Comparison of whole genome sequencing with Illumina short-reads and Oxford Nanopore of the E.Coli O157:H7 strain

Abstract

The pathologic, enterohemorrhagic O157:H7 strain of E. Coli is pathogenic to humans, spreading by contaminated food or water. It produces grave symptoms, often leading to hospitalizations. Its treatment is complex, as it can release Shiga toxins in response to medications. Thus, accurate detection of food and diseased patients is crucial for outbreak prevention and control. Here, we compare the use of Illumina short-read and Oxford Nanopore sequencing for the analysis of the genome of the E.Coli O157:H7 strain. An accurate tool for genomic inquiry is needed for its future detection and understanding. Examining the reads with modern software revealed a wider base detection by Oxford Nanopore, with higher contig and assembly length. However, Illumina assembly was superior in its accuracy, with fewer mismatches and misassemblies. Hence, the choice of the technique must be multifactorial and dependent on the primary goal of the investigation. Finally, such a comparison should be repeated after further improvements of Oxford Nanopore.

Conclusions

The addition of the Oxford Nanopore scaffold to the short-read Illumina during assembly improved the overall assembly length and GC content, also giving more insight into base-pair order on the large scale. Still, Illumina without Oxford Nanopore produced an assembly with fewer discrepancies with the reference genome and remains superior for applications requiring high accuracy.

Introduction

Escherichia coli is a widespread bacteria, present in the intestines of animals under healthy conditions, including in humans. Most, but not all, of E.Coli strains are harmless to humans. Ingestion of the pathogenic strains in varied threshold quantities causes disease. E. Coli O157:H7 strain is of particular notice, since as little as 10 – 100 colony-forming units, around 10,000 fewer than for other pathological strains (*Rahal et al, 2012*). Being foodborne, it is often ingested with contaminated water, dairy, or meat, as in the case of the Oregon and Michigan food poisoning, when it was first isolated (*Riley et al, 1983*). It is an enterohemorrhagic bacteria, able to produce Shiga toxins, likely in response to stress. Moreover, treatment with certain antibiotics caused greater toxin release (Grif et al, 1998; Zhang et al, 2000). The toxins degrade the epithelial layer, leading to hemorrhage and bloody diarrhea. Noteworthy, only one of the two possible toxins is neutralized by antisera to Shiga toxin from Shigella dysenteriae (Rahal et al, 2012). The possibility of severe presentation of the infection can cause hospitalization, possibly with hemolytic uremic disorder (HUS), kidney failure, or even death (Manasa, 2016; Rahal et al, 2012). Hence development of prevention methods, an accurate diagnosis, and effective treatment are important.

Understanding the action of the O157:H7 strain is complex due to its differential expression patterns in hosts (Lowe et al, 2009; Rashid et al, 2006) and varied toxin combinations (Grif et al, 1998; Rahal et al, 2012). The development of next-genome sequencing technologies shed light on its plasmid and its distinguishing features. As outlined by Orlek and colleagues (2017), that effort was led by PacBio in 2014, with the addition of Illumina in 2015 and Oxford Nanopore in 2016. Unraveling the genetic features of the strain can directly aid in diagnosis, food quality screening, and development of treatments. Currently, the presence of bacteria is confirmed with plate culture, reaction to sorbitol, serologic typing, and testing for Shiga genes stx1 and stx2 (Human Foods Program, 2024). Alternatively, sequencing, especially with Oxford Nanopore, would provide a faster and more sensitive alternative. Indirectly, it can also help in counteracting the antibiotic resistance growing within our population (Murray, 2022; Ventola, 2015) by identifying the resistance genes and studying the genetic evolution of bacteria, as done before (Ohnishi et al, 1999; Orlek, 2017; Perna et al, 2001).

Therefore, the following study examines the results of second-generation Illumina and third-generation Oxford Nanopore sequencing of the E. Coli O157:H7 strain genome. In order to examine the quality of their application, it compares the single-base and overall quality of the reads between the methods, as well as the validity of their concomitant use.

Materials and methods

Analysis overview

The data for the report were next genome sequencing reads of E. Coli O157:H7 strain acquired with either Sanger Illumina 1.9 MiSeq or with the Oxford Nanopore Sequencing. The reads were analyzed on the Galaxy

server (Galaxy Version 3.15.3) according to the analysis pipeline in ***Figure 1***, which covered the assembly, trimming, quality, completeness, and accuracy assessment in FastQC (version 0.12.1; Andrews, 2010), TrimGalore (version: 0.6.7; Krueger, 2023), SPAdes (version: 3.15.4; Bankevich et al, 2012) and Quast.

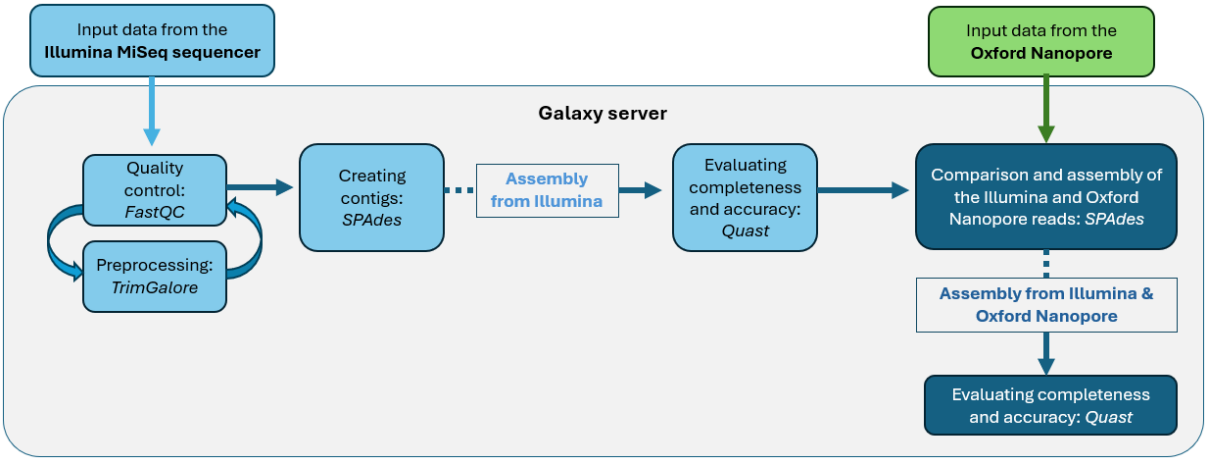


Figure 1. Pipeline of the analysis of the sequencing data. *The reads from Illumina MiSeq sequencer underwent preprocessing and assembly – quality control in FastQC before and after trimming of the reads in TrimGalore, matching of the contigs in SPAdes and the assessment of the assembly in Quast. SPAdes was run with default parameters, including an automatically detected k-mer length based on sequence length, with assembly and error corrections (BayesHammer, Nikolenko et al, 2012). The assembled contigs from SPAdes are compared to the provided E. Coli O157:H7 strain EDL933 reference genome to assess the completeness and accuracy of the assembly. After this point, the result was compared against the reads from Oxford Nanopore. Finally, both reads were assembled and examined in Quast. All software was accessed via the Galaxy server (Galaxy Version 3.15.3).*

FastQC

Firstly, FastQC provides an overview of the accuracy of the read – a per-base quality score obtained from the QC filter of the sequencer for each read. The scores are shown as Phred Quality (Q), based on the base calling error probability (P), and calculated as: $Q = -10\log_{10}P$. The software displays the Q score per base pair (bp) at a set position in read and a distribution of read based on the score. Secondly, one can assess the quality of the reads based on the nucleotide content, namely the GC content, the distribution of specific nucleotides throughout the read, and the percentage of unidentified bases.

TrimGalore

The software detects the remainder of the adapter sequence at the 3' end of the read and the low-quality reads, therefore increasing the quality of the final result. The minimum adapter overlap used was 1 bp. Discarded were reads of Phred Quality Score below 20; the Cutadapt algorithm detects the bases from the end of the sequence and stops computing when it finds an accepted value (Martin, 2010). The paired reads of the length below 100 bp were also removed.

SPAdes

SPAdes assemble the genome based on the overlaps of contigs from trimmed forward and reverse sequence reads. The software was run with default parameters with assembly and error corrections (BayesHammer, Nikolenko et al, 2012). The initial error correction removed rare k-mers at the contig assembly stage, as their sparsity likely results from sequencing error. Post-assembly error correction compared contigs with their consecutive reads' consensus base and changed the contig's base into the consensus base when they were not equal. The hereby received assembly is assessed further in Quast (***Figure 1***), by comparing it to the reference genome. The lower threshold for contig length was set at 1000 bp.

Results

Illumina preprocessing & assembly

The E.Coli O157 H7 sequence comprised a total of 104.2 Mbp. Its GC content was equal to 50% with normal distribution across reads, a value consistent with this bacteria. The overall quality of the Illumina reads was identified as good, with mean Phred Q equal to 33 and 9 for forward and reverse strands respectively. The quality of the read was lower towards the end of the read due to signal decay or phasing, which resulted in failed per base signal quality for both strands. Apart from the expected initial unequal distribution of bases due to the biased sequence composition, the reads have an overrepresentation of C bases at the end, likely due to the presence of adapters and lower quality of reads. For the forward strand, FastQC has identified an overrepresented sequence of 386 bp as a TruSeq Adapter. Almost all the reads (98.3% and 98.0% for forward and reverse strands, respectively) were unique, with a high uniformity of length.

Table 1. Descriptive statistics of reads of E. Coli O157 H7 from Illumina MiSeq sequencing pre- and post-trimming, and from Oxford Nanopore. *The reads were obtained from Illumina MiSeq sequencer and analysed with FastQC on the Galaxy server. No sequences were flagged as poor quality.*

Strand	Forward		Reverse	
Trimming status	before	after	before	after
Total number of reads	347,436	333,828	347,436	333,828
Total bases (Mbp)	104.2	90.8	104.2	74.5
Sequence length	35-301	100-301	35-301	100-301
GC content	50%	50%	51%	50%

Trim Galore detected 82,304 reads with adapters for the forward strand and 85,661 reads for the reverse strand – 23.7% and 24.7% of the total, respectively. Furthermore, 9.9% (10,353,880 bp) of the forward strand base pairs did not pass the quality threshold (Q = 20). Even more reads were subpar for the reverse strand - 28,860,397 bp (24.7%). The most common base preceding adapters was C, consistent with the inequal base distribution. Finally, 3.9% of paired reads (13608 bp) were removed because of their short length.

The quality of the trimmed sequences exceeded their predecessors for both forward and reverse sequences (***Figure 2***). The discrepancy was most visible in the increased minimum sequence length, more proportional split across bases towards the end of the sequences, and a higher proportion of the reads with good average

base quality. The GC content of the sequence didn't change significantly, as displayed in ***Supplementary Figure 1***. The quality of the end of the sequence remains low and was omitted due to the

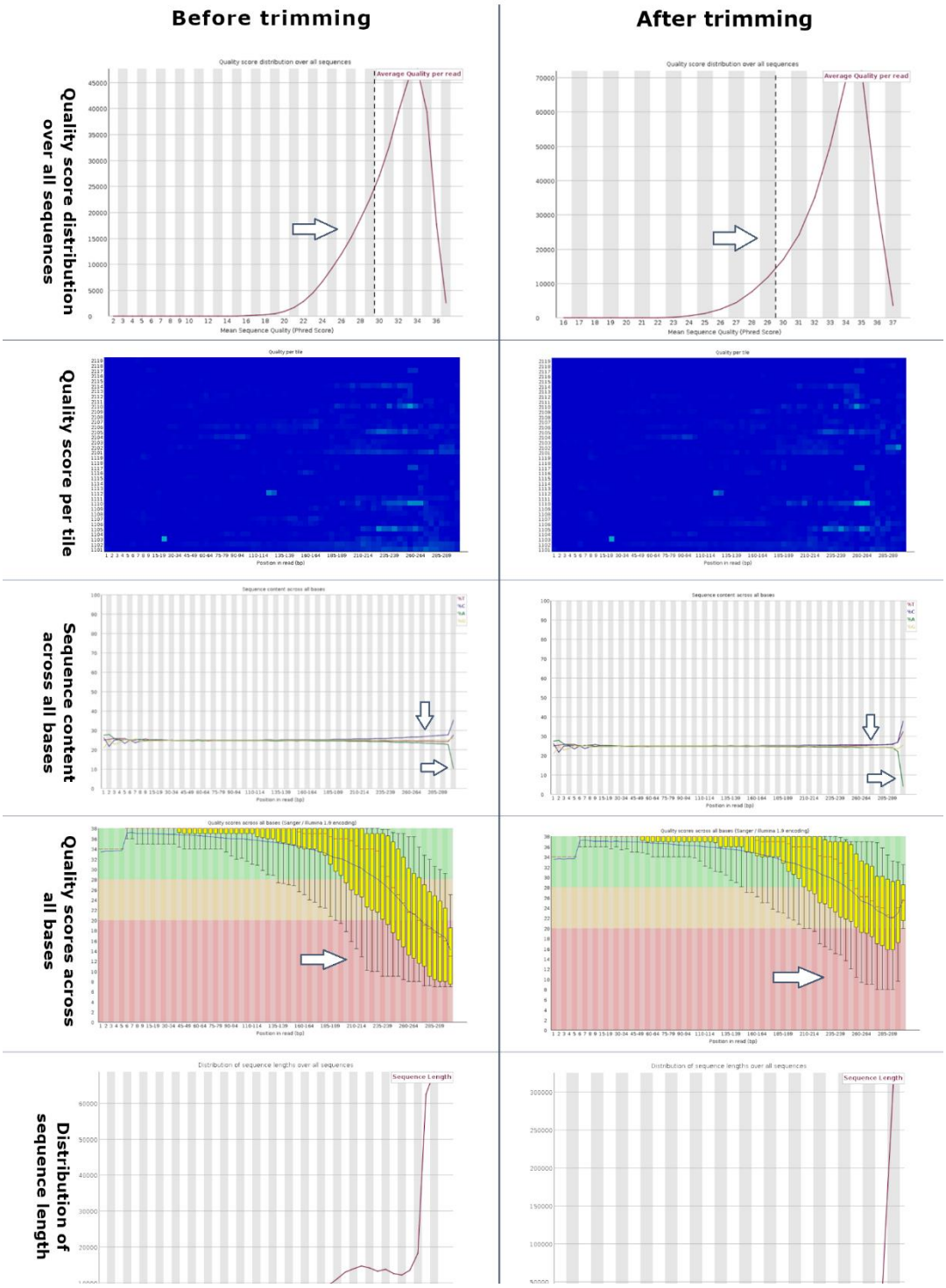


Figure 2. Quality measures of the reads of the forward strand of E. Coli O157 H7 before and after trimming. *The reads were obtained from the Illumina MiSeq sequencer, analysed with FastQC and trimmed with Trim Galore on the Galaxy server. The main discrepancies were marked with arrows. After trimming, the graph of the quality score (Phred Score) distribution was screwed to the right, with the higher proportion of reads exceeding 30. The quality per tile changed slightly due to the removal of some reads. The sequence content became more proportionally split into bases. Due to the removal of the adapters, the proportional of A bases at the end sharply decreased. The Trim Galore removed the base pairs of sequence length below 100, with the difference visible on the distribution of sequence length.*

cessation of computing at the subsequent high-quality read, as explained in the *Materials and Methods* section.

Using the trimmed reads, SPAdes assembled the contigs using the k-mer length of 21, 33, 55, 77, 99, and 127. Read error correction changed a total of 1604737 bases in 470290 reads. Similarly, the Illumina reads were assembled with the Nanopore read of the same genome using SPAdes.

Genome Assembly Evaluation

The reference sequence of the E. Coli O157:H7 provided consisted of 5,639,399 bp, 50.4% of them being GC. The assembly of Illumina and of Illumina and Oxford Nanopore from SPAdes were compared against it in Quast. The number of statistics was the same or similar for both the assemblies. Both had the duplication ratio of 1 and only 1 indel per 1 kbp. There was only one unaligned contig of 3433 bp for both assemblies. More common were the differences between the assemblies, outlined in ***Table 2***. The assembly of Illumina with Oxford Nanopore had a longer total length with a GC content slightly closer to the reference genome than that of Illumina alone. It also had fewer, longer contigs (***Figure 3B***). However, the longest of them, accounting for 79% of the total assembly length, was relocated. As outlined further on the coverage histogram in ***Figure 3A***, the assembly with Oxford Nanopore had a very high coverage depth for its long contig. For Illumina alone, the reads responsible for that depth were split into many shorter contigs of lower depth, providing a higher resolution of data.

Table 2. Comparison of the Quast statistics for the assembly of Illumina alone and with Oxford Nanopore (ON). A.Overall statistics*. The overall length of the assembly was longer for the Illumina & ON read than for Illumina alone, which equalled to 95.1% and 97.0% of the reference genome, respectively. The assembly consisted of much more contigs of shorter length for the latter – the longest contig accounted for 10% of its sequence. For Illuina with ON it was as much as 79%, hence the N50 value was equal to its length. Although similar, the GC content was slightly higher, and closer to the value of the reference genome, for Illumina & ON.* ***B. Misassemblies report.*** *The number of contig relocations was higher for ON, with one of them being the longest contig. Moreover, ON had a higher number of indels; most of them were over 5 bp long, which translated into a larger discrepancy in total indel length.*

A. Quast statistics	Illumina	Illumina + Nanopore
# contigs	95	12
Largest contig	505,119	4,233,969
Total length	5,219,547	5,389,753
% covered by the largest contig	10%	79%
N50	148,373	4,233,969
Genome fraction (%)	95.1	97.0
GC (%)	50.3	50.4
B. Misassemblies report		
# misassemblies: contig relocations	1	3
Misassembled contigs length	98,856	4,658,887
# local misassemblies	3	2
# mismatches	325	850
# mismatches per 100 kbp	6	16
# indels	58	68
Indels length	64	117

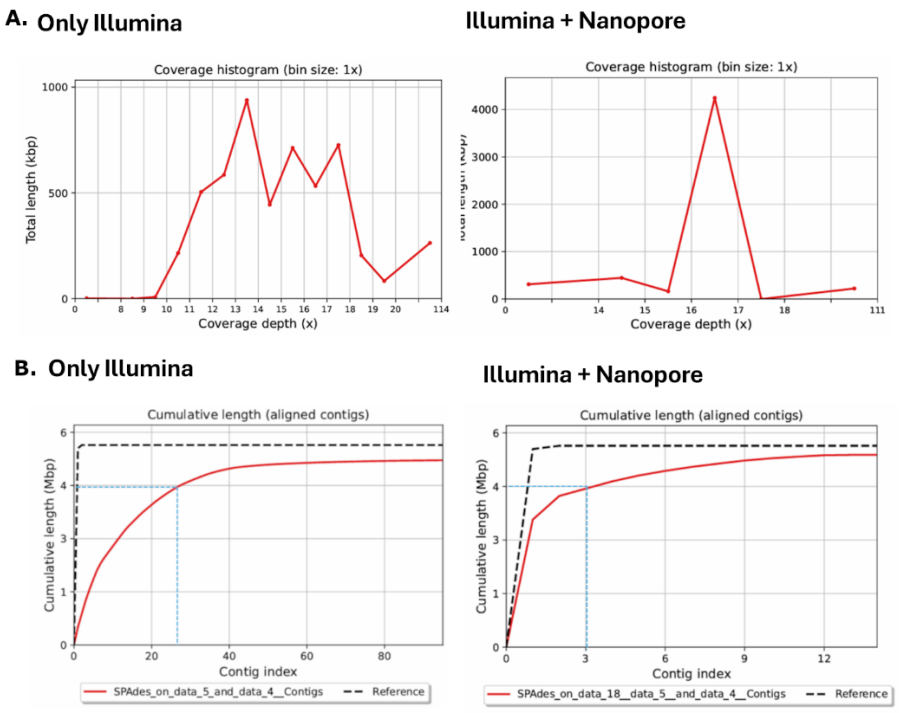


Figure 3. Visualisation of the coverage depth (A) and cumulative length of the assembly by contig (B) for either Illumina alone or Illumina with Oxford Nanopore (ON). *The graphs were produced by Quast on the Galaxy server.* ***A.*** *The assembly with Oxford Nanopore had a very high coverage depth for its long contig. For Illumina alone, the reads responsible for that depth were split into many shorter contigs of lower depth, providing higher resolution of data.* ***B.*** *The same cumulative length is achieved with fewer contigs for the assembly with ON. The length of 4 Mbp was marked with a dashed, blue line – such length was obtained with around 22 contigs for Illumina alone vs. 3 contigs for Illumina with ON. The assembly of the Illumina alone accounts for 95% of the reference genome (black, dashed line), as compared to 97% for Illumina & ON.*

Discussion

The investigation of the E.Coli O157:H7 genome using Illumina pair-end short reads alone and with the Oxford Nanopore revealed a number of differences in the quality of the assemblies.

Firstly, the addition of the Oxford Nanopore read allowed for the assembly to increase in total length of the read. Reads with fewer, longer contigs a wider ordering of base pairs, and hence also genes, in respect to each other. That might serve to detect long genes in its genome, such as for a 16S rRNA gene frequently used for microbial identification in laboratories. A successful differentiation of the O157:H7 strain would aid in detecting and controlling food contamination events (*Ohnishi et al, 2000)*. Though largely replaced with mass spectrometry, such a technique could also be used in the clinic (*Church et al, 2020*). However, the larger read and assembly size came with a trade-off of accuracy. The misassembles and mismatched were more common, also affecting the largest contig. Its relocation was likely due to a misread of a repeated base in the Nanopore. Although the quality of Illumina's sequence was also poor towards the end of the read, trimming and the interpretation of the quality per tile allowed for an overall good assembly of the

genome. For applications requiring a precise detection of a short genetic sequence, provided it is not located at the replicon's ends, Illumina assembly might be more suited.

Overall, the choice of technique employed for bacterial genome sequencing is multifactorial, largely dictated by the need for accuracy and contig length. The presence of a reference genome is also crucial, as in its absence the relocated read can't be moved, leading to incorrect interpretations. Finally, it's worth noting that Oxford Nanopore is still a relatively recent development and its accuracy is likely to be improved with time.

References

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc

Bankevich, A., et al (2012). "SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing," Journal of Computational Biology, 19 (5), pp. 455-477. doi:10.1089/cmb.2012.0021]

Church, L.D. *et al.* (2020) "Performance and Application of 16S rRNA Gene Cycle Sequencing for Routine Identification of Bacteria in the Clinical Microbiology Laboratory," *Clinical Microbiology Reviews*, 33(4), p. 10.1128/cmr.00053-19. Available at: https://doi.org/10.1128/cmr.00053-19.

Devi, M. (2016) "Pathology, Ecology and Infection of E.coli," *Research and Reviews Journal of Medical and Health Sciences RRJMHS*, 5.

Grif, K. *et al.* (1998) "Strain-Specific Differences in the Amount of Shiga Toxin Released from Enterohemorrhagic Escherichia coli O157 following Exposure to Subinhibitory Concentrations of Antimicrobial Agents," *European Journal of Clinical Microbiology and Infectious Diseases*, 17(11), pp. 761–766. Available at: https://doi.org/10.1007/s100960050181.

Gurevich, A. et al. (2013) "QUAST: Quality assessment tool for genome assemblies," Bioinformatics, 29(8), pp. 1072–1075. Available at: https://doi.org/10.1093/bioinformatics/btt086.

Human Foods Program (2024). Equivalent Testing Methodologies for Spent Sprout Irrigation Water. [online] U.S. Food and Drug Administration. Available at: https://www.fda.gov/food/laboratory-methods-food/equivalent-testing-methodologies-e-coli-o157h7-and-salmonella-spent-sprout-irrigation-water-or..

Krueger, F. (2023) „FelixKrueger/TrimGalore: v0.6.10 - add default decompression path". Zenodo. doi: 10.5281/zenodo.7598955.

Lee Ventola, C. (2015) *The Antibiotic Resistance Crisis Part 1: Causes and Threats*.

Lowe, S.R.M. *et al.* (2009) "Escherichia coli O157:H7 Strain Origin, Lineage, and Shiga Toxin 2 Expression Affect Colonization of Cattle," *Applied and Environmental Microbiology*, 75(15), pp. 5074–5081. Available at: https://doi.org/10.1128/AEM.00391-09.

Martin, M. (2010) Algorithm details, Algorithm details - Cutadapt 0.1 documentation. Available at: https://cutadapt.readthedocs.io/en/stable/algorithms.html#quality-trimming-algorithm (Accessed: 18 November 2024).

Murray, C.J. *et al.* (2022) "Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis," *The Lancet*, 399(10325), pp. 629–655. Available at: https://doi.org/10.1016/S0140-6736(21)02724-0.

Nikolenko, S.I., Korobeynikov, A.I. and Alekseyev, M.A. (2012) “BayesHammer: Bayesian clustering for error correction in single-cell sequencing.” Available at: https://doi.org/10.1186/1471-2164-14-S1-S7.

Ohnishi, M. *et al.* (1999) “Chromosome of the Enterohemorrhagic Escherichia coli O157:H7; Comparative Analysis with K-12 MG1655 Revealed the Acquisition of a Large Amount of Foreign DNAs,” *DNA Research*, 6(6), pp. 361–368. Available at: https://doi.org/10.1093/dnares/6.6.361.

Ohnishi, M. *et al.* (2000) “Comparative Analysis of the Whole Set of rRNA Operons Between an Enterohemorrhagic Escherichia coli O157:H7 Sakai Strain and an Escherichia coli K-12 Strain MG1655,” *Systematic and Applied Microbiology*, 23(3), pp. 315–324. Available at: https://doi.org/https://doi.org/10.1016/S0723-2020(00)80059-4.

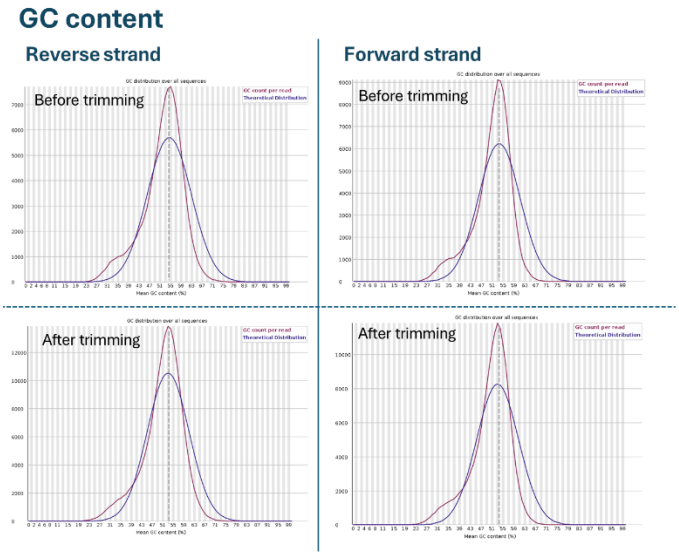
Orlek, A. *et al.* (2017) “Ordering the mob: Insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids,” *Plasmid*, 91, pp. 42–52. Available at: https://doi.org/https://doi.org/10.1016/j.plasmid.2017.03.002.

Perna, N.T. *et al.* (2001) “Genome sequence of enterohaemorrhagic Escherichia coli O157:H7,” *Nature*, 409(6819), pp. 529–533. Available at: https://doi.org/10.1038/35054089.

Rahal, E.A. *et al.* (2012) “Escherichia coli O157:H7—Clinical aspects and novel treatment approaches,” *Frontiers in Cellular and Infection Microbiology*, 2. Available at: https://www.frontiersin.org/journals/cellular-and-infection-microbiology/articles/10.3389/fcimb.2012.00138.

Rashid, R.A. *et al.* (2006) “Expression of putative virulence factors of Escherichia coli O157:H7 differs in bovine and human infections,” *Infection and Immunity*, 74(7), pp. 4142–4148. Available at: https://doi.org/10.1128/IAI.00299-06.

Riley, L.W. *et al.* (1983) “Hemorrhagic Colitis Associated with a Rare Escherichia coli Serotype,” *New England Journal of Medicine*, 308(12), pp. 681–685. Available at: https://doi.org/10.1056/NEJM198303243081203.



Supplementary Figure 1. GC content of the forward and reverse strand before and after trimming of the E. Coli O157 H7 sequence reads. The reads were obtained with Illumina MiSeq and analysed with FastQC on the Galaxy server, pre- and post-trimming with Trim Galore. All readings present with a normal distribution, unaffected by trimming. There is an increased proportion of low GC-content reads, likely due to the decreased quality of the reads towards the end of the recording.