Automation Proposal for the Intermediate Steps in the 16S FFPE Samples Analysis Pipeline

Elsa Martin-De Arribas, Kelly Conde-Pérez, Pablo Aja-Macaya, Juan A. Vallejo, Margarita Poza, and Susana Ladra

Universidade da Coruña, Centro de Investigación CITIC, 15071 A Coruña, Spain
meiGAbiome, INIBIC-CICA - Universidade da Coruña - CIBERINFEC-ISCIII. Hospital Universitario, A Coruña, Spain.
Correspondence: elsa.mdearribas@udc.es

DOI: *https://doi.org/10.17979/spudc.000024.35*

Abstract:
In the day-to-day work of bioinformatics, the use of integrated software packages, which encompass a wide range of tools, enables the development of pipelines for omics data analysis. Within the various existing pipelines, we focus on the analysis of the 16S rRNA gene as it allows for the study of diversity and taxonomy of prokaryotic microorganisms such as Bacteria and Archaea. However, these pipelines often involve a sequence of multiple tools that require intermediate steps before further processing can proceed, as in the case between Cutadapt and DADA2. In fact, in a typical pipeline, the values for DADA2 input arguments 'trunc-len-f' and 'trunc-len-r' are extracted from the output of Cutadapt. The best approach for selecting optimal values (aka the trimming positions) is graphically visualizing Cutadapt output and manually selecting the most accurate trimming position length. Therefore, we propose the automation of this specific intermediate step between Cutadapt and DADA2 tools, by selecting values displayed in the graphs that meet the filtering criteria. This automation has been incorporated into a custom pipeline for the analysis of the microbiome in 16S paired-end samples from colorectal cancer patients, and could potentially serve as a standardization approach in these processes.

1 Introduction

In the realm of biological sciences, the study of microbial communities persists as a crucial element, with relevant spanning diverse fields, such as agriculture, marine environment, and medicine. While any microbiome-related field of study is of interest, our specific focus lies within the biomedical sphere, particularly in cancer research. This paper, in particular, centres on the examination of formalin-fixed paraffin-embedded (FFPE) microbiome samples from colorectal cancer (CRC) patients (Conde-Pérez et al., 2023). Sequencing of these samples was conducted using the 16S Amplicon Metagenomic Sequencing technique, a next-generation, high-throughput methodology that emphasizes amplifying specific, short, targeted regions known as amplicons. This methodology is a standard in molecular laboratories, supported by several dedicated bioinformatics tools and platforms for the analysis of 16S rRNA gene amplicon data (Straub et al., 2020).

One of the best-known and widely used examples is the QIIME2 platform (Bolyen et al., 2019), which stands for Quantitative Insights Into Microbial Ecology. Within this platform, numerous integrated tools are essential for such studies. In our case, we developed a specific pipeline for the analysis of these FFPE samples, adding a new qiime2-based implementation called Sidle (Debelius et al., 2021). Sidle was a Python version developed within the QIIME2 environment, based on a pre-existing algorithm originally created by Fuks et al. in 2018 and called Short MUltiple Reads Framework (SMURF) algorithm (Fuks et al., 2018). The original purpose was to enable the reconstruction of multiple short, fragmented amplicons against a known database, but Sidle adds a novel tree-building solution. This enhancement significantly improves the resolution of the reconstructed community compared to single amplicons.

Inside the Sidle pipeline, two pivotal tools play an important role in the analysis: Cutadapt and Divisive Amplicon Denoising Algorithm (DADA2). As described in its paper, Cutadapt finds and removes unwanted sequences from high-throughput sequencing reads, such as adapters, primers, poly-A tails, and more (Martin, 2011). Subsequently, DADA2 (Callahan et al., 2016) takes over, returning the reads that have been retained after quality filtering and merging steps. However, transitioning from one tool to another involves the manual selection of the positions of the forward and reverse sequences that will be trimmed by DADA2 for subsequent read processing. That manual selection requires incrementing the time of selecting and testing the best optimal positions to pass the filtering, merging, and removing chimeras sequences in the sample reads. For that reason, we propose a novel automatic selection and testing of optimal positions in the DADA2 tool.

2 Background

2.1 16S Amplicon Metagenomic Sequencing

Over the last decade, the ribosomal 16S rRNA gene has been established as a phylogenetic marker due to its ubiquitous presence among members of both Bacteria and Archaea domains. Its structural composition is characterized by a combination of "conserved" and "variable" regions: the conserved regions are universally shared across all microorganisms, while the variable regions serve as distinctive "tags", facilitating the precise taxonomic or phylogenetic classification of each microorganism. Consequently, universal primers are designed based on the adjacent conserved regions to the targeted variable regions, enabling differentiating the taxa present (Straub et al., 2020).

Hence, one of the best-known approaches in metagenomic studies is sequencing the 16s rRNA gene or the 18S rRNA gene. However, since the 18S rRNA gene is present in eukaryotic organisms and fungi, it will not be considered in this study. Although there are various types of next-generation sequencing, we will focus on amplicon sequencing. Amplicons are small fragments of DNA or RNA obtained through Polymerase Chain Reaction (PCR) or other techniques that yield multiple copies of these fragments. The PCR amplicon-based sequencing has been employed in multiple biological fields, from agriculture to tumour immunology (Straub et al., 2020).

In our study, we employ a set of five primers specifically designed for amplicon sequencing. These primers play a crucial role in PCR, where they target and amplify the 16S rRNA gene from the bacterial DNA in our samples. It generates paired-end sequences, which provide valuable information about the amplicon fragments at both ends. These fragments are distinguished thanks to our carefully selected primer sets, which enable us to differentiate and analyse the sequences generated from the forward and reverse strands of the target gene regions. This approach enhances our ability to study microbial communities with greater accuracy and depth.

2.2 Colorectal cancer, microbiome and FFPE samples

Colorectal cancer (CRC) ranks as the second leading cause of cancer-related death globally (Wang et al., 2012; Wong et al., 2019), with the highest incidence in Spain in 2023 with 42,721 cases, followed by breast cancer with 35,001 cases, and lung cancer with 31,282 cases (Sociedad Española Oncológica Médica (SEOM), 2023). Traditionally, CRC treatments include surgery and chemotherapy (sometimes combined with radiotherapy). However, recent developments focus on gut microbiome-targeted therapies due to the gut microbiota's role in CRC (Conde-Pérez et al., 2023; Wang et al., 2012; Wong et al., 2019).

The gut microbiota, residing in the gastrointestinal tract, influences energy metabolism, immunity, and carcinogenesis responses. Through functional studies, alterations in microbial composition and ecology have been revealed in individuals with CRC, pinpointing the roles of several bacteria in colorectal carcinogenesis, such as *Fusobacterium nucleatum*, certain *Escherichia coli* strains, and *Bacteroides fragilis* (Wang et al., 2012; Wong et al., 2019). Additionally, a recent study has proposed the translocation of particular bacteria, *Parvimonas micra* from the subgingival cavity to the gut (Conde-Pérez et al., 2023). This discovery is groundbreaking, as the data suggest that this bacteria may migrate as a part of a synergistic consortium with other periodontal bacteria.

Our research delves into the bacterial microbiota of CRC patients using the 16S amplicon sequencing, deepening our CRC understanding. The microbiota is extracted from the formalin-fixed paraffin-embedded (FFPE) samples, which entails preserving and processing colon or rectum tissue sections. FFPE's preservation protocol, while precise, affects DNA integrity during long-term storage, prompting the need for improved preservation methods (Guyard et al., 2017). These samples play a pivotal role in CRC diagnosis, characterization, and research through microbiota analysis, further advancing our CRC insights (Conde-Pérez et al., 2023).

2.3 Workflow

The principal aim of a bioinformatic pipeline is maximizing the reliability and confidence in community data analysis, which becomes particularly challenging when dealing with microbial communities. Among the studies comparing the existing technologies applied in bioinformatic pipelines, Straub's study assessed a popular list of analysis pipelines, i.e., Mothur, QIIME (version 1 and 2), and MEGAN against 24 datasets (Straub et al., 2020). QIIME2 was identified as the most accurate option due to its denoising capabilities (DADA2 and Deblur), surpassing OTU clustering methods used by the other pipelines. Despite the extra computational time required by DADA2, it effectively retains sequences below specific length thresholds and captures unique site-specific sequences, highlighting the critical importance of precise pipeline selection in microbiome research.

Taking all of these criteria into consideration, the FFPE samples pipeline was developed following the recommendations of the Sidle author (Debelius et al., 2021). In a nutshell, our pipeline workflow begins by importing FASTQ format samples into the QIIME2 environment. Given the unique characteristics of FFPE samples, where our objective is to detect the maximum number of microorganisms, we define five variable regions. Next, we demultiplex the samples into the five defined regions (in our case) using the Cutadapt tool, which leverages primers to differentiate and remove them from the sequences. Following this, manual trimming positions are selected based on Cutadapt's output for each region, which can be computationally intensive and time-consuming. These positions, defining parameters such as 'trunc-len-f' and 'trunc-len-r' for right-side trimming and 'trim-left-f' and 'trim-left-r' for left-side trimming, are then used in DADA2. Subsequently, DADA2 is executed, producing three output files for each region. One of these files contains statistics extracted for each sample during the tool's step, resulting in a final percentage of reads that have passed all selection criteria. The final steps involve aligning the region in order to reconstruct a portion of the original 16S rRNA targeted gene and detect the microorganisms present in the samples (Conde-Pérez et al., 2023).

2.4 Related studies

Upon conducting a comprehensive review of the literature related to this topic, a new tool called FIGARO was identified in the form of a preprint (Weinstein et al., 2019). Its primary objective is to optimize cut-off positions for maximizing read retention post-filtering, resembling DADA2 in approach. However, FIGARO's GitHub repository has not seen updates since its initial release, potentially causing installation challenges. Additionally, we found the nf-core/ampliseq framework (Ewels, *et al.*, 2020; Straub et al., 2020), a bioinformatics analysis pipeline. It supports denoising across various amplicons (16S, ITS, CO1, 18S), allows taxonomic assignment, and works with both paired-end and single-end data from multiple sequencing platforms. Notably, no instances of analysis with FFPE samples were found.

In a brief comparative benchmarking of these two options and the approach presented in this paper, it becomes evident that FIGARO (Weinstein et al., 2019) is not a readily applicable tool within the QIIME2 environment. Conversely, the rationale behind their optimal positions of trimming selection approach is well-founded. On the other hand, the nf-core offers the flexibility to develop a pipeline tailored to a specific sample type using Nextflow or one of the multiple pipelines shared by its contributors. However, it is important to note that no existing pipelines were identified that refer to the sample type utilized in our study, and proficiency in Nextflow is a prerequisite for its utilization (Ewels, *et al.*, 2020; Straub et al., 2020).

3 Proposal

The previous section emphasizes the underlying issue with FFPE samples due to DNA degradation over time due to the tissue preservation techniques employed (Guyard et al., 2017). To maximize the usability of reads from FFPE samples, this study focuses on the DADA2 tool. Due to the complexity of the samples, the study was designed using five different sets of primers to amplify five regions of the 16S rRNA gene, which would later be reconstructed thanks to QIIME2-implementation Sidle. The pipeline used the QIIME2 ecosystem platform installed in a Conda environment, where all the required tools were available. Firstly, paired-end samples in FASTQ format were imported to QIIME2 and were separated into the five regions by Cutadapt, thanks to the five sets of primers. From here on, the general process will be counted, but it was carried out in each of the regions. Cutadapt usually returns a qza extension file output, which would be the input file for other tools, but, for dynamically visualizing the results and data, QIIME2 provides a qzv extension file. These types of files are designed to be opened in the web-based interactive viewer called QIIME2 View, which offers a user-friendly manner of exploring data.

Henceforth, assuming that the qzv file is actually a compressed file consisting of a set of files containing the graphics and data to be displayed, files containing the data for forward and reverse directions were analysed. These two files showed the same structure: a data frame with eight columns and rows as length sequence, representing the data in two interactive graphical plots with quality values as the y-axis and sequence length as the x-axis. Various Python scripts were created to analyse the data and extract the most optimal positions for forward and reverse directions. Some filter criteria were established during the data exploration (each direction was evaluated independently):

1. Eliminate low-quality sequences by multiplying the maximum count by 0.2.
2. Select a minimum starting length, typically set at 100.
3. Get statistical analysis of each quality columns (2%, 9%, 25%, 50%, 75%, 91% and, 98 %) and count column.
4. Transform quality values within columns to a scale of 0, 1, or -1.
5. Detect local maxima points within columns 50% and 75% using the Python library scipy.signal and the find_peaks function.
6. Test optimal positions for DADA2 input parameters, ensuring they are greater than the minimum length requirement (here 100).
7. Set left-side values to zero based on previous results.
8. Ensure optimal positions satisfy DADA2's overlap formula.

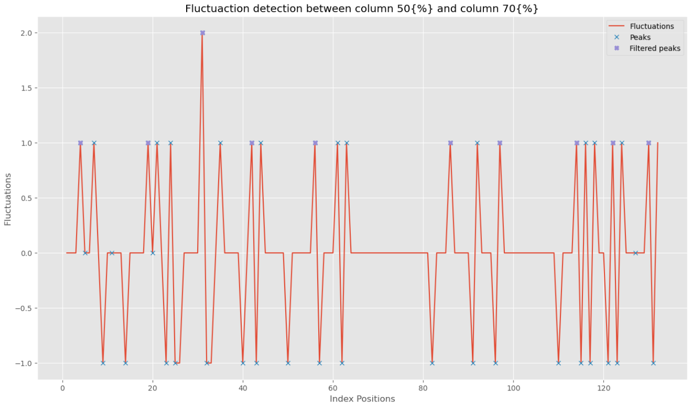

Figure 1: Forward strand graph for 16S rRNA gene amplified R1 region

Upon completing the testing of optimal positions for both directions and obtaining three DADA2 output files, we analyse the denoising_stats output file. This file provides denoising statistics, detailing the entire DADA2 denoising process, including total input reads, passing reads after quality filtering, merged sequences, and the number of non-chimeric reads retained. The aim is to retain as many non-chimeric reads as possible for each sample, although the specific target may vary depending on the study's goal and characteristics.

4 Preliminary Results

This study proposes an ad hoc approach to automate an intermediate step between Cutadapt and DADA2 tools, detecting changes along the quality values of base positions in both read directions. Traditionally, manual selection of optimal positions requires careful examination of data to identify fluctuations in quality levels. The proposal detects small and large changes and indicates the most optimal positions for trimming in the right section of the graph, ensuring they are higher than position 100, as in Sidle, post-dada2-trimming is performed before alignment. This makes it easier to perform the step of alignment and reconstruction of the final sequence of the 16S rRNA gene, formed by the regions under study (more than 60% of the region is reached in the reconstruction).

In a brief example of how it works, for the R1 region, manually selected position values for the forward and reverse strands were 114 and 115, respectively. However, the algorithm suggested positions of [114, 122, 130] for the forward strand and [112, 126] for the reverse strand. Test results indicated that the optimal trimming positions were 130 for the 'trunc-len-f' argument (right side of the forward strand) and 126 for the 'trunc-len-r' argument (right side of the reverse strand). As shown in Figure 1, the blue cross represents all potential optimal positions, while the highlighted purple cross signifies the final optimal values for the forward direction, with the x-axis representing the index position and the y-axis indicating detected fluctuations. In this pipeline, the 'trim-left-f' and 'trim-left-r' parameters were set to zero for both strands. It is evident that the manually selected positions were overly conservative. This process was applied to five regions, and the automated selection proved more accurate, except for the R2 and R4 regions, where there were closely spaced fluctuations.

5 Discussion

The study of the microbiome is still a pending subject. Although a great deal of progress has been made in a short time that it has really been booming, it is still necessary to standardize the processes. In this case, the work focused on a complicated type of sample, since due to the treatment it receives before sequencing, the starting DNA is highly degraded. However, these types of samples can now be analysed thanks to the SMURF tool implementation within QIIME2, called Sidle. Despite this, throughout the workflow defined in Sidle, a great loss of information has been observed simply because the values of the trimming parameters in DADA2 have been incorrectly defined.

In a promising development, this limited-scale study introduces an initial ad hoc approach to streamline an intermediate step between the Cutadapt and DADA2 tools. Our primary goal is to efficiently identify quality changes in both read directions, going beyond traditional manual selection by detecting subtle and significant changes. The approach determines the most suitable trimming positions within the right section of the quality profile graph, with a criterion set beyond position 100. This choice aligns with the Sidle workflow, where a pre-alignment step called post-dada2-trimming is applied, splitting sequences at position 100 across all samples from all regions. This approach streamlines subsequent alignment and reconstruction steps for the final 16S rRNA gene sequence, covering over 60% of the targeted gene during reconstruction.

There remains a substantial amount of work ahead, considering that the results presented are in their preliminary stages. To ensure the robustness and applicability of our approach, several key tasks are on our agenda. Firstly, we will evaluate the tool's performance on datasets containing FFPE samples, as well as normal datasets, i.e. V3V4 samples. Secondly, we have future plans to expand our testing to encompass 18S and ITS samples, though this lies further down the road. Additionally, we aim to benchmark the optimization outcomes against other established tools in the field. Furthermore, we will explore the utility of our approach in the context of the Deblur tool, which is akin to DADA2 but selected based on specific process requirements and sample types. Lastly, our strategy involves the meticulous analysis of DADA2's verbose output during runtime, enabling us to proactively halt the process if anomalies such as sample removal in filtering steps emerge, ensuring the integrity of the workflow.

6 Conclusions

Microbiota's role in diseases like autoimmunity, cancer, neurological, and renal diseases has been studied over the past two decades. Tools like Sidle, integrated into QIIME2, allow for a pre-established pipeline for analysing FFPE paired-end sample sequences. The pipeline involves other tools as Cutadapt for primer removal and DADA2 for denoising, with parameters 'trunc-len-f' and 'trunc-len-r' determining read truncation. We suggest a new automated approach, based on the data output (normally visualized graphically to choose the positions) and following the quality-filtering indications from Cutadapt. In conclusion, this paper presents preliminary findings, and our ongoing work aims to optimize this procedure, incorporate enhancements, and potentially develop machine learning or deep learning models for full automation, with plans for benchmarking against other tools.

Funding

The development of the project is supported in part by CITIC as a Research Center of the Galician University System and is financed by the Consellería de Cultura, Educación e Universidade de la Xunta de Galicia through the ERDF (80%), Operational Program ERDF Galicia 2014–2020, and the remaining 20% by the Secretaría Xeral de Universidades (Grant. ED431G 2019/01), as well the Xunta de Galicia/ERDF-EU [ED431C 2021/53], and the Recovery, Transformation, and

Resilience Plan (PRTRC17.I1), The present study also received funding from the Instituto de Salud Carlos III (ISCIII, Spain) through the project PI20/00413 and co-funded by the European Union (EU) to M.P., while K. C-P. was granted by Asociación Española Contra el Cáncer (AECC).

Acknowledgments

The authors of this paper extend their sincere appreciation to the collaborative efforts and contributions of the meiGAbiome Group, as well as the entire team of medical and anatomopathologists. Finally, we are deeply grateful to the patients whose selfless donations have made this and numerous other studies possible.

Bibliography

E. Bolyen et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37(8):852–857, 2019.

B. Callahan et al. DADA2: High-resolution sample inference from illumina amplicon data. *Nature Methods*, 13(7):581–583, 2016.

K. Conde-Pérez et al. Parvimonas micra can translocate from the subgingival sulcus of the human oral cavity to colorectal adenocarcinoma. *Molecular Oncology*, 2023.

J. Debelius et al. A comparison of approaches to scaffolding multiple regions along the 16S rRNA gene for improved resolution, 2021.

Ewels, *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology*, 2020.

G. Fuks et al. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. *Microbiome*, 6(1), 2018.

A. Guyard et al. DNA degrades during storage in formalin-fixed and paraffin-embedded tissue blocks. *Virchows Archiv*, 471(4):491–500, 2017.

M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, 17(1):10–12, 2011.

Sociedad Española Oncológica Médica (SEOM). Las cifras del cáncer en españa 2023. *https://seom.org/images/Las_cifras_del_Cancer_en_Espana_2023.pdf*, 2023.

D. Straub et al. Interpretations of environmental microbial community studies are biased by the selected 16s rrna (gene) amplicon sequencing pipeline. *Frontiers in Microbiology*, 11, 2020.

T. T. Wang et al. Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, 6(2):320–329, 2012.

M. Weinstein et al. Figaro: An efficient and objective tool for optimizing microbiome rrna gene trimming parameters, 2019.

S. Wong et al. Gut microbiota in colorectal cancer: mechanisms of action and clinical applications. *Nature Reviews Gastroenterology & Hepatology*, 16(11):690–704, 2019.