Forming Bioinformatics Pipelines: Characterization Analysis of *Clostridioides difficile* Illumina Reads through De-Novo Omics Assemblies and Protein Prediction.

James Martinez

College of Science, Northeastern University

BINF 6308: Bioinformatics Computational Methods 1

Dr. Stefan Kaluziak

December 16, 2022

Introduction

The following project pipeline had been primarily created with the organism, *Clostridioides difficile*, in mind. *Clostridioides difficile* is a low GC content gram-positive bacteria primarily associated with life threatening and infectious diarrhea and other gastrointestinal diseases such as colitis. Increased infectivity, lethality, and multi-drug-resistance of newer strains such as *Clostridioides difficile 630,* have left researchers on a scramble to identify new potential protein targets to prevent its spread. To assist real-world lab testing through in-silico methods, the following pipeline was created to: (1) produce a de-novo Genome Assembly from a strain's reads, (2) align RNA-seq data to generate a Transcriptome Assembly, and (3) assign Gene Annotations through protein predictions. The hope is to refine the current pipeline for future usage towards (4) protein function analysis and pathway association, to direct lab scientists towards significant areas to study.

The data we have collected to test the first iteration of the pipelines comes from various sources. For Genome Assembly, SRR22386611, was a Illumina paired reads run from a DNA library generated at Hangzhou Medical College of the strain NB647, little else is described. For Transcriptome Assembly, SRR21284802, SRR21284803, SRR21284804, SRR21284805, SRR21284806, SRR21284807, were part of a Illumina paired reads run from an experiment at Tufts University School of Medicine on *Clostridioides difficile 630* wild type variants and ddl gene in-frame mutation variants (Belitsky, 2022). The purpose of the experiment was to investigate the relation between the ddl gene and Van operons, on the vancomycin resistance potential of *Clostridioides difficile 630* through peptidoglycan formation; the end results showed knockout did not significantly alter low levels of vancomycin resistance, but however did support pathway interaction between the genes for peptide D-Ala-D-Ala assimilation into

peptidoglycan instead of D-Ala-D-Ser. For future refinement of the part of the pipeline, files for the reference guided transcriptome assembly were provided for the Refseq genome file of *Clostridioides difficile 630*: GCF_000009205.2_ASM920v2_genomic.fna; ASM920v2; RefSeq GCF_000009205.2. The following pipeline runs from the scratch directory and is composed of four scripts to be called in sequential order: (1) sbatch_assemble_genome.sh , (2) sbatch_alignRNAseq.sh, (3) sbatch_trinity.sh, and (4) sbatch_transdecoder.sh. As a heads up before starting, the pipeline optimally stores all data in its execution directory, meaning that while the analysis does copy data to "/home", it is best organized in "/scratch". Finally some files that are called in the program are saved in share class workspaces rather than with the user, so such links were hard coded in; and the Refseq genome if not downloaded on one's own may use the path: "/home/martinez.jam/GCF_000009205.2_ASM920v2_genomic.fna".

Results

The Genomic Assembly was completed in one step using the command line tools and programs: SRA Toolkit 3.0.2: "fasterq-dump", Trimmomatic-0.39-2 (Bolger, 2014), Spades 3.15.4 (Bankevich, 2012), QUAST 5.2.0 (Gurevich, 2013). Using the fasterq-dump we split the read data from the SRR22386611 run, into paired left and right reads, and sent the unpaired to their own directory. Using Trimmomatic, the paired reads were trimmed with mostly default parameters, such as using the phred 33 scoring system, TruSeq3 adapters to be trimmed, setting leading and trailing to 20, sliding window to 4 and 30, and finally making read minimum length to 36. Those trimmed paired reads were then assembled into mapped contigs using Spades, that once completed were assessed using QUAST. The following graphs are meant to visualize our end results:

Report

	contigs
# contigs (>= 0 bp)	237
# contigs (>= 1000 bp)	37
# contigs (>= 5000 bp)	31
# contigs (>= 10000 bp)	28
# contigs (>= 25000 bp)	25
# contigs (>= 50000 bp)	21
Total length (>= 0 bp)	4049016
Total length (>= 1000 bp)	4004362
Total length (>= 5000 bp)	3987331
Total length (>= 10000 bp)	3967028
Total length (>= 25000 bp)	3917116
Total length (>= 50000 bp)	3760954
# contigs	43
Largest contig	549417
Total length	4008790
GC (%)	28.37
N50	208886
N75	138362
L50	7
L75	12
# N's per 100 kbp	0.00

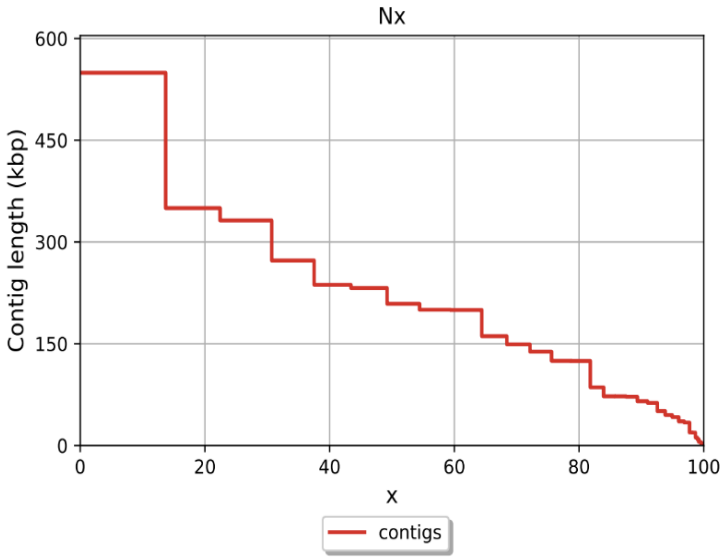

Figure 1 & 2. QUAST Assessment of SPAdes Results

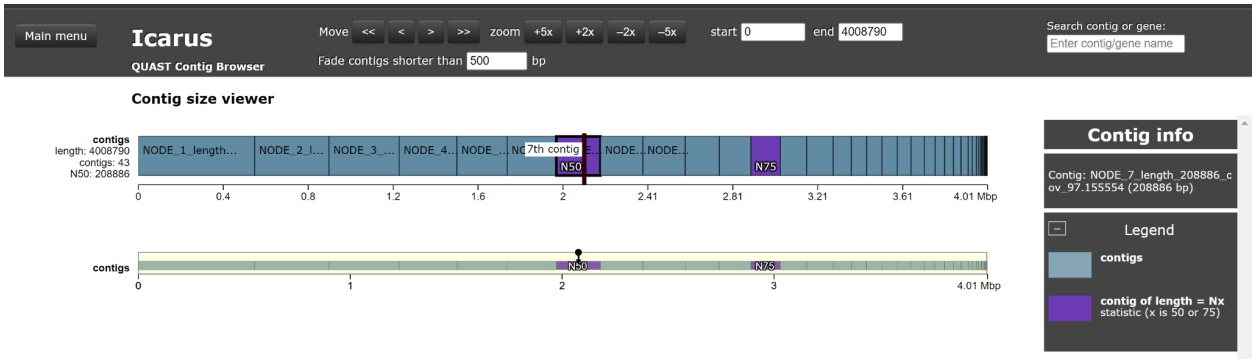

Figure 3. Icarus View of Aligned Genome Assembly

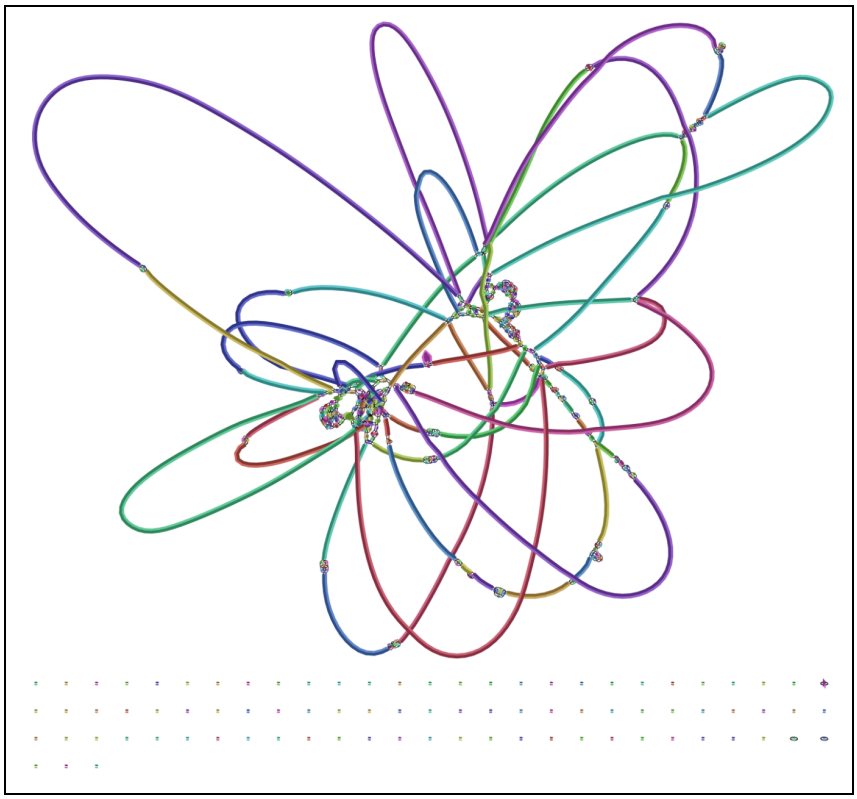

Figure 4. Bandage De-Bruijn Graph Representation of contigs.fasta File From SPAdes

The QUAST results supported the genome assembly for *Clostridioides difficile* strain NB647 using SPAdes to be good, indicative by containing 43 contigs, but more importantly having a generally high N50 score of 208,886 at the 7th contig. “N50 is a metric widely used to assess the contiguity of an assembly, which is defined by the length of the shortest contig for which longer and equal length contigs cover at least 50 % of the assembly”. (Alhakami 2017) Comparison with other assemblers can be used to determine if SPAdes was truly the best option for the assembly, however on its own it is clear that as of now that only the Trimmomatic step needs consideration. Proper trimming needs the full experimental details such as the correct adapter library to identify these sequences in the reads.

The Transcriptome Assembly consisted of two parts, focused on transcriptome assembly based on reference and de-novo methods. All next steps use reads purely from the Tufts

University RNA-seq experiment shown in the introduction, For the reference assembly, the first part required usage of SRA Toolkit 3.0.2: “fasterq-dump”, Trimmomatic-0.39-2, GMAP and GSNAP Version 2021-12-17 (Wu, 2016), and SAMtools v1.4 (Li, 2009). Similar to before, fasterq-dump was used to retrieve the read runs from SRR21284802, SRR21284803, SRR21284804, SRR21284805, SRR21284806, and SRR2128480. GMAP used the Refseq genome file GCF_000009205.2_ASM920v2 for *Clostridioides difficile 630* to create an index database to align the reads too in the future. Trimmomatic used the exact parameters as before used in the genome assembly to trim the RNA-seq reads, again because the adapter sequences could not be identified, yet much of the quality was similar. GSNAP is then used with the paired reads and previously made GMAP database to try to align the reads into a sam file, in this case even though we are looking at bacteria, the N -1 splice site search flag was kept by instructor recommendation. SAMtools was used to convert the alignment files into binary format and then index them to create a comprehensive order for the reads to be fully assembled.

Part 2 of the transcriptome step contains usage of the previous part’s sorted BAM files to create the reference based assembly, and then using the original reads generating the de-novo assembly, through Trinity. The tools used in this step are SAMtools v1.4 and Trinity v2.15.0 (Haas, 2013). The reference assembly method continued by merging all the previously indexed BAM files into a singular one, and then calling Trinity on it with a parameter to set the max intron size to be 10,000. The de-novo assembly used a list of both the left and right paired reads originally created from the fasterq-dump at the time of Trinity’s execution, other important parameters set were to sequence type to fastq and the CPU set to 4 to allow high parallelization and speed up of the program. In both cases, the final transcriptomes were assessed using the built in TrinityStats.pl program to generate gene and transcript counts with contig statistics.

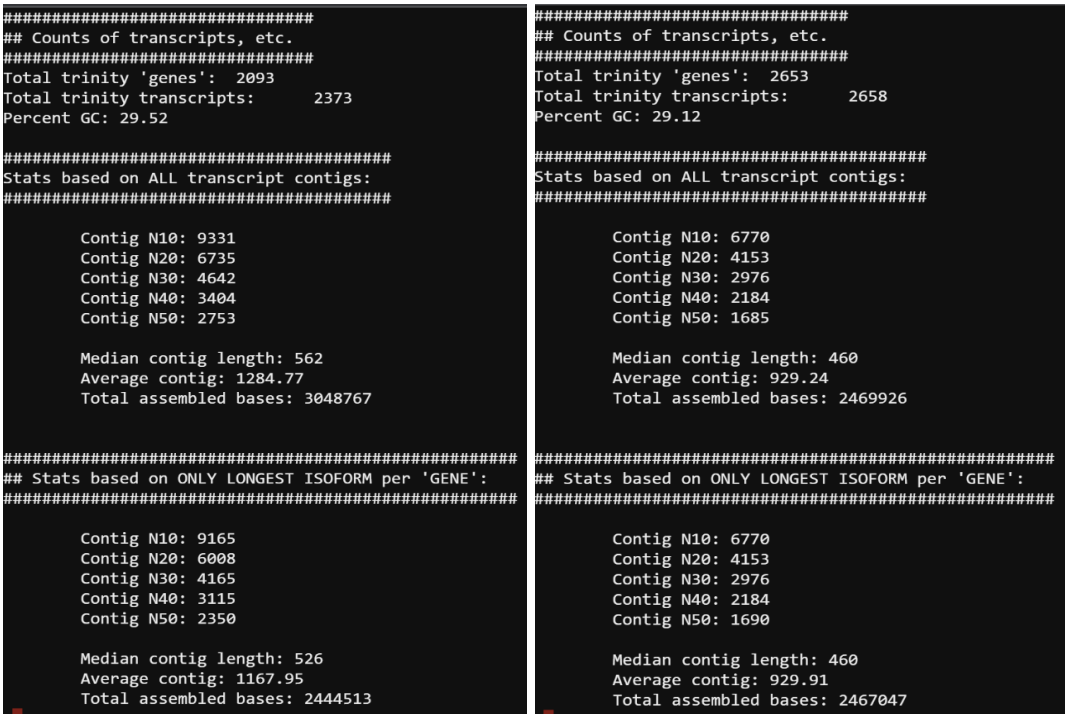

Figure 5 & 6. Comparison of De-Novo Vs. Reference Based Transcriptome Assemblies

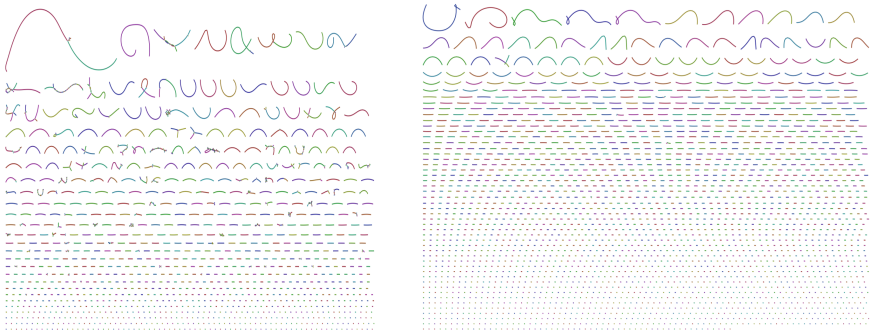

Figure 7 & 8. De Novo vs. Reference Fasta File Visualization Through Bandage

Visualization of the end statistics seem to imply that both transcriptomes generated did contain much different results, mainly seen through the longest isoform per gene contig N50 of 2350 for the de novo over the 1690 value for the reference. Although a transcriptome's N50 is much different from a genome's N50 (better to use ExN50), by using the longest isoforms an assumption can be made supporting de-novo performed best, so it was chosen for protein prediction. Quick Analysis of the end result de novo fasta file was performed below.

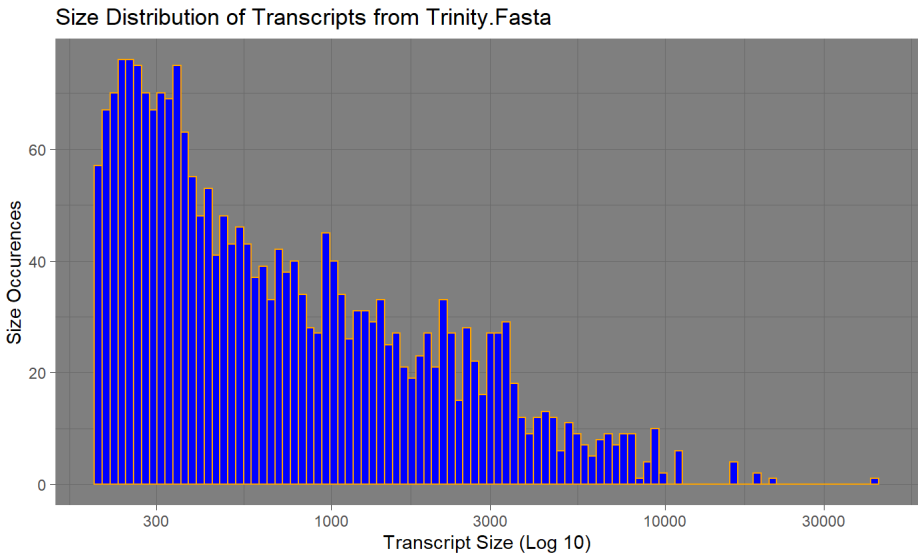

Figure 9. De Novo Transcriptome Assembly Transcript Size Distribution

The reference based transcriptome however, does seem to have high specificity due to the amount of transcripts being nearly equivalent to the known genes, potentially better for analyzing much similar strains.

The part three Protein Prediction step of the pipeline uses the 2,373 transcripts from the Trinity.fasta file to identify potential open reading frames that can correspond to potential proteins. The programs executed in this step were BLAST 2.9.0 (McGinnis, 2004), TransDecoder 5.6.0, and HMMER v3.3.2 (Finn, 2011). Beginning with long_orfs.py, the program was kept at default since it only needed the Trinity.fasta file to generate the longest ORF's from the transcriptome; at run-time no changes seemed necessary. The resulting .pep file containing all the longest ORFs was then compared against the Swissprot database to identify potential matches, with parameters set to return only at max one significant match, with an e value minimum of 1e-5, and to be returned in outfmt 6 tabular output. Besides assigning cores for setting parallelization for the script's execution, a finding of note was that in some cases the max_target_seqs doesn't always give the best e-value results, rather it selects the first based on

what the program identified as significant. Because this issue also occurs with its counterpart parameter num_alignments, it was noted that future iterations would simply need manual assessment to make sure the result is always the highest scoring pair. The end result of the blastp program was 2413 similar proteins. The pfam.Scan script remained much similar to it's default also, it used hmmscan to search for protein domains using the .pep file of longest ORF's, against the Pfam database, where it was then formatted a default tabular format, resulting in 24,002 domains. The three previous programs were used in the execution of Transcoder.Predict in the predictProteins.sh file, with the .pep file noticeably being self tracked in memory. Besides adding back the original Trinity.Fasta file, the other two optional retain_pfam_hits and retain_blast_hits parameters were added to refine the predicted protein search. The resulting .pep file resulted in 3087 coding sequences that could be used to identify potential similar proteins. The final step was to perform another blastp run, where the parameters were changed to have num_alignments return only 1 and the e-value set 1e-10 to filter insignificant matches. The end result was 2138 significant predicted proteins from the original *Clostridioides difficile* transcriptome, and was saved to alignPredicted.txt.

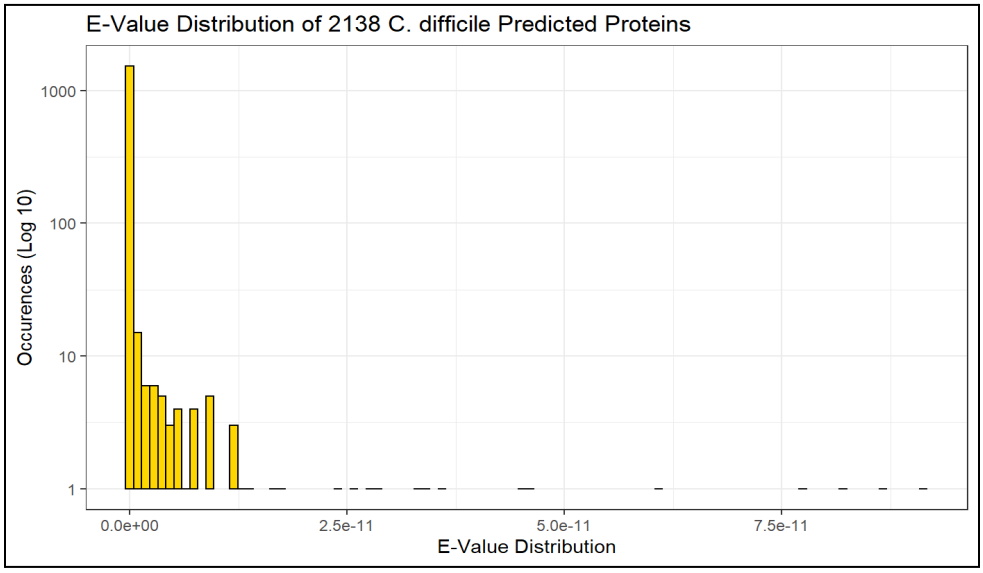

Figure 10. E-value Significance Distribution of Final Aligned Protein Predictions

E-values closer to 0 are indicative of significant results, more precisely it's the chances that a sequence could be randomly selected from the total set of sequences in the database. Since we used Swissprot these significance values were based on a total of 500,000+ sequences. It is expected that for less characterized samples that the e-values would be much further away, however since the observed *Clostridioides difficile* 630 has been widely researched, many proteins for it already exist, therefore making it likely to discover nearly identical matches or very similar ones. Using the set of significant predicted proteins, the next users can have a basis of Protein ID's to search against a functional annotation database to begin characterizing the proteome.

Conclusions & Future Consideration

The iteration of our genome assembly was a seemingly good attempt, the N50 value of 208,886 bp's correlating with the 7th contig inferred its good fit. Future iterations of the genome assembly step could be improved by altering the Trimmomatic step. The current iteration uses default parameters due to lack of additional annotations to the original sequence run, however next time identifying the adapters used in a run could assist in their precise trimming. Alteration of the minimum length and score of read would also be necessary when dealing with lower quality runs, which wasn't addressed in this pipeline due to high quality.

Tweaks of the transcriptome aspect of the pipeline can be performed to improve future iterations. For instance when viewing *Clostridioides difficile* strains with high sequence similarity to existing strains, reference guided assembly seems to be a viable method of exploration. Comparisons of our N50 value for both transcriptome assemblies show the de-novo achieving the higher value, however N50 is less meaningful for transcriptomes and might be

better analyzed by ExNx values. Judging from other metrics such as genes to transcript produced, smaller GC content, and high removal of bases, evidence supports the reference assembly providing a possible specific/better representation. However the best way to determine significance would be to investigate ExN50, BUSCO, and FastQC programs to generate extra visualizations to support comparisons of the transcriptome.

While we ended the pipeline with 2138 significant predicted proteins, the value is a culmination of the use of mostly default parameters used so far throughout the pipeline. Reviewing each script sequentially, there is actually very little that we would attempt to change. To increase the scope of high similarity pairs, since a few ORF's couldn't find similar proteins, it may help to change the blastPep.sh program to use a larger, specialized database such as NCBI Bacteria Database rather than Swissport. Increasing the e-value setting would be necessary as a consequence though, since the larger size might introduce low quality proteins as results while appearing "significant". Using other programs such as PSI-blast and DELTA-blast also could be suggested to find more distantly similar proteins. The final consideration for future implementation would include a conversation of direct de-novo genome assembly to protein prediction. In this case we see that a majority of the pipeline stays the same towards first creating a de-novo genome, then using that genome to create a reference transcriptome. Assuming the Transdecoder steps can only work with the Trinity transcripts, blastp, and hmscan, the only changes would possibly be to additionally use blastx(translated nucleotide blast search) and nhmmscan(dna search against DNA profiles) to cross reference with the original protein prediction results.

The user in the future will most likely continue from the end of the pipeline, meaning the predicted proteins will have functionality analysis performed on them. In this case aligning

protein results to pathway databases such as KEGG, can provide the user with protein functionality, orthologous proteins, and interaction pathway maps.

References

Alhakami, H., Mirebrahim, H., & Lonardi, S. (2017). A comparative evaluation of genome assembly reconciliation tools. Genome biology, 18(1), 93. https://doi.org/10.1186/s13059-017-1213-3

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology, 19(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Belitsky, B. R. (2022). VanG- and D-Ala-D-Ser-dependent peptidoglycan synthesis and vancomycin resistance in Clostridioides difficile. Molecular Microbiology, 118, 526– 540. https://doi.org/10.1111/mmi.14980

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics, 30(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic acids research, 39(Web Server issue), W29–W37. https://doi.org/10.1093/nar/gkr367

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics (Oxford, England), 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McGinnis, S., & Madden, T. L. (2004). BLAST: at the core of a powerful and diverse set of sequence analysis tools. Nucleic acids research, 32(Web Server issue), W20–W25. https://doi.org/10.1093/nar/gkh435

Wick, R. R., Schultz, M. B., Zobel, J., & Holt, K. E. (2015). Bandage: interactive visualization of de novo genome assemblies. Bioinformatics (Oxford, England), 31(20), 3350–3352. https://doi.org/10.1093/bioinformatics/btv383

Wu, T. D., Reeder, J., Lawrence, M., Becker, G., & Brauer, M. J. (2016). GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality. Statistical Genomics, 283–334. https://doi.org/10.1007/978-1-4939-3578-9_15