THE EFFECT OF QUALITY TRIMMING ON JOINING PAIRED-END READS IN MICROBIOME DATA ANALYSIS

Kristýna Heřmánková

Bachelor Degree Programme (3), FEEC BUT

E-mail: xherma30@stud.feec.vutbr.cz

Supervised by: Karel Sedlář

E-mail: sedlar@feec.vutbr.cz

Abstract: The main goal of microbiome research is to determine the microbial composition of a target sample. In successfully performed research, a cause of a disease can be found or pathogens in an environmental sample can be revealed. The correct evaluation of microorganisms present in the analysed sample is therefore required, but unfortunately not that often reached. The accuracy of the resulting composition can be affected already in pre-processing steps of the analysis. The frequent issue that can negatively affect the reliability of research is data loss. This loss is most common in the case of the paired-end reading method. The occurrence of data loss can be caused by the need of joining this pair of sequences into one continuous sequence. The need of joining reads with low-quality ends together can lead to counts reduction of sequences in an available dataset, and therefore unrealistic results can be obtained. This paper shows, how quality trimming can affect the number of lost sequences during the reads joining step.

Keywords: NGS, quality trimming, 16S rRNA, amplicon

1 INTRODUCTION

For purpose of the microbial study, which aims to determine the taxonomic diversity of a sample, the amplicon sequencing approach is used. The principle of this approach lies in sequencing only a target gene or its regions instead of the whole genome. This gene, called the marker gene, should be specific enough, to distinguish ideally each species. For the domain of *Bacteria*, the 16S rRNA gene is used. The suitability of this gene for taxonomy classification lies in its presence in every organism and its hypervariable, and also conserved regions. The most significant information on taxonomic diversity would be reached by sequencing the whole 16S rRNA gene. Unfortunately, current sequencing technologies, which are capable sequence its length of approximately 1600 bp, are still in the development process because of their high sequencing error rate. In this case, the next-generation sequencing technologies (NGS), which generate reads of significantly shorter length, are used. This means, that only a few regions of the 16S rRNA gene can be sequenced.

2 METHODS

The widely used sequencing technology for amplicon sequencing purposes is Illumina. This technology offers a few platforms for various applications. For reaching the longest possible reads, Illumina also offers paired-end sequencing. This method generates two reads from both ends of one sequence. Each read can be of length up to 300 bp, depends on the selected platform, so the target sequence can be almost twice longer than sequencing with the single-end method.

For a longer resulting sequence obtained from the paired-end method, overlapping of generated reads is required. The length of an overlapping region can vary. If the whole length of reads overlaps, the resulting sequence is not longer, but can only reaches higher quality. If reads overlap only partially, the long continuous sequence is reached. This requires the process of joining these two reads together.

Because of the different orientation of the sequencing process, reads are joined with their ending tails. The longer the sequenced region is, the higher probability of error occurrence is. With the higher error rate, the number of low-quality bases increases. This can lead to the failure of the joining step. The following text describes, how reads can be treated to prevent unsuccessful reads joining.

2.1 QUALITY TRIMMING

Before the quality trimming was done, specific primers were removed. This step was performed in QIIME 2 (Version 2020.8) [1] (Quantitative Insight Into Microbiome Environment), which is the most used open-source platform for microbiome analysis.

A Phred quality score defines the accuracy of base calling. This accuracy is measured for each base in the read and the corresponding score is recorded. FASTQ format then stores both, the sequence of nucleotides and the related score. Generally, bases with the quality under the value of 20 are considered low-quality, but the value can vary depending on usage. Table 1 shows the selected Phred scores, and corresponding accuracies calculated using formula:

$$A = 1 - 10^{\frac{-Q}{10}} \quad (1),$$

where A is the base call accuracy of the Phred score Q.

Table 1: The base call accuracy and the corresponding score.

Phred quality score	Base call accuracy
10	90 %
20	99 %
30	99,9 %
40	99,99 %

For trimming low-quality ends of paired-end reads, Trimmomatic (Galaxy Version 0.38.0) [2] [3] was used. This tool is designed for trimming sequences from Illumina platforms. Trimmomatic offers trimming of single-end or paired-end reads in FASTQ format and allows a few options, how to trim these reads. In this paper, the TRAILING method was used. The method only needs target files with paired-end reads and a quality threshold. The TRAILING method begins at the end of reads and tracks each base quality. If the quality is under the selected threshold, the base is removed. Trimming continues towards the beginning of the read and stops when the base quality is at least equal to the threshold. This is done for both paired reads.

2.2 JOINING READS

Data from Trimmomatic were imported back to the QIIME 2 and a joining step was performed. A few parameters can be set for joining reads. The –p-maxdiffs parameter can be important in the case of successful reads joining. The default value of this parameter is 10, which means that only 10 mismatches in the overlapping region are allowed. If a match is recorded, the resulting sequence has at this position higher quality. If there is a mismatch between reads, the base with higher quality is chosen and the resulting quality of the base decreases [4]. Because of the subsequent analysis, sequences still need to be of high quality, and thus the default value of this parameter was used.

2.3 MATERIALS

Sequencing data used in this paper were sequenced with the Illumina MiSeq platform for generating paired-end reads of length 2x300 bp. These data consist of three bacterial communities artificially prepared in Veterinary Research Institute with the purpose to provide testing datasets for a chimera detection tool. Sequences are amplicon sequence variants (ASVs) of 16S rRNA of bacteria present in the sample. The overlapping region is approximately 140 bp long, so the resulting continuous sequence is of length about 460 bp.

3 RESULTS

Table 2 contains numbers of paired-end reads in available datasets before the mentioned exact process was performed.

Table 2: Datasets and their counts of reads.

Dataset	Number of reads
P2	103 606
P3	71 610
P4	64 029

In Figure 1A, the impact of allowed mismatches in overlapping regions is shown. The default value of 10 was chosen, because of a relatively high number of joined sequences, and also high-quality sequences for further analysis.

As Figure 1B shows, sequences without the quality trimming step have a bigger issue with joining pairs together. Unlike the untrimmed reads where only 68 % of all reads in available datasets passed the joining step, 90 % of quality trimmed reads were successfully joined. The reason why this happens can be, as mentioned previously, that low-quality ends in overlap region can mismatch. If lots of mismatches are recorded, sequences are not allowed for passing the joining step. On the opposite, if the match between bases is recorded, the resulting sequence is of high quality.

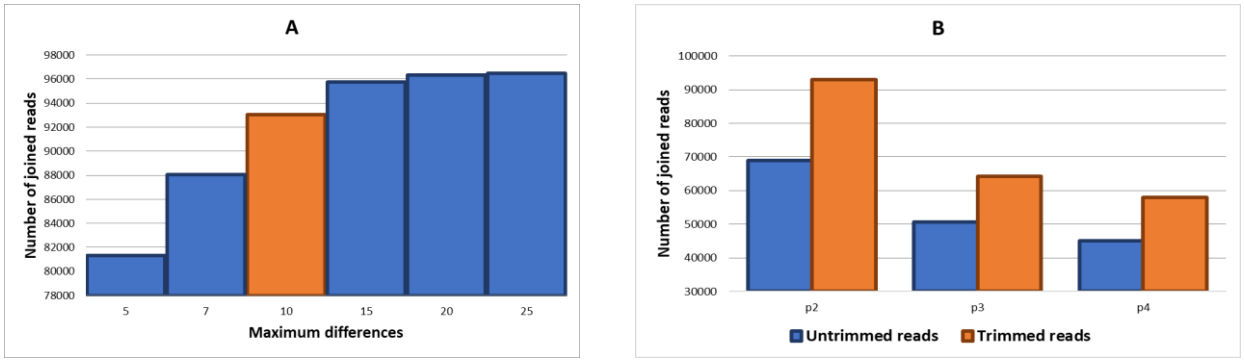

Figure 1: (**A**) Histogram of maximum numbers of mismatches that can occur in the overlapping region. (**B**) Visualization of differences between untrimmed and trimmed reads in available datasets.

Figure 2 shows a few quality thresholds used in the Trimmomatic tool for quality trimming and their impact on counts of successfully joined reads. Value of threshold ensures which bases at the

end of reads should be retained and which are supposed to be trimmed. The orange highlighted threshold was selected as it achieved the highest number of joined reads.

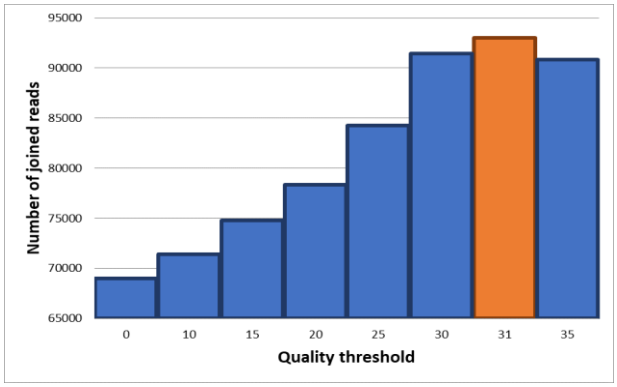

Figure 2: The difference between counts of successfully joined reads with the chosen threshold.

4 CONCLUSIONS

The further analysis of microbiome studies lies in subsequent processing steps as dereplication, where every ASV is identified, chimera detection, which should reveal spurious organisms, and operational taxonomic unit (OTU) picking, where ideally every OTU represents one species. In all these steps, the removal of insufficient sequences is performed, and thus the unnecessary data loss needs to be inhibited.

Also, the result of microbial research, which aims to diversity analysis of a sample, depends on the number of sequences that went through the whole analysis. The higher the number of lost sequences is, the probability of incorrectly revealed microbial community increases. On the opposite, if the number of appropriately treated sequences is the highest possible, results will be probably more realistic.

As results showed, quality trimming has a positive impact on obtained reads for further analysis, and therefore should be included as the pre-processing step on Illumina paired-end reads.

REFERENCES

[1] BOLYEN, Evan, Jai Ram RIDEOUT, Matthew R. DILLON, et al., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology* [online]. **37**(8), 852-857. ISSN 1087-0156. DOI: 10.1038/s41587-019-0209-9

[2] BOLGER, Anthony M., Marc LOHSE a Bjoern USADEL, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. **30**(15), 2114-2120. ISSN 1460-2059. DOI: 10.1093/bioinformatics/btu170

[3] AFGAN, Enis, Dannon BAKER, Bérénice BATUT, Marius VAN DEN BEEK, Dave BOUVIER, Martin ECH, John CHILTON, Dave CLEMENTS, Nate CORAOR, Björn A. GRÜNING, Aysam GUERLER, Jennifer HILLMAN-JACKSON, Saskia HILTEMANN, Vahid JALILI, Helena RASCHE, Nicola SORANZO, Jeremy GOECKS, James TAYLOR, Anton NEKRUTENKO a Daniel BLANKENBERG, 2018. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Research*. **46**(W1), W537–W544. ISSN 13624962. DOI:10.1093/nar/gky379

[4] EDGAR, Robert C. a Henrik FLYVBJERG, 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. *Bioinformatics*. **31**(21), 3476-3482. ISSN 1367-4803. DOI:10.1093/bioinformatics/btv401