# Project 1: Final report

# Group: nkob

# Members: Heather Baldwin, Ida Holásková, Denzel Middleton

# Date: October 10, 2019

# Whole Genome Sequencing of *Neisseria gonorrhoeae*

# INTRODUCTION

In 2012, 78 million incidences of gonorrhea worldwide were reported to the World Health Organization (WHO) (*1*). Gonorrhea is a sexually transmitted disease (STD) transmitted by the bacterium *Neisseria gonorrhoeae (NG).* The *NG* colonizes in warm, moist regions of the female reproductive tract as well as the urethra in both males and females. Widespread and improper use of antibiotics have advanced antibiotic resistance of *NG* to the extent of extensively drug-resistant strains being identified (*2*). The objective of this study was to assemble the *NG* isolate, SRX6631047, and compare size (bp) of the assembled genome to the known reference *NG* genome available on NCBI.

# RESPONSE TO REVIEWERS

Authors would like to express gratitude to reviewers for their thorough review and constructive suggestions to improve our proposal. We took the following steps to advance the quality of the proposal: Specific objective of genome assembly was defined more clearly and narrowed down to only obtain highly contiguous genome of *NG* and compare the size (bp) to the reference genome. Source of data and sequencing methods (generated *vs.* retrieved) were clarified and mock methods not performed by our group were removed. The error with regard to calculation of coverage was corrected. The error was a result of misunderstanding as there are more than one definition of coverage (*3*, *4*). Lee and coworkers (*4*) utilized two types of coverage calculations; one, as a percentage of assembled genome to the reference genome and second; the coverage in terms of redundancy – a number of reads that align to, or ”cover” a known reference (reported as “median depth of coverage”). The *NG* genome we selected to assemble should have 107% coverage with respect to reference genome. After screening the quality of sequences, we were able to determine the “genome coverage”, defined as how repeatedly, on average, a reference sequence is covered by bases from the reads. Genome coverage is an important statistic because multiple observations per base are needed to obtain a reliable call. It is calculated by multiplying the number of sequences by two (due to paired-end sequencing) and by the length of the sequences, and the product divided by the reference genome size. The obtained genome coverage in our project yielded 124x based on raw, untrimmed reads (more details in results). It is difficult to find an optimal “threshold” of lowest accepted genome coverage, as it may depend on various parameters, however, the group that performed recent and detailed whole genome sequencing on *NG* isolates from about 400 patients in New Zealand, excluded from their analysis all isolates with lower than 45x coverage (*4*). Thus, we are confident that 124x genome coverage is a good starting point of the assembly pipeline.

More details on the version of the assembler software are now provided. The velvet assembler was chosen not only for its high speed and precision, but also for its reliability (since it’s been around and used for a long time). Velvet is available to use on Spruce and it can be used for the short paired-end *NG* whole genome sequencing (WGS) dataset. With respect to expected outcomes, more details on quality control (FASTQC) on various trimming combinations (using Trimmomatic), and numerous tests of k-mer size (Velvet optimization) were added and provided in the current report to resolve possible inaccuracies in the assembled genome.

# RESULTS

**Data retrieval-**

The dataset (SRX6631047) was obtained after a SRA search on NCBI for *NG.* Further investigation of the dataset revealed the sample was collected by the CDC from a clinical patient that tested positive for *NG* infection.

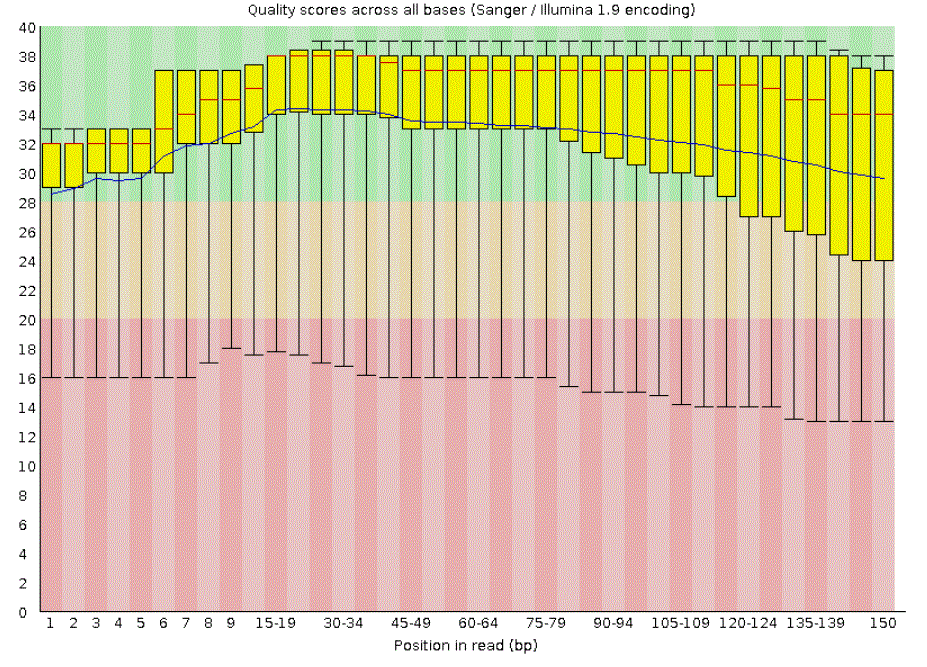
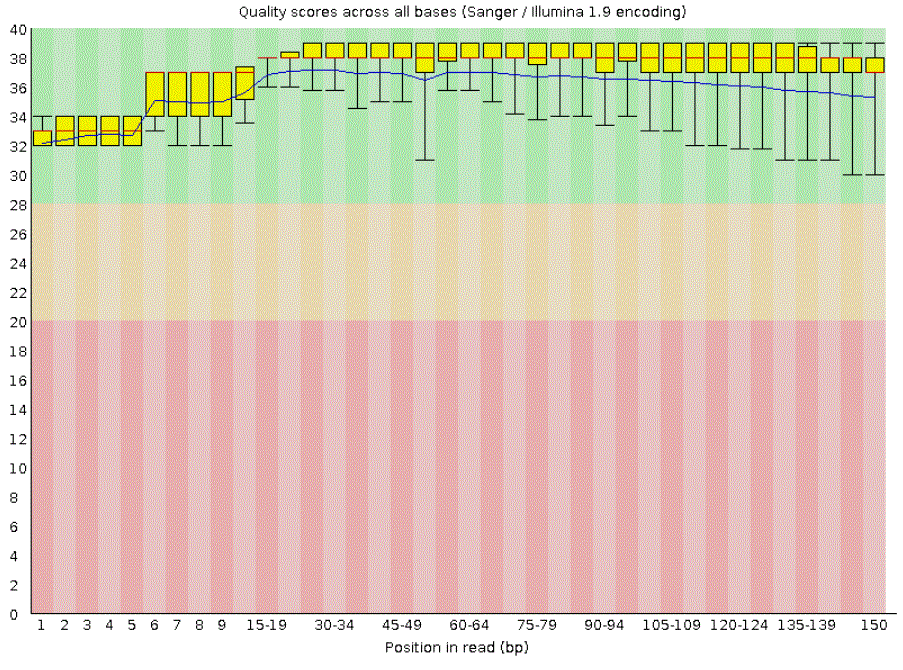
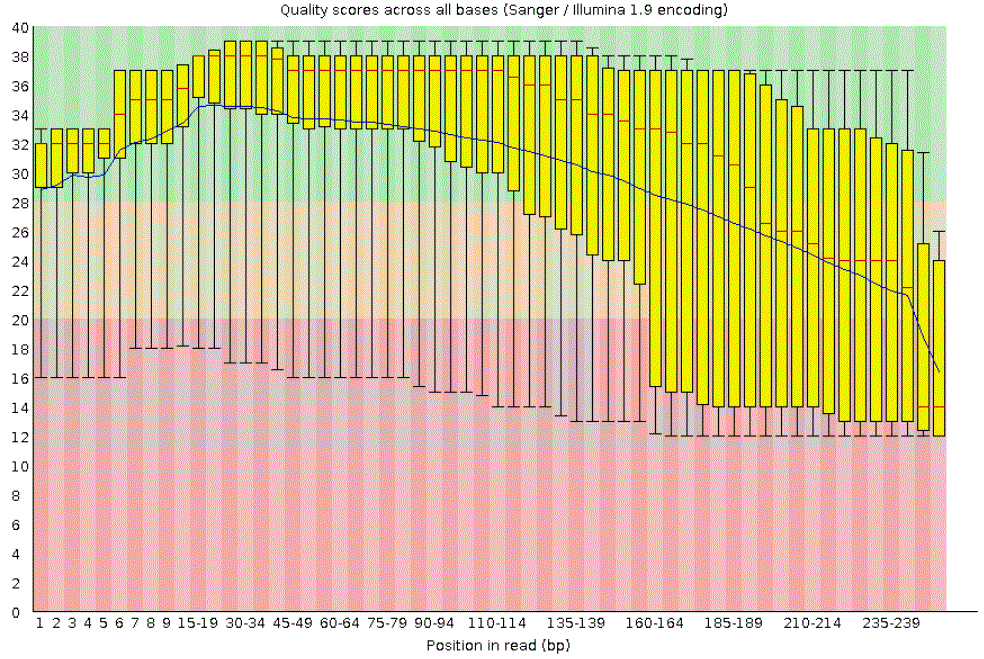
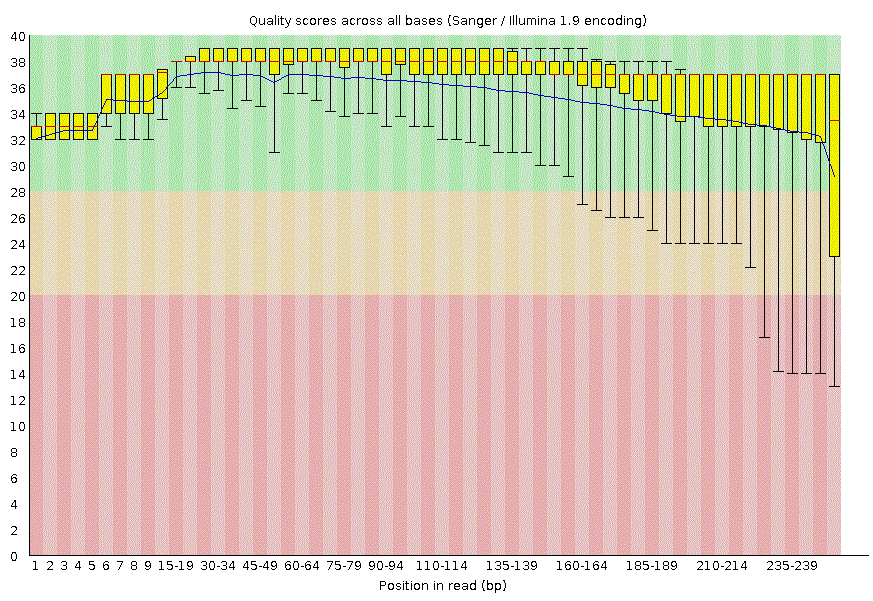
**Quality Control Processing-**

*Fastqc (v0.11.7)*

The original data set obtained from NCBI after quality control revealed a need for trimming. As observed in **Figure 1A** (left read) and **B** (right read), the bases >160 bp position were found in the red zone (low quality reads). With a possible threat to our mean coverage and to ensure the best outcome for our assembly, our group decided that in moving forward, the labor needed to be divided. One member of the group focused on trimming the data to improve its quality, another member started attempting the assembly on originally proposed dataset, while the third focused on finding a new dataset that had better quality scores. **Table 1** demonstrates the various outputs of Fastqc after trimming attempted by our group to identify the version of our dataset with the best quality score and reasonable genome coverage. The attempt to find a new data set on NCBI (6 new datasets) resulted in the same problem our group ran into with our original dataset—the left read from fastqc would comeback with much higher quality reads than the right. With all of the above in mind, our group felt the best course of action was to move forward with the original dataset (SRX6631047) trimmed at 150 bp due to a reasonable mean genome coverage (**64x)** and sensible quality score results, represented in **Figure 1C (left read) and D (right read)** and **Table 1.** Specifically, one can observe a marked improvement from before (**A, B**) and after (**C, D**) trimming by shortening the interquartile range (IQR) of the quality scores (Q), (IQR =length of the yellow bar) spanning mostly in the green (preferred, Q> 28) area, while both average Q (blue line) and median Q (red line) are both well above Q28 after the trimming. The GC content (53%) was very similar to the published reference genome value of 52%. All other quality control criteria present on the fastqc output were satisfactory.

*Trimmomatic (v0.38)*

Multiple variations of trimming were employed with the goal of improving the quality of reads while maintaining reasonable genome coverage. After initial quality control on raw data using FASTQC, we detected the increased number of low-quality reads, especially the base pairs (bp) in greater than position 160. Both the mean and median quality scores (Q) were reaching into the red zone indicating Q less than 20, especially in the right reads. Thus, we employed the trimmomatic command CROP to take off all the bp on greater than 160 position. After the trimming, another FASTQC was performed to check the improvement of the data after trimming. We repeated dozens of various trimming options, including trimming off the bp smaller than the threshold of a chosen size (MINLEN); we also thought of excluding the bp of lower Q from both ends, using LEADING and TRAILING code as well as MAXINFO, trying to make a meaningful compromise between quality (strictness) and genome coverage. As obtaining higher quality reads led to small genome coverage (for instance keeping only 160 pb reads and excluding bases at the end of the read that had lower than Q22, lead to only genome coverage 48x; **Table 1**), we advanced with 150 bp.



B.

A.

**Figure 1:** **Quality scores across all bases before (A, B) and after trimming to 150 bp (C, D).** Position in read (bp) is on horizontal axis and Q, the quality score on the vertical axis. Colored regions represent the preferred (Q>28, green), less preferred (20<Q<28 orange) and low quality (red) reads. Bars represent IQR, the whiskers 10th and the 90th percentile of all bases in the position category.

D.

C.

**Table 1: FASTQC results with or without trimming the lower quality sequences.**

Five basic commands in Trimmomatic were utilized. CROP:# trimmed away all bases that are higher in position than the specified number; MINLEN (minimum length) trimmed away the bases with position less than the specified number; LEADING cut bases off the start of the read if quality score is below threshold; TRAILING (trail) cut bases off the end of a read, if quality score is smaller than the specified threshold value; MAXINFO:*x*:*y* should balance read length and error rate to maximize the value of each read, the *x* is the target read length, followed by the *y* representing the “strictness”, used in calculation of error sensitivity of the read, with possible value between 0 and 1. A low value of this parameter (<0.2) favors longer reads, while a high value (>0.8) favors read correctness;

150<crop<150 = CROP:150 MINLEN:150.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Trimmomatic Parameters** | **Total Seq.** | **Seq. Length** | **Cov. Depth (X)** | **IQR of Q (yellow bars) reaching lower quality area (Q<20, red, 20<Q<28, orange)** |
| **Untrimmed** | 533485 | 251 | **124** | 19 bars in red |
| **150<crop<150** | 460036 | 150 | **64** | 7 bars in orange |
| **160<crop** | 533485 | 160 | **79** | 8 bars in orange, 1 in red |
| **160<crop<160** | 450092 | 160 | **67** | 8 bars in orange, 1 in red |
| **170<crop<170** | 439482 | 170 | **69** | 9 bars in orange, 3 in red |
| **160<crop<160, trail<22Q** | 320481 | 160 | **48** | 1 bar in orange |
| **160<crop<130, trail<33Q** | 455650 | 160 | **68** | 3 bars in orange |
| **170<crop<130, trail<30Q** | 469353 | 170 | **74** | 8 bars in orange |
| **180<crop<160, leading<10Q, trail<30Q** | 422975 | 180 | **71** | 9 bars in orange |
| **180<crop<150, leading<15Q, trailing<30Q** | 443354 | 180 | **74** | 9 bars in orange |
| **200<cropped<100, trailing<30Q** | 499678 | 200 | **93** | 12 bars in orange, 3 in red |
| **200<crop<150, leading<10Q, trailing<10Q** | 460036 | 200 | **85** | 8 bars in orange, 9 in red |
| **200<crop<150, leading<10Q, trailing<25Q** | 459608 | 200 | **85** | 8 bars in orange, 8 in red |
| **230<crop<230, trailing<30Q** | 108660 | 230 | **23** | 8 bars in orange, 1 in red |
| **maxinfo:170:0.8** | 533485 | 251 | **124** | 8 bars in orange, 3 in red |

**Processing-**

*Velvet Assembly**(v1.2.10)*

Velvet was utilized for the de novo genome assembly. The untrimmed (raw) data was assembled using Velvet and the k-mer size of 15. This resulted in a de Bruijn Graph (DBG) with 243,749 nodes and N50 of 28, which is unacceptable. Utilizing the *NG* dataset that was trimmed to 150 bp and excluded sequences less than 150 bp, the assembly using Velvet was performed. First, odd k-mers ranging from 15-31 were used to assemble the *NG* genome. This resulted in N50 values ranging from 12 (k-mer 15) to 94 (k-mer 31) (**Figure 2; inset**). Due to obtaining low N50 values, *NG* was assembled again with Velvet while utilizing larger k-mer sizes with odd values ranging from 75-93. The resulting N50 values ranged from 5.8 kb (k-mer 75) to 10.2 kb (k-mer 93) (**Figure 2)**.

**Post Processing-**

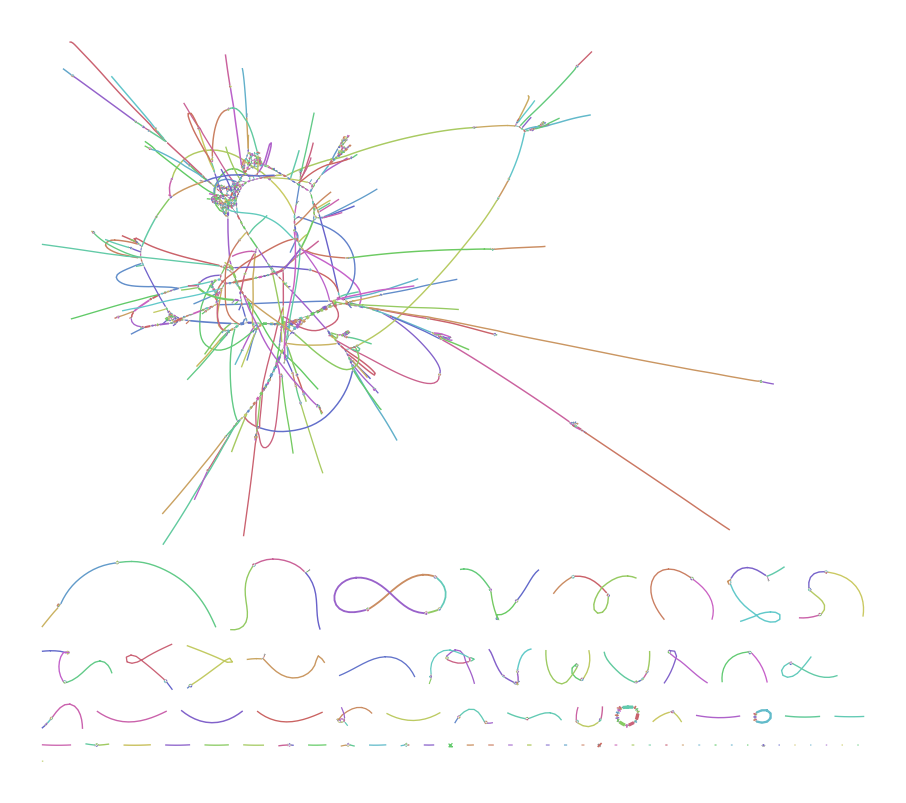
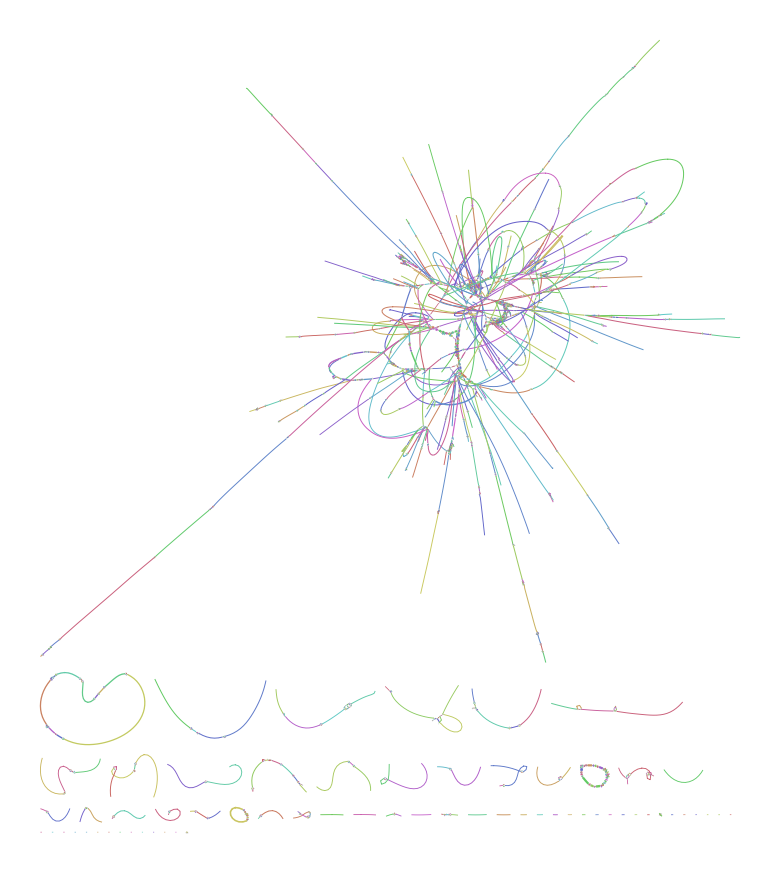
At the completion of the *NG* assembly, a N50 vs. k-mer graph was constructed in order to identify the optimal k-mer size (**Figure 2**). Our group decided a k-mer size of 81 resulted in the best DBG as N50= 9.9kb, median depth 22.4 x, dead end= 10.53%, longest node= 64,280, number of contigs= 1,325, total length= 2,158,307 as compared to other k-mer sizes observed. For example, k-mer size of 75 resulted in N50= 5.8 kb, dead end= 6.16%, longest node= 24,181, number of contigs= 2,063, and total length= 2,174,156. When comparing the above-mentioned k-mer sizes, a k-mer of 81 resulted in a larger N50 value, longest node, and the least number of contigs, which depicts a better assembly than k-mer size of 75. Bandage software was used to better visualize the DBG graphs (**Figure 3**). When comparing the *NG* reference genome length (2,153,922 bp; *NG* reference genome available on NCBI) to our assembled *NG* genome (2,158,307 bp), our assembled genome was 4,385 bp larger (100.2%) than the reference genome.



**Figure 2: The effect of k-mer length on N50 size.**

The k-mer lengths of 15 to 31 resulted in low N50 values which is depicted in the inset. Therefore, larger k-mer sizes of 75 to 93 resulted in higher N50 values. Increasing *k* from 75 to 81 increased N50 from 5.8 to 9.9 kb. However, the k-mer length of 83 decreased N50 to 9.7 kb. The N50 increased to 10.2 kb with k-mer length of 93.

1. **B.**



**Figure 3. DBG graphs obtained from utilizing Bandage software to visualize the effect of k-mer length and N50 values.** (A) The DBG graph using the k-mer length of 75 and (B) the DBG graph using k-mer length of 81.

**DISCUSSION**

The objective of this experiment was to assemble the genome of *NG* isolate and to compare the assembled genome size (bp) to the reference genome size (bp) found on NCBI. The pipeline implemented varies from the proposed pipeline due to new knowledge discovered in necessary software for the assembly process (**Figure 4**). The proposed pipeline included two additional steps; the use of Assemblathon script for the number of contigs and Centrifuge software was proposed to use in order to eliminate contamination with other bacterial species found within our sample. Assemblathon script deemed unnecessary due to Velvet output providing the number of contigs and Centrifuge was also unnecessary due to our isolate being 97.7% pure.

The process of quality control with the Fastqc program revealed variation within our quality scores across all bases. Nkob consistently ran into complications with the right read having too many bases in the less preferred low-quality area as compared to the left read. This maybe a consequence of Illumina's sequencing-by-synthesis method, essentially, the more cycles run, the noisier the signal or even a product of improper library construction. Trimommatic command line tool, utilizing the paired end mode, enabled us to trim the Illumina (FASTQ) data after we detected the presence of low-quality bases. The issue with lower quality of right reads could possibly be addressed by further optimizing trimming using stricter criteria. For instance, our right reads could be trimmed to a smaller length (~150 bp) than left reads (~170 bp) and then utilize another software to pair them before QC and assembly. However, we would need to make sure we would achieve appropriate genome coverage.

To generate the DBG graph, Velvet software was utilized. Initially, the untrimmed data were assembled using a N50 value of 15. This did not result in a high-quality assembly as the dataset included longer reads that were of poor quality at the 3’ end. After performing trimming, the *NG* genome was re-assembled. However, this time multiple k-mer sizes were utilized in a single command line through the discovery of the command “*m, M, s”.* Odd values of *k* in increments of 2, ranging from 15 to 31 were used for the initial alignment of the data trimmed to 150 bp. This resulted in reduced N50 values which ranged from 12 to 94. In an effort to raise the N50 values, larger values of *k* (odd values ranging from 75-93 in increments of 2) were used. This resulted in higher quality DBGs as N50 values ranged from 5855 to 10224. To determine the optimal value of *k* for our dataset, a N50 value vs. k-mer size plot was generated that depicts the increase in N50 with higher k-mer values to a certain point. We decided the *k* value of 81 resulted in the highest quality DBG graph as the N50 was 9.9 kb and contained 1325 nodes. The *k* value of 93 resulted in a DBG graph with fewer nodes and a higher N50 value, however, it had a higher number of dead ends and a lower median coverage compared to graph generated with the *k* value of 81. It was unexpected that the highest N50 value did not generate the highest quality DBG. After investigation, it was clear that after the N50 peaked at 9.9 kb (*k =* 81), the DBG graph began to decline in quality despite the N50 values once again reaching 10 kb (*k* = 93).

Overall, the final N50 value of 9.9 kb could be improved in the future through further modifying our methods including optimizing trimming of the paired end data and determining the effect of k-mer sizes larger than 93. In addition, we learned the importance of running the assembly earlier in the process, the importance of optimizing k-mer length to obtain larger N50 values, and the value of dividing labor amongst team members to successfully and thoroughly complete the task.

# PIPELINE CODE

# 

**Figure 4. Software pipeline utilized to obtain the whole genome sequence of *NG.***

*Fastqc Code (Untrimmed Data)*

fastqc FastA.gz FastB.gz

#fastqc was utilized to address the quality of the raw paired end dataset selected

*Trimmomatic Code (including fastqc of output)*

trimmomatic PE -threads 6 FastqA.gz FastqB.gz -baseout ng\_cro150.fastq CROP:150

#the paired end reads were cropped to 150 bp by trimming the 3’ end of the reads

fastqc ng\_cro150\_1P.fastq ng\_cro150\_2P.fastq

#the quality was assessed again by using fastqc

trimmomatic PE -threads 6 ng\_cro150\_1P.fastq ng\_cro150\_2P.fastq -baseout ng\_cro150min150.fastq MINLEN:150

#reads less than 150 bp in length were then excluded from the dataset

fastqc ng\_cro150min150\_1P.fastq ng\_cro150min150\_2P.fastq

#fastqc was ran again to ensure that the trimmomatic parameters set improved the quality of the data set

*Velvet Code*

velveth Trimmed150Longkmer 75,95,2 -shortPaired -fastq -separate ng\_cro150min150\_1P.fastq ng\_cro150min150\_2P.fastq

#velveth was ran on the trimmed/cropped data to create the files for velvetg to interpret using the k-mer lengths of odd numbers ranging from 75 to 95

velvetg Trimmed150Longkmer -cov\_cutoff 4 -min\_contig\_lgth 100 -ins\_length 150 -exp\_cov 64

#velvetg was used to create the DBG. The coverage cutoff was set at 4 to exclude low coverage nodes. Minimum contig length was designated to be 100 bp to eliminate short contigs and the insert length was set to 150 since reads were trimmed to 150. The expected coverage was designated to be 64 as this was our expected coverage of the trimmed dataset.

**REFERENCES**

1. L. Newman, J. Rowley, S. V. Hoorn, N. S. Wijesooriya, M. Unemo, N. Low, G. Stevens, S. Gottlieb, J. Kiarie, M. Temmerman, Global Estimates of the Prevalence and Incidence of Four Curable Sexually Transmitted Infections in 2012 Based on Systematic Review and Global Reporting. *PLOS ONE*. **10**, e0143304 (2015).

2. J. W. Tapsall, F. Ndowa, D. A. Lewis, M. Unemo, Meeting the public health challenge of multidrug- and extensively drug-resistant Neisseria gonorrhoeae. *Expert Review of Anti-Infective Therapy; London*. **7**, 821–34 (2009).

3. R. C. Deonier, M. S. Waterman, S. Tavaré, Eds., in *Computational Genome Analysis: An Introduction* (Springer New York, New York, NY, 2005; https://doi.org/10.1007/0-387-28807-4\_4), pp. 99–119.

4. R. S. Lee, T. Seemann, H. Heffernan, J. C. Kwong, A. Gonçalves da Silva, G. P. Carter, R. Woodhouse, K. H. Dyet, D. M. Bulach, T. P. Stinear, B. P. Howden, D. A. Williamson, Genomic epidemiology and antimicrobial resistance of Neisseria gonorrhoeae in New Zealand. *Journal of Antimicrobial Chemotherapy*. **73**, 353–364 (2018).