Marine Genomics 61 (2022) 100914

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Benefits of merging paired-end reads before pre-processing environmental metagenomics data

Midhuna Immaculate Joseph Maran, Dicky John Davis G. *

Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai, India

ARTICLE INFO

Keywords:
Metagenomics
High throughput sequencing
Environmental DNA
Sequence alignment
Quality processing
Trimmomatic

ABSTRACT

Background: High throughput sequencing of environmental DNA has applications in biodiversity monitoring, taxa abundance estimation, understanding the dynamics of community ecology, and marine species studies and conservation. Environmental DNA, especially, marine eDNA, has a fast degradation rate. Aside from the good quality reads, the data could have a significant number of reads that fall slightly below the default PHRED quality threshold of 30 on sequencing. For quality control, trimming methods are employed, which generally precede the merging of the read pairs. However, in the case of eDNA, a significant percentage of reads within the acceptable quality score range are also dropped.

Methods: To infer the ideal merge tool that is sensitive to eDNA, two Hiseq paired-end eDNA datasets were utilized to study the merging by the tools – FLASH (Fast Length Adjustment of SHort reads), PANDAseq, COPE, BBMerge, and VSEARCH without preprocessing. We assessed these tools on the following parameters: Time taken to process, the quality, and the number of merged reads.

Trimmomatic, a widely-used preprocessing tool, was also assessed by preprocessing the datasets at different parameters for the two approaches of preprocessing: Sliding Window and Maximum Information. The pre-processed read pairs were then merged using the ideal merge tool identified earlier.

Results: FLASH is the most efficient merge tool balancing data conservation, quality of reads, and processing time. We compared Trimmomatic's two quality trimming options with increasing strictness with FLASH's direct merge. The raw reads processed with Trimmomatic then merged, yielded a significant drop in reads compared to the direct merge. An average of 29% of reads was dropped when directly merged with FLASH. Maximum Information option resulted in 30.7% to 68.05% read loss with lowest and highest stringency parameters, respectively. The Sliding Window approach conserves approximately 10% more reads at a PHRED score of 25 set as the threshold for a window of size 4. The lowered PHRED cut off conserves about 50% of the reads that could potentially be informative. We noted no significant reduction of data while optimizing the number of reads read in a window with the ideal quality (Q) score.

Conclusions: Losing reads can negatively impact the downstream processing of the environmental data, especially for sequence alignment studies. The quality trim-first-merge-later approach can significantly decrease the number of reads conserved. However, direct merging of pair-end reads using FLASH conserved more than 60% of the reads. Therefore, direct merging of the paired-end reads can prevent potential removal of informative reads that do not comply by the trimming tool's strict checks. FLASH to be an efficient tool in conserving reads while carrying out quality trimming in moderation. Overall, our results show that merging paired-end reads of eDNA data before trimming can conserve more reads.

1. Introduction

Metagenomics, also referred to as environmental genomics, is the study of genetic material recovered directly from environmental samples (skin and gut samples, soil and water samples) that can be preserved, extracted, amplified, sequenced, and categorized based on its

Abbreviations: DNA, Deoxyribonucleic Acid; eDNA, Environmental DNA; FLASH, Fast Length Adjustment of SHort reads; rRNA, Ribosomal RNA; SW, Sliding Window; MI, Maximum Information; NGS, Next Generation Sequencing.

* Corresponding author at: Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research, Chennai 600116, India.
E-mail address: dicky@sriramachandra.edu.in (D.J. Davis G.).

https://doi.org/10.1016/j.margen.2021.100914
Received 16 July 2021; Received in revised form 18 November 2021; Accepted 18 November 2021

sequence (Bouchot et al., 2014). In traditional genomics, cultivated clonal cultures and early environmental gene sequencing cloned specific genes (e.g. 16S and 18S rRNA genes) are essential for producing a profile on diversity in a natural sample. Though widely used, this method excludes a vast majority of biodiversity (Hugenholz et al., 1998). Since metagenomics provides the most coverage in the estimation of the organisms in a sample, it is a powerful tool for biodiversity studies. The major limitation of environmental DNA (eDNA) studies is the degradation of eDNA in the environment, as often, only small segments of genetic material remain (Seymour, 2019). Nevertheless, eDNA still has numerous applications in conservation, monitoring, and ecosystem assessment.

The affordability of Next-Generation Sequencing (NGS) technology has paved the way to generate a plethora of sequenced data. One method of sequencing is the paired-end read technology, which generates reads from the two ends of target DNA fragments that are then merged to obtain the original sequence in full length. This technology promises great advantages: it produces twice the number of reads to that of single-end read sequencing, and sequences aligned as paired reads enable better detection of genomic rearrangements and repetitive sequence elements, gene fusions, and insertion-deletion (Indel) variants, which is not possible with single-read data (*Metagenomic Analysis of Environmental Water Samples With the NextSeq® 500 System*, 2015).

However, NGS sequencing is with its limitations. It is known that the read quality drops significantly towards the 3′ end due to sequencer limitations (Fuller et al., 2009). Short reads are usually discarded because they occur multiple times within the target sequence and therefore give very vague and misleading genomic information. Therefore, quality trimming tools are used before merging the reads. Trimmomatic is a multithreaded command-line tool that is used for quality trimming and removal of adapters from FASTQ data for paired-end and single-end reads (Bolger, 2014). It offers two methods of quality processing; Sliding Window (SW) quality filtering and Maximum Information (MI) quality filtering. The former method scans from the 5′ to the 3′ end of the read and removes bases from the terminating portion when the average quality of a group of bases drops below a specified threshold. The latter method is a novel technique, according to the authors, where the trimming process becomes increasingly strict as it progresses through the read, rather than applying a fixed quality threshold.

Merge tools such as FLASH (Fast Length Adjustment of SHort reads) merge paired-end reads by overlapping them from fragment libraries shorter than twice the length of reads (Magoč, 2011). FLASH performs error-correction before merging reads. VSEARCH is another open-source merge tool for processing genomics and metagenomics nucleotide sequence data (Rognes, 2016). It performs an array of functions, from file conversion to global alignment and operational taxonomic unit (OTU) clustering, including paired-end reads. The algorithm computes the optimal ungapped alignment of the overlapping region of the forward sequence and the reverse-complemented reverse sequence.

Another commonly used merge tool is BBMerge, which is also an overlap-based tool for merging short high-throughput shotgun sequencing reads (Bushnell, 2017). BBMerge allows simple adjustment of merging sensitivity to process large datasets of different sequence types. PANDAseq is a 16S rRNA gene amplicons assembly and error correction tool that corrects errors probabilistically with the overlap data from the paired-end reads, and when the overlap between the forward and reverse reads is of the minimal overlap threshold, the uncalled or miscalled bases are corrected using the complementary sequence (Masella, 2012). COPE is a free tool that connects pair-end reads by using kmer frequency information to authenticate possible overlaps of reads, which is also used in error correction in reads (Liu, 2012). These tools are used in the preprocessing and merging steps and therefore play a crucial role in downstream processing.

While it is imperative to conserve only the data of the highest quality, it is to be noted that eDNA which is exposed to external factors could possibly be compromised, especially marine eDNA which is reported to degrade about 1.7 times faster in an inshore environment than the offshore (Collins et al., 2018). Therefore, a stringent quality assessment would drop a significant portion of the read, thus making it shorter, which is eventually discarded. Therefore, moderate retention of average quality bases would make these short reads sufficiently long enough to be informative, risking the preservation of errors. Therefore, when reads are quality processed, and then merged, the sequences are unwittingly processed twice. The reason being, while these merge tools do not explicitly mention their ability to merge adapter trimmed reads to assemblages without quality trimming, it is evident from their default settings that such processing is carried out.

2. Materials and methods

It has been shown that quality-based trimming of NGS data increases the alignment of reads (MacManes, 2014). However, this increased mappability of reads remaining after trimming comes at the expense of a dramatic reduction in the absolute number of aligned reads, as a consequence of some reads failing to pass minimum quality criteria during trimming. We predicted that this loss of information would impact analyses' downstream of alignment; in particular, sequence alignment. To assess this impact, we first obtained two Antarctic seawater metagenomic eDNA datasets (SRR3952299 and SRR3952300) from NCBI's Sequence Read Archive (SRA) (*NCBI-SRA*, n.d.). The files contained unmerged adapter-free raw reads (each read of length 250 nt, making the total pair length of 500 nt) sequenced using Illumina Hiseq Technology. FASTQC tool was used to analyse the quality and the number of reads throughout the experiment (Andrews, n.d.).

To identify the optimal merge tool, the raw reads of each dataset were merged individually using the merge tools: VSEARCH v2.10.4, FLASH v1.2.11, PANDAseq v2.11, COPE v1.1.0 and BBMerge v38.33 without pre-processing, shown in (Fig. 1). The resulting merged sequences were visualized using FASTQC. For this study, we used a quad-core i5 Intel® Core™ Linux machine, clocking at 2.20GHz, AMD Radeon graphics, and 12 GB RAM. Steps were taken to ensure that no background processing occurred during the tests that could influence the results.

The output from each of the merge tools was manually checked for the number of reads conserved, the quality of assemblages, and the time taken. We tested the accuracy of merging using reads chosen at random from the merged file (T) as a template and aligning the corresponding raw reads manually (M). The forward read (M1) and the reverse read (M2) were examined for their overlapping regions after the latter was processed in EMBOSS Revseq (*ResearchGate*, n.d.). The manually merged reads were then assessed against the T file for any misaligned merges. We used FASTQC to identify the most applicable tool for eDNA analysis. The time taken for each run of every tool was noted. Additionally, the

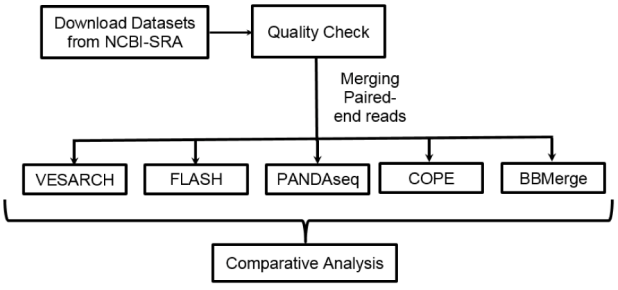

Fig. (1). Flow chart to identify the optimal merge tool without pre-processing eDNA metagenomics data.

Fig. 1. Flow chart to identify the optimal merge tool without pre-processing eDNA metagenomics data.

RAM used was constantly monitored.

To identify the best methodology for processing eDNA datasets, we quality trimmed the datasets using Trimmomatic's v0.38 two trimming approaches: Sliding Window and Maximum Information, shown in (Fig. 2). We tested the rate of conservation of reads by studying the range of values that were tested on each approach. The Sliding Window approach was studied for the optimal quality value for maximum conservation of reads with a good quality score and for the optimal window, i.e., the number of bases in a read to be evaluated at a given time, in succession. Analysis of the Maximum Information approach was also carried out in the same fashion. The quality trimmed reads with the highest conservation of reads and the quality from each approach were then merged using the merge tools. A minimum overlap of 10 bases was set and other parameters were at default for all the merge tools. The reads and assemblages were assessed using FASTQC.

3. Results

3.1. FLASH, the most optimal merge tool for eDNA data

To estimate the magnitude of read loss by the standard steps of preprocessing, the datasets were subjected to the following tests: first, the raw reads, after quality check, were merged with all five merge tools and were assessed. They were then subjected to trimming using Trimmomatic. Second, the raw datasets were pre-processed first with Trimmomatic, which involved a gradient increase in strictness in both Sliding window and Maximum Window options and the measurements were recorded. The trimmed datasets were ultimately merged with the FLASH. The results of reads that were quality processed using Trimmomatic are outlined in the Table 1. For identifying the most efficient tool, we assessed the following parameters: Time taken to process, the quality, and the number of merged reads.

We found FLASH to be one of the well-balanced tools for merging. With a very small mismatch ratio and error rate (<0.40 for a 700,000 merged pairs and < 1%, respectively (Magoč, 2011)), FLASH conserves an average of 71% of the reads in the datasets and carries out the process within an approximate span of 30 min. FLASH uses less RAM space during the run and is very compatible with non-customized computers. The quality across bases for both datasets is excellent with a PHRED quality score above Q34 with per sequence quality score of 37 for both datasets. Though VSEARCH processes faster, its stringent parameters merge a comparatively lower number of reads when compared with PANDASEQ and FLASH. VSEARCH outputs good quality assemblages with an approximate PHRED score of 39 across the distribution of assemblages. Merging paired-end reads with FLASH set at a minimum overlap of 10 bp (default setting) was sufficient to conserve good quality

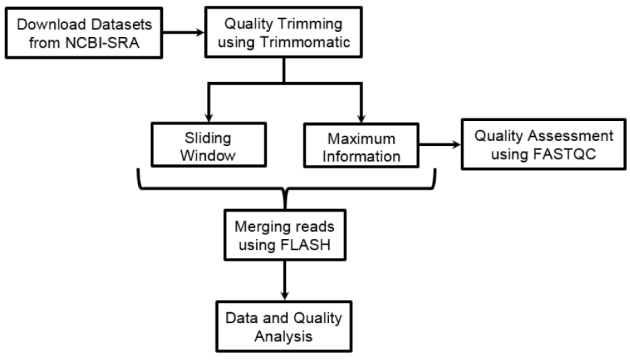

Fig. (2). Flow chart to identify the optimal merge tool with quality trimming methods of Trimmomatic tool for eDNA metagenomics data.

Fig. 2. Flow chart to identify the optimal merge tool with quality trimming methods of Trimmomatic tool for eDNA metagenomics data.

Table 1
Comparison of popular merge tools based on their conservation number, quality of the output, and the time taken for processing two HiSeq eDNA datasets.

FILE	Merge Tool	Number of Reads Conserved	Average of Mean Quality Across Bases	Time Taken to Process (Min)
	BBMERGE	24,051,205	38.5	43
	COPE	9,515,163	39	48
SRR3952299	FLASH	27,012,070	37.5	35
	PANDASEQ	32,856,304	34	65
	VSEARCH	22,779,640	38.5	30
	BBMERGE	25,892,483	39	88
	COPE	7,166,053	37	40
SRR3952300	FLASH	28,478,153	37	26
	PANDASEQ	32,959,795	35	78
	VSEARCH	23,481,879	39	30

merges. For those reads which have no overlaps or poor overlaps, FLASH outputs the forward and reverse reads into two separate FASTQ files for normal assembly.

BBMERGE, a popular tool for NGS processing, is with an agility that is slightly behind FLASH and VSEARCH. However, it conserves fewer reads when compared with FLASH and PANDAseq; its conservation rate is similar to that of VSEARCH. The output files have an excellent PHRED score of 39 across bases for all datasets.

PANDAseq conserves the maximum number of merged reads in all datasets, however, it takes the longest time to process. Unlike other tools, with the exception of COPE, a drop in the average quality of bases in the middle section of the reads is observed. However, the average quality score across bases is maintained at 37 and above. COPE's parameters are observed to be strict by default and yield the least number of merged data with a 78.8% loss of data. The time taken for the run is similar to that of BBMERGE.

The reads of datasets were subjected to random manual alignment tests against the assemblages produced by each of the tools. All the tools use overlapping bases of forward and reverse bases to merge the reads accurately.

3.2. Analysis of trimmomatic methods

Trimmomatic offers two methods of quality filtering, namely, Sliding Window (SW) and Maximum Information (MI) quality filtering. While the former is a standard method used by many trimming tools, the latter is a novel method that is a characteristic feature of Trimmomatic. The Maximum Information quality filtering provides the user the freedom to set the strictness value between 0 and 1, with 1 being the highest strictness value. The Sliding Window approach allows users to set the preferred PHRED value for a user-defined number of bases (window) at a given time. The number of bases in the read that is to be evaluated at a given point is also user-defined.

To identify the loss of reads that occurs during preprocessing, we tested both methods of quality trimming at different quality stringency levels and window sizes/ number of bases to examine. We studied the results produced by MI to identify the ideal quality setting that allows for maximum conservation of reads while maintaining an acceptable quality score across bases. The identified quality setting was used to select the number of bases to be read between the manual prescribed 40 bases to 90 bases shown in (Fig. 3a and b). Similarly, the Sliding Window approach was studied by identifying the most applicable quality setting and then its window size shown in (Fig. 4a and b).

We found that under Maximum information, the least quality setting produces the maximum number of reads with an average score of 38 of mean quality across bases. We find a median increase of 3037 reads when the number of bases to be scanned is fixed between 40 and 90 bases. The Sliding Window shows a similar result with its quality trend across a set window. At PHRED score Q5 quality setting, we find the mean quality across bases to average 38. Window sizes tested from 1 to

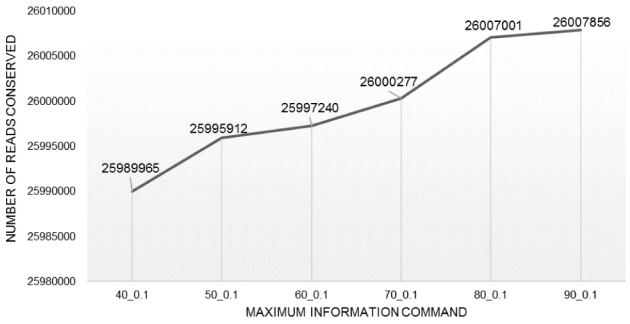

Fig. (3a). Graphical representation of reads conserved in dataset SRR3952299 on increasing length of target for analysis using Maximum Information Command.

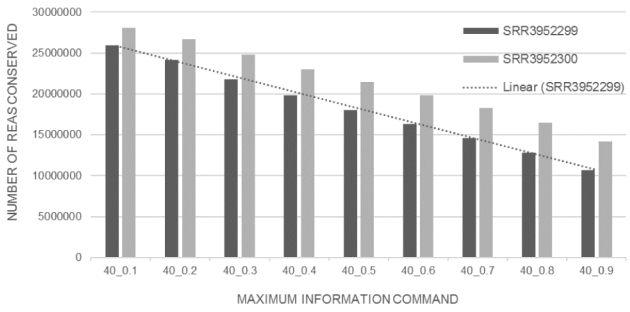

Fig. (3b). Graphical representation of conservation of reads in both datasets on increasing strictness using Maximum Information Command.

Fig. 3. a Graphical representation of reads conserved in dataset SRR3952299 on increasing length of target for analysis using Maximum Information Command. b. Graphical representation of conservation of reads in both datasets on increasing strictness using Maximum Information Command.

10 showed a negligible increase in the conservation of reads.

4. Discussion

The necessity of quality trimming has been debated among researchers (*ResearchGate*, n.d.; *Biostars*, n.d.). However, these do not directly address the issues pertaining to environmental DNA. Our study, therefore, converges into the unique problem that is brought forward by environmental DNA, namely, DNA sequences exposed to harsh environmental conditions. While the previous research may aid in understanding the outcomes of quality trimming on data, we expound the importance of a delicate balance between conservation and quality of reads. Therefore, to our knowledge, this is the pioneer study on the standardization of trimming thresholds on paired-end Illumina reads for metagenomic analysis. We have, however, constricted only to the Illumina Hiseq datasets from one experiment for replication's sake.

Generally, trimming tools, depending on the quality parameter threshold Q set, may dramatically reduce the dataset size or retain random reads. Therefore, the parameters have been left to the researcher's discernment to balance the read loss and dataset quality. However, researchers who choose to compute with manual -specified/default parameters may unwittingly influence the final merged data. For this reason, we have carried out this comparative analysis of different merge tools and gradient studies to understand the trade-off between read loss and quality.

We found that the reads in raw files of all datasets show a significant drop in the quality of reads towards the 3′ end, but are free of any

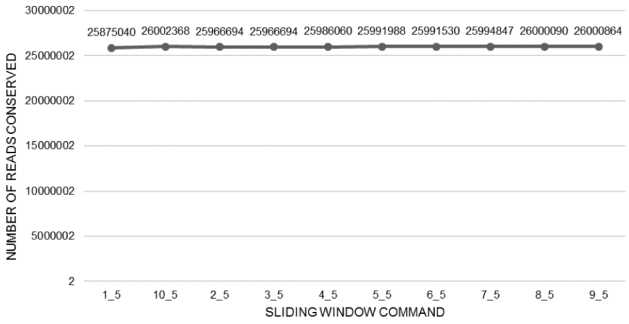

Fig. (4a). Graphical representation of reads conserved in dataset SRR3952299 on increasing length of target for analysis using Sliding Window Command.

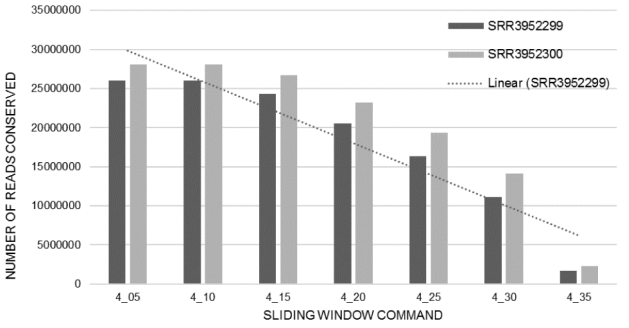

Fig. (4b). Graphical representation of conservation of reads in both datasets on increasing strictness using Sliding Window Command.

Fig. 4. a Graphical representation of reads conserved in dataset SRR3952299 on increasing length of target for analysis using Sliding Window Command. b. Graphical representation of conservation of reads in both datasets on increasing strictness using Sliding Window Command.

adapter contamination. Overall, we believe that our results show how trimming the paired-end environmental DNA datasets before merging can decrease the final number of merged reads. The results of the merged datasets were well above the required quality score across bases. Of the 5 merge tools, we found that FLASH optimally balances both maximal conservation of reads and the quality across bases, while COPE is found to be too stringent for eDNA data.

We focussed on the metagenomics datasets obtained by sequencing with Illumina's Hiseq technology. The implications of our recommendations on reads produced by other sequencing methods are beyond the scope of this paper. We would, in the future, analyse the benefits of direct merging of reads without pre-processing on data from other types of sequencing methods.

5. Conclusion

Multiple merge tools are available for merging paired-end reads from NGS data. However, it is imperative to identify the tool that is suited for the data of interest. Environmental DNA obtained could potentially be degraded due to exposure to unsuitable environmental elements. Therefore, multiple widely-used merge tools were tested for their robustness to merge and retain maximum number of reads sequenced with Illumina's Hiseq. FLASH is found to be the most efficient tool at conserving reads and at time taken to process the data.

Quality trimming of Hiseq data reads with default parameters ensures that the quality drop towards the end of the reads, as expected with Illumina sequencing, are trimmed. However, the parameters are set for good to best quality DNA samples while eDNA samples collected may have varying percentage of degraded samples. Therefore, trimming

reads with default parameter prior merging could effectively reduce the number of overlapping bases required to merge paired-end reads, thus leading to loss of reads. We have also observed that merge tools indirectly trim reads while merging. So, we recommend direct merging of reads. Should the researcher prefer to quality trim with Trimmomatic prior merging, we recommend lowering the strictness to ensure maximum retention of good quality reads.

Ethics approval and consent to participate

Not applicable.

Human and animal rights

No Animals/Humans were used for studies that are the basis of this research.

Declaration of competing interest

The authors declare no conflict of interest, financial or otherwise.

References

Andrews, S.. FastQC: A Quality Control Tool for High Throughput Sequence Data. http://www.bioinformatics.babraham.ac.uk. (Accessed 12 November 2018).
Biostars. https://www.biostars.org/p/225683/ (Accessed 5 Aug 2019).
Bolger, A.M., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Oxford J. https://doi.org/10.1093/bioinformatics/btu170.
Bouchot, J.-L., Trimble, W.L., Ditzler, G., Lan, Y., Essinger, S., Rosen, G., 2014. Advances in machine learning for processing and comparison of metagenomic data. Comput. Syst. Biol. https://doi.org/10.1016/b978-0-12-405926-9.00014-9.
Bushnell, B., 2017. BBMerge – accurate paired shotgun read merging via overlap. PLoS One. https://doi.org/10.1371/journal.pone.0185056.
Collins, R.A., Wangensteen, O.S., O'Gorman, E.J., et al., 2018. Persistence of environmental DNA in marine systems. Commun. Biol. 1, 185. https://doi.org/10.1038/s42003-018-0192-6.
Fuller, C., Middendorf, L., Benner, S., et al., 2009. The challenges of sequencing by synthesis. Nat. Biotechnol. https://doi.org/10.1038/nbt.1585.
Hugenholz, P., et al., 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. 180 (18), 4765–4774 (PMC 107498).
Liu, B.C.O.P.E., 2012. An accurate k-mer-based pair-end reads connection tool to facilitate genome assembly. Bioinformatics. https://doi.org/10.1093/bioinformatics/bts563.
MacManes, M., 2014. On the optimal trimming of high-throughput mRNA sequence data. Front. Genet. https://doi.org/10.3389/fgene.2014.00013.
Magoč, T., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Oxford J. https://doi.org/10.1093/bioinformatics/btr507.
Masella, A.P., 2012. PANDAseq: paired-end assembler for illumina sequences. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-13-31.
Metagenomic Analysis of Environmental Water Samples With the NextSeq® 500 System. https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/appnote-metagenomics.pdf, 2015–. (Accessed 5 August 2019).
NCBI-SRA. https://www.ncbi.nlm.nih.gov/sra (Accessed 12 Nov 2018).
ResearchGate. https://www.researchgate.net/post/Quality_trimming_of_16S_MiSeq_data_before_or_after_merging_paired_reads (Accessed 5 Aug 2019).
Rognes, 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4. https://doi.org/10.7717/peerj.2584 e2584.eCollection.
Seymour, M., 2019. Rapid progression and future of environmental DNA research. Commun. Biol. 2, 80. https://doi.org/10.1038/s42003-019-0330-9.