMSlooksy used for downstream analysis and visualisation.

**OCMSlooksy**

OCMSlooksy is an RShiny application dedicated for downstream data visualization and analysis of microbiome data. The app is designed for researchers to interactively explore their 16S amplicon sequencing data through analysis methods commonly used in microbiome studies. As such, the app is organized into analysis modules (Microbiome Profile, Alpha-Diversity, Beta-Diversity, Differential Abundance and Pairwise Feature Comparison), each of which contains one or several related analyses that fall into the respective module’s theme. The analysis modules wrap R packages commonly used in microbiome data analysis (e.g. vegan\(^8\), DESeq2\(^9\), propr\(^10\)) and display analysis results in interactive tables and figures. Table 1 outlines the analyses available in each module. All analysis results can be exported as individual tables or figures. At the same time, all results from an analysis module, along with corresponding data preparation steps can be saved as a downloadable report.

The source code is available at GitHub and is archived with Zenodo\(^11\).

**Operation**

**OCMS\(_{16S}\)**

OCMS\(_{16S}\) requires python 3.6 or later and runs on unix-based systems. Software dependencies can be installed via conda or manually (details at https://ocms-16s.readthedocs.io/en/latest/dada2.html#installation). Databases required for DADA2 to perform taxonomic assignment are available at https://benjineb.github.io/dada2/training.html.

The pipeline runs the standard DADA2 workflow (Figure 1A), taking parameters to the various DADA2 functions from an editable YAML configuration file (pipeline.yml). The configuration file is created by running:

```bash
ocms_16s dada2 config
```

Once edited with desired parameters, to run the pipeline from start to finish i.e. raw reads to annotated ASV counts, consists of a single command line statement:

```bash
ocms_16s dada2 make full -v5 -p80
```
In this example, OCSM_16S DADA2 is processing the data through submitting 80 jobs to the cluster (-p80) meaning that 80 samples will be processed in parallel. If you do not have access to a cluster, you can also run it locally using the --local flag. A pipeline.log file is automatically created when the pipeline is executed so you can track the progress of the pipeline.

To create the SQLite database for downstream analysis in OCMSlooksy simply type:

ocms_16s dada2 make build_db

This will create the database in the current working directory with a name that corresponds to the given name in pipeline.yml.

**OCMSlooksy**

OCMSlooksy is available to download as an R package to run locally using R version 3.6 or later. Once the OCMSlooksy R package is installed, the app is run from within R:

OCMSlooksy::run_app()

In-depth instructions for using OCMSlooksy can be found on its wiki and an overview of the app’s workflow is shown in Figure 1B. Getting started with the app requires an SQLite database and a metadata file containing sample information. 16S rRNA gene amplicon sequencing data generated from sequence processing pipelines other than OCMS_16S can still be analysed with OCMSlooksy, using the ‘OCMSlooksy::create_db’ function to create the required SQLite database file from count and taxonomy tables. If data processing was performed by OCMS_16S, a visual summary of sequence processing is available under QC Report.

The first step in performing analysis is setting which samples to include in all subsequent analyses. One of the goals of the app is to be transparent on the processing steps applied to 16S rRNA gene amplicon data. As such, data preparation steps are executed within each analysis module, and any data manipulation (e.g. feature filtering, count normalization/transformation) occurring within an analysis module is independent and isolated from other analysis modules. As seen in Figure 1B, analysis modules provide different analysis tools to explore the data and make different types of statistical analyses.

### Use case

We present a previously published gut microbiome data of dextran sodium sulphate (DSS)-induced colitis in wild-type (WT) and gelatinase B/matrix metalloproteinase-9 knock-out (MMP-9 KO) mice, as the use case. This study reflects a 2 × 2 experimental design commonly used in microbiome studies: two treatment conditions (control and DSS) across two host genotypes (WT and MMP-9KO). Furthermore, we use this data set to demonstrate how OCMS_16S and OCMSlooksy can be used in conjunction to bring a researcher from raw reads to exploratory analysis and hypothesis testing while showing the practicalities of coping with an imperfect data set during bioinformatic analysis. The use case data is already pre-loaded onto the app as the “example data” should users want to try out the app without going through OCMS_16S first.

Table 1. Summary of analyses available in OCMSlooksy, as organised into analysis modules.

| Analysis Module                | Analysis                                      | Description                                                                 |
|--------------------------------|-----------------------------------------------|----------------------------------------------------------------------------|
| Microbiome Profile            | Microbiome profiles                           | Bar plot microbiome profiles of samples or groups of samples                |
|                                | Profile Sparsity                              | Visualize microbiome profiles in terms of presence/absence of taxa          |
| Alpha-diversity               | Alpha-diversity                               | Richness, evenness, and diversity index measures with statistical testing  |
| Beta-diversity                | Sample dissimilarity                          | Measures sample dissimilarity using Bray-Curtis distances, with statistical testing when applicable |
| Principal Coordinate Analysis (PCoA) | Non-supervised multivariate analysis that uses sample distances to visualize samples as scores on a two-dimensional plot |
| Principal Component analysis (PCA) | Non-supervised multivariate analysis to visualize samples as scores and features as loadings on a two-dimensional plot |
| Permutational Multivariate Analysis of Variance (PERMANOVA) | Partitions total variation as a response to one or more experiment factors, with statistical testing |
| Differential abundance        | Differential abundance with DESeq2⁹           | Applies DESeq2, a method for differential analysis to detect fold changes in count data when comparing groups of samples. Limited to comparing up to two experiment factors, each with at least two levels, and their interaction term. |
| Pairwise feature comparison   | Pairwise feature comparison                   | Compares features in a pairwise fashion, measuring their association with proportionality metric, rho¹⁰. |

Page 4 of 11
Processing sequences with OCMS_16S

A tutorial on downloading and processing the sequences in this use case is provided at readthedocs. Briefly the operational steps are provided below. This assumes you have installed OCMS_16S and its dependencies.

```bash
mkdir ocms_16s_tutorial
cd ocms_16s_tutorial
bash <path-to-OCMS_16S/tutorial/download_files.sh>

# fetch DADA2 reference databases
wget https://zenodo.org/record/2541239/files/RefSeq-RDP16S_v2_May2018.fa.gz
wget https://zenodo.org/record/2658728/files/RefSeq-RDP_dada2_assignment_species.fa.gz

# Run pipeline
ocms_16s dada2 config

At this point you need to edit the pipeline.yml (see https://ocms-16s.readthedocs.io/en/latest/tutorial.html).

ocms_16s dada2 make full -v5 -p40

Once the pipeline is completed you can build the database for input into OCMSlooksy.

ocms_16s dada2 make build_db
```

Figure 1. OCMS 16S rRNA amplicon sequencing analysis workflow. A) Workflow of OCMS_16S, pipeline for sequences processing. B) Workflow for OCMSlooksy app for exploratory analysis of microbiome data.
Exploring data with OCMSlooksy

Visualizing results of DADA2 processing

In contrast to existing data analysis and visualisation applications, results of sequence processing are propagated from the processing pipeline (OCMS_16S) to OCMSlooksy. Therefore, OCMSlooksy can be used to view both the parameters that were passed to DADA2 as well as the result of each step of the DADA2 workflow. The Trim and Filter, and Denoise Sequences sections show how many reads remain in each sample after the respective steps. In this case, sample read depths range from ~16,000 to ~51,000 reads (Figure 2A) and Sequence Rarefaction confirms that the sequence depth is adequate i.e. maximum available richness is reached by ~10,000 reads (Figure 2B). One way of assessing processed sequences includes visualizing the number of ASVs observed in each sample (Figure 2C) where the number of ASVs per sample to be expected depends on the nature of the sample. The number of samples a given ASV is observed in (i.e. prevalence) is another useful measure to assess the data at this stage. When visualised as the number of ASVs observed with increasing prevalence (Figure 2D), microbiome data often forms an exponential decay. This shows that few ASVs are common (high prevalence) while many ASVs are considered rare, and only observed in a small handful of samples. OCMSlooksy also displays ASV prevalence against mean relative abundance, giving insight into the

Figure 2. Quality control measures reporting on performance of sequence processing. (A) Number of reads remaining after Trim and Filter, and Denoise Sequence steps in the sequence processing pipeline. (B) Sequence rarefaction curve show samples all contain adequate read depth achieved for all samples. (C) Number of amplicon sequence variants (ASVs) observed in each sample. (D) Distribution of ASV prevalence, which is the number of samples a given ASV is observed in. Low ASV prevalence indicates the ASV is rare, while high ASV prevalence indicates the ASV is observed in many samples. (E) Distribution of ASV prevalence as a function of mean relative abundance shows that rare ASVs tend to make up a very small proportion of the microbiome profile. On the other hand, frequently observed ASVs tend to also be the most abundant.
proportion of the microbiome represented by a given ASV (Figure 2E). Figure 2E shows that the most prevalent ASVs also tend to be the most dominant members of the microbiome in terms of relative abundance.

**Identifying problematic samples from their microbiome profiles**

Once the processing steps have been assessed, OCMStoolksy provides an opportunity to filter samples (Filter Samples step), which sets the samples to be included in subsequent analyses. Reasons to subset the data for analysis could be due to sample quality or it could be for performing analysis directed towards specific research questions. In our first pass of analysis when we are taking a look-see at the data, we include all samples in our analysis.

The Microbiome Profiles analysis module provides tools to visualize the microbiome on a per-sample basis. Looking at the phylum level, we see that 15 of the 40 samples predominantly consist of an unclassified phylum (Bacteria_unclassified) (Figure 3A and 3B). This indicates that these samples may contain little information and/or sequences generated from spurious (i.e. chimeric or off-target) amplicons. Given that these observations are likely due to technical issues rather than biological characteristics, we have justifiable reason to go back

![Microbiome profile tools highlight subset of samples that can be justifiably excluded from analysis](image)

**Figure 3.** Microbiome profile tools highlight subset of samples that can be justifiably excluded from analysis. Amplicon sequence variants (ASVs) were aggregated at the phylum level and filtered for phyla with a minimum of 1 read count and 2 sample prevalence. (A) Relative abundance profiles at the phylum level show 15 of 40 samples are ~80–90% unclassified phylum. Problematic samples are highlighted with black arrows. (B) Presence-absence heatmap of phyla in each sample (1=present and 0 = absent) show subset of samples (highlighted with red rectangle) only contains four phyla, three of which are considered unclassified at the phylum level.
to the Filter Sample module and exclude these samples. One of the advantages of using the app as a data visualization tool is that problem-samples are readily apparent and can be removed early in the data analysis process.

**Characterising diversity**
Once problematic samples have been removed, we can assess the community level characteristics of the samples in respective analysis modules. The Alpha-Diversity analysis module calculates the several alpha-diversity metrics for each sample and provides a figure to compare alpha-diversity across sample groups with statistical testing (Figure 4A). Using this analysis module, we visualize the distribution of microbiota richness across each treatment and genotype group. The Beta-Diversity analysis modules includes several analysis tools, including PCoA, which facilitate multivariate analysis. Plotting parameters are customizable to highlight experimental variables and provide researchers with some control over how the data is visualized (Figure 4B).

**Assessing Differential Abundance**
The Differential Abundance analysis module employs DESeq2 to identify taxa that are differentially abundant across sample groups. For simplicity, the app limits model design to comparing two groups in up to two experimental variables, with an optional interaction term. The taxa that are differentially abundant for each contrast are displayed in a volcano plot, from which ASVs can be selected to display the abundance levels

---

**Figure 4. Characterizing microbiome diversity.** (A) Alpha-diversity analysis module calculates alpha-diversity metrics, such as richness, evenness and Shannon's Diversity Index, with comparison using Kruskal-Wallis test. Samples that were predominantly bacteria_unclassified phyla were excluded from this analysis. Analysis was performed at the amplicon sequence variant (ASV) level and all ASVs were included in the analysis. (B) Beta-diversity analysis module includes PCoA using Bray-Curtis distances. This analysis tool is used to illustrate that sample microbiota cluster by treatment method, with minimal influence from the host genotype. Samples that were predominantly bacteria_unclassified phyla were excluded from this analysis. ASVs were aggregated at the genus level and filtered with count cut-off of 1 and sample prevalence of 2. DSS=dextran sodium sulphate; WT=Wildtype; MMP-9 KO=gelatinase B/matrix metalloproteinase-9 knock-out.
across groups (Figure 5). Alternatively, significant ASVs, as determined by the cut-off thresholds set, are plotted according to their log2-fold change (Figure 6).

**Conclusion**

OCMS_16S and OCMSlooksy are sister tools that aim to alleviate the bioinformatics bottleneck in microbiome studies. Several bioinformatics tools exist in this space, but often there is still a disconnect between processing sequences and data analysis. OCMS_16S wraps DADA2 functionality that is executed in a single command line statement following the same philosophy as Dadaist2. However, as OCMS_16S and OCMSlooksy have been developed in parallel, more information is available to the downstream visualisation tool than currently available applications. For example, OCMSlooksy allows users to explore DADA2 processing parameters and results of processing steps. In our experience this is extremely useful in collaborative projects where one individual processes the data and another picks up the processed data for downstream analysis. As an RShiny app, OCMSlooksy presents microbiome data in an interactive platform that allows researchers to explore the data while performing analyses often employed in microbiome studies. While designed to take the output from OCMS_16S, OCMSlooksy can be used independently from OCMS_16S as input files for the app can be generated with helper functions included in the R package. The analysis tools implemented in

---

**Figure 5.** Screenshot of the Differential Abundance analysis module, which employs DESeq2 to identify differentially abundant taxa in a simple 2-level, 2-factor experiment design. In this case, the observational variable was set to Treatment, to compare control vs. dextran sodium sulphate (DSS)-treated mice, and Genotype was set as the covariate, to account for the Wildtype and MMP-9-/- genotypes used in the experiment. No interaction term was included in the model. Samples that were predominantly bacteria_unclassified phyla were excluded from this analysis. Amplicon sequence variants (ASVs) were aggregated at the genus level and filtered with count cut-off of 1 and sample prevalence of 2.
OCMSlooksy, while not exhaustive, provide an overview look at the microbiome such that early data exploration and analysis are accessible by researchers, regardless of their bioinformatic expertise. The aim with providing interactive data visualization tools is that technical or experimental issues are readily apparent, and the initial data exploration and analysis conducted with the app inform more directed and specialised analysis beyond the use of the app.

Reproducibility
A reproducible example of using OCMS_16S in the use case is available in the tutorial. The DADA2-formatted reference database used in the use case was [RDP trainset 16] and is available at Zenodo.

Data availability
Data presented in the use case were from the study by Magali de Bruyn et al.13. Raw sequencing reads were downloaded from the European Nucleotide Archive (PRJEB21337, http://www.ebi.ac.uk/ena/data/view/PRJEB21337) and associated sample metadata was downloaded from Supplementary File in 10.1038/s41522-018-0059-0.

Software availability
Source code for OCMS_16S and OCMSlooksy are available from: https://github.com/OxfordCMS.

Archived source code at time of publication:
OCMS_16S: https://doi.org/10.5281/zenodo.6066297.
OCMSlooksy: https://doi.org/10.5281/zenodo.582567011.

License: MIT

Documentation for OCMS_16S is available at: https://ocms-16s.readthedocs.io/en/latest/index.html.

Vignette for OCMSlooksy is available at: https://github.com/OxfordCMS/OCMSlooksy/wiki

References
1. Callahan BJ, McMurdie PJ, Rosen MJ, et al.: DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016; 13(7): 581–583. PubMed Abstract | Publisher Full Text | Free Full Text
2. McMurdie PJ, Holmes S: Shiny-phyloseq: Web application for interactive microbiome analysis with provenance tracking. Bioinformatics. 2015; 31(2): 282–283. PubMed Abstract | Publisher Full Text | Free Full Text
3. Reeder J, Huang M, Kaminker JS, et al.: MicrobiomeExplorer: an R package for
the analysis and visualization of microbial communities. Bioinformatics. 2021; 37(9): 1317–1318. PubMed Abstract | Publisher Full Text | Free Full Text

4. Dhariwal A, Chong J, Habib S, et al.: MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 2017; 45(W1):W180-W188. PubMed Abstract | Publisher Full Text | Free Full Text

5. Cribbs AP, Luna-Valero S, George C, et al.: CGAT-core: a python framework for building scalable, reproducible computational biology workflows (version 2; peer review: 1 approved, 1 approved with reservations). F1000Res. 2019; 8: 377. Publisher Full Text

6. Ansorge R, Birolo G, James SA, et al.: Dadaist2: A Toolkit to Automate and Simplify Statistical Analysis and Plotting of Metabarcoding Experiments. Int J Mol Sci. 2021; 22(10): 5309. PubMed Abstract | Publisher Full Text | Free Full Text

7. Ilott N: OxfordCMS/OCMS: (v0.0.2). Zenodo. 2021. http://www.doi.org/10.5281/zenodo.6066300

8. Oksanen J, Blanchet FG, Kindt R, et al.: vegan: Community Ecology Package. 2012. Reference Source

9. Love MI, Huber W, Anders S: Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014; 15(12): 550. PubMed Abstract | Publisher Full Text | Free Full Text

10. Quinn TP, Richardson MS, Lovell D, et al.: propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. Sci Rep. 2017; 7(1): 16252. PubMed Abstract | Publisher Full Text | Free Full Text

11. schyen, Ilott N: OxfordCMS/OCMSlooksy: First release of OCMSlooksy (v0.0.1). Zenodo. 2022. http://www.doi.org/10.5281/zenodo.5825670

12. Anderson MJ: Permutational Multivariate Analysis of Variance (PERMANOVA). In Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd, 2017; 1–15. Publisher Full Text

13. de Bruyn M, Sabino J, Vandeputte D, et al.: Comparisons of gut microbiota profiles in wild-type and gelatinase B/matrix metalloproteinase-9-deficient mice in acute DSS-induced colitis. NPJ Biofilms Microbiomes. 2018; 4: 18. PubMed Abstract | Publisher Full Text | Free Full Text