UPPSALA
UNIVERSITET

Factors Affecting How Well Bacterial Whole Genome Sequencing Reads Assemble

Linda Mustafa

Master's Degree Project in Medical Research, 30 credits, Spring 2021
Department: Medical Biochemistry and Microbiology
Supervisor: Bo Segerman

Table of Contents

Abbreviations

AMR
Antimicrobial Resistance

BAM
Binary Alignment Map

BWA
Burrows-Wheeler Aligner

C,jejuni
Campylobacter Jejuni

cgMLST
Core Genome Multilocus Genotyping

DNA
Deoxyribonucleic Acid

NGS
Next-Generation Sequencing

N50
The shortest contig length that needs to be included for covering 50% of the genome

PCR
Polymerase Chain Reaction

PFGE
Pulsed-Field Gel Electrophoresis

QC
Quality Control tool

SAM
Sequence Alignment/Map format

SNP
Single-Nucleotide Polymorphism

SMRT
Single-Molecule Real-Time

SPAdes
St. Petersburg genome assembler

WGS
Whole Genome Sequencing

Abstract

Recently Whole Genome Sequencing (WGS) has become the new high-resolution tool used to trace the source of foodborne outbreaks. There are often only a few genetic differences that can distinguish closely related bacterial isolates, and variability in data quality between different laboratories may influence the results. In this project, a data set from ten laboratories where the same bacterial samples were sequenced using different library preparation kits and sequencing methods in an interlaboratory study, has been used. Factors that could be responsible for the different performance in terms of how well the raw WGS data from the different labs assembles were investigated. The raw data from the different labs assembled very differently. One lab showed adapter sequences in their reads and filtering them improved the assembly substantially. All labs utilizing the transposase-based library preparation kit Nextera, had base composition bias in the beginning of the reads. For many labs, as the coverage was increased, the number of contigs first increased and then decreased. This was due to low number of contaminating reads from other species. However, these contaminations were barely visible in the plots generated by Kraken/Krona. Filtering out contigs with very low coverage removed this problem. Two labs performed much worse than the others. Some of their reads showed quality drop towards the ends, whereas their data also had the longest read length. However, quality trimming the read ends did not improve the assembly. These two labs had higher GC content in their reads compared to the other labs, the reason for this needs further investigation.

Keywords

WGS, NGS, Illumina, *C,jejuni*, SPAdes, FastQC, MultiQC, BWA, Kraken/Krona, Trimmomatic.

Popular scientific summary

We all need food to survive. Nevertheless, food can effectively be contaminated with foodborne risks during processing, production, distribution, shipping, and preparation. According to the World Health Organization, about 600 million people get sick and 420,000 die, due to eating contaminated food each year. Viruses, bacteria, and parasites are the most common causes of foodborne diseases. Whole Genome Sequencing (WGS) is a novel, state of-the-art, accurate and rapid technology that is molecule-based and can be utilized globally to identify and distinguish pathogenic bacteria that may cause foodborne infection outbreaks. *Campylobacter* bacteria species are a major cause of foodborne infection derived from bacteria around the world and *C.jejuni* is one of the most common *Campylobacter* species.

In order to develop these WGS-based identification and distinguishing strategies, it is important to investigate the quality of the resulting WGS assemblies and the underlying mechanisms of quality variation in great details. Here it is suggested that variation in the quality of the data may influence comparisons combining data from different labs. In this study, a set of WGS sequences produced in an interlaboratory comparison between different labs that have sequenced the same foodborne pathogen isolates to identify factors that could be responsible for the different performance, was examined. These labs used different instruments and different library preparation kits (NEBNext, Nextera and TruSeq library prep kits).

The results show that raw data from the different labs assembled very differently despite originating from the same sample. This was due to different experimental protocols in individual laboratories or to contamination with other species. These contaminations were hardly visible in the plots generated by common bioinformatics tools. But instead, we could detect the contamination by filtering out low coverage contigs (contigs refers to “a set of overlapping DNA segments that together represent a continuous sequence”). While comparing the quality of each library kit, we found that the best performing NEBNext using lab data gives better assembly than the best Nextera using lab data. There was also a difference in results using different variants of the same library preparation kit.

The Nextera library preparation kit with the variant XT showed worse assembly results. The labs that used NEBNext produced better assembly results with the kit variant FS.

One lab was not able to produce identical replicates, but instead sequenced the same sample on three different Illumina machines: iSeq, MiSeq and NextSeq. This provided data for a comparison between these three machines. The per base sequence quality showed high quality along the whole reads for iSeq and MiSeq, while the NextSeq data showed a slight drop in sequence quality towards the ends of reads. Despite this quality drop, the data produced by NextSeq gave fewer contigs and better assembly than the others for both strains Campy1 and Campy2. The reason for this is unknown, and somewhat surprising, since the per base quality was worst in the NextSeq data as compared to the others.

In summary this study shows that the final assembly quality can be affected by a large number of technical factors, and it is important for each lab to learn how to control their data for these factors.

Introduction

Foodborne diseases are known to cause significant human morbidity and mortality especially in young children and immunocompromised individuals (1). About 1.8 million children die worldwide everyday due to foodborne diseases (2). Viruses, bacteria, and parasites are the main causes of foodborne diseases (3). Recent technological advances have made it possible to create a safer process of producing foods by accurately identifying, monitoring and preventing microbial and chemical hazards in foods. Pathogenic bacteria isolated from food can be compared to each other with high resolution genetic methods to trace spreading and to identify sources of food contaminations. Linking the genetic data with data from patient isolates can connect outbreaks to food sources. *Campylobacter* bacteria species are a major cause of foodborne illness derived from bacteria worldwide and *C.jejuni* is one of the most common C*ampylobacter* species (4,5).

Older genetic typing methods, e.g., pulsed-field gel electrophoresis (PFGE), give a medium resolution DNA fingerprint that can be used to compare bacterial isolates, and can thus be used to investigate and control outbreaks. However, it has limitations related to the resolution when used in molecular characterization and subclassification of bacterial pathogens and delivers substandard precision. (6). Recent developments in our ability to sequence an entire bacterial genome in a timely and cost-efficient manner have given us a new high-resolution tool that is providing information about a pathogen at the DNA and gene level with previously unseen precision (7,8). The process for doing this is called whole-genome sequencing (WGS). Nowadays high-resolution bacterial typing can be made routinely with WGS, which has replaced many traditional molecular approaches such as PFGE and serotyping. Also, antimicrobial resistance (AMR) and virulence profiling can be performed with WGS data. (9,10). Thus, WGS is ideal for using in international and national surveillance systems for the public health and food safety. In addition to what WGS has contributed with in terms of detection and response to outbreaks, there will likely be a revolution in microbiological source attribution of sporadic foodborne diseases and an expansion of our knowledge of the epidemiology of various infectious diseases in the coming years. WGS technology decodes almost all of the information in an entire bacterial genome, in a timely and cost-effective manner. It is a tool with a high ability to deduce sources of contamination in addition to contributing to the ability of food safety officials to detect, respond to, and even prevent outbreaks of foodborne disease much more accurately and quickly compared to any other method used so far (11).

The WGS technology has developed from Sanger sequencing (12), which was a good method for very precise sequencing of relatively short DNA fragments (about 1000 base pairs). The consequent evolution of next-generation sequencing (NGS) made it possible to sequence whole bacterial genomes rapidly and in a short time and WGS could therefore be applied to public health and food safety (13,14). High throughput bacterial WGS is mainly done with short read sequencing, Illumina being the most commonly used such technology. Illumina sequencing is the most used short read NGS technology and is dependent on adding reversible terminators to identify the single bases as they are integrated into DNA strand. First the DNA molecules are attached to short adaptor sequences and then the DNA molecules are amplified on a solid surface which allows the formation of local clusters. The reversible terminating nucleotides (A, T, C, G) are added and those nucleotides are fluorescently labelled and connected to a blocking group. The competition for binding site between these four nucleotides occurs on the DNA template and nonincorporated molecules are washed away. The fluorescent colour of each newly incorporated nucleotide is then detected by a camera. Finally, a laser removes the fluorescent label and the blocking group. Repetition of these steps allows for identification of each base in the sequence. The process is repeated in cycles. The advantage of Illumina is that it enables to sequence millions of DNA fragments at one go and the sequences are obtained rapidly (15). The rapid development of NGS sequencing technologies has led to emergence of many new sequencing platforms (16). Ion Torrent/Ion proton is

another platform that can be used for bacterial WGS. The system depends on the use of standard sequence chemistry, but with a detection system based on semiconductors. This sequencing method is dependent on the detection of hydrogen ions that are released during DNA polymerization, in contrast to the methods that are used in other sequencing technologies (17, 18). Ion Torrent/Ion proton suffer from inaccuracy in identifying the length of homopolymers, i.e., repeats of the same nucleotide (19,20). Thus, the homopolymeric errors lead to local alignments that are inaccurate and become an issue when analysing genomic variations (21).

Recently long read sequencing has become popular. It is especially good for creating reference genomes which are completed or close to completed. Long read sequencing can produce complete or near complete genomes of bacteria. The technologies can generate sequences up to 30 kilobase pairs originating from a single DNA molecule. The quality of each individual base in the reads is relatively low, so the technology relies on making a consensus of many reads and is therefore often combined with Illumina sequencing to obtain high base quality. The longer read length obtained can resolve repetitive regions that are difficult to assemble with short read data and therefore long read sequencing leads to better assembly of the sequences (22). There are two main technologies available for long-read sequencing: PacBio single-molecule real-time (SMRT) sequencing and Oxford Nanopore sequencing. These two techniques apply different methodologies; however, both are able to sequence long stretches of DNA (23).

WGS analysis can be divided into a wet-lab part where DNA is prepared and sequenced and a bioinformatic part where data is analysed. Bioinformatics analysis consists of extensive computational steps for the appropriate analysis of the massive amounts of data generated from NGS technologies. Bioinformatics, respectively in the context of molecular pathology and genomics, uses statistical and computational mathematical tools to organize, collect, and analyse the large and complex data of genome sequencing (24). Bioinformatics analyses are performed in multiple steps where the output of the first analysis serves as input for the next step of analysis and so on. These multi-software /multi-step analyses are commonly referred to as pipelines (25). Almost all bioinformatics analyses of WGS check the quality of the raw data as the first step. FastQC is a quality control tool that is frequently used to assess the quality of the sequencing run and to identify problems in the raw data such as inclusion of adapter sequences and poor quality of data. If such problems exist, Trimmomatic is a bioinformatic tool for read trimming that can be used to remove low quality sequences from the end or the beginning of a sequence and excise adapter sequences (26).

After trimming, the next step is usually to perform genome assembly using a de novo genome assembly program. SPAdes assembly (short for “St.Petersburg genome assembler”) (27), is one of the most frequently used assembly programs for bacterial genomes. SPAdes has three modules, the first one using higher algorithms that depend on Bayesian sub-clustering and Hamming graphs to correct errors of the quality-controlled sequences. This leads to improved quality of input sequence reads and thereby also improved genome assembly of NGS dataset (28). The second module is the assembly module that uses corrector-error reads and then implements the assembly in an iterative aspect based on Bruijn graphs. The third module is a mismatch corrector which uses the output from the second module and reduces significant numbers of mismatches by mapping the reads back to the assembly using the Burrows-Wheeler Aligner (BWA), eventually creating very accurate contigs (29).

Mapping the WGS reads against a reference genome sequence is an important step in many WGS assembly pipelines. One of the software packages that performs mapping is BWA (30). BWA aligns sequence reads against a reference genome sequence using a Burrows-Wheeler Transform that allows for mismatches and gaps. BWA uses SAM (Sequence Alignment/Map) format as standard output for sequence alignments.

Contaminations can disturb WGS analysis and sometimes the taxonomic classification tool Kraken is used to identify possible contaminations (31). The output from Kraken can be visualized in bioinformatics

visualization tools such as Krona. Krona uses JavaScripts and HTML implementation that allows for interactive charts that can be observed with any web browser, and no need to install any software or further plug ins. This web-based architecture also allows each layout to be a separate document, making it easy to share via email or publish to a standard web server (32). The WGS technology gives high resolution data, but variation in the quality of the data may influence comparisons combining data from different labs.

Aim

In this project we have used a set of WGS sequences produced in an interlaboratory comparison between different labs that have sequenced the same foodborne pathogen isolates (two *Campylobacter jejuni* isolates). We have used the raw data from theses labs to compare different quality parameters. The goal was to identify factors that could be responsible for the different performance in terms of quality of the resulting WGS assemblies when using data from the different labs. To measure how well the raw data assembles, we have used an approach where the amount of raw data is titrated, thereby creating curves that show how assembly QC parameters, such as number of contigs and N50 values, change when coverage is increased.

Materials

The data set was retrieved from *German-Wide Interlaboratory Study* where the reproducibility and accuracy of NGS (next-generation sequencing short read) data was compared between ten laboratories participating in food safety analysis (state laboratories, research institutes, universities and companies) from Austria and Germany (33). DNA samples of three different bacterial species were sequenced (*Campylobacter jejuni*, *Salmonella enterica*, and *Listeria monocytogenes)* in duplicates, according to local sequencing protocols. The participating laboratories had used different sequencing platforms of Illumina (iSeq, MiSeq, NovaSeq, NextSeq) and one instrument (S5) of Ion Torrent sequencing technology. For all data sets sequence quality parameters were determined and thereafter compared between these ten laboratories. Core Genome Multilocus Genotyping (cgMLST) and SNP typing were conducted to evaluate the reproducibility of sequence data collected for individual samples. In this project we took the *Campylobacter jejuni* data and compared the results from different laboratories using bioinformatics tools. The nine labs that had used Illumina to perform the WGS were included in the analysis, whereas the one lab that used Ion Torrent was excluded. Each lab sequenced two *Campylobacter* strains and used different instrument/library preparation kits. The reference sequences had been produced in the same interlaboratory study by PacBio sequencing and are available from the Bioproject PRJNA638266 (33). The Illumina data used in this study is available from the Bioproject PRJEB37768 (34).

Methods

The raw fastq-files were analyzed with FastQC (version 0.11.9) (35) and the FastQC-reports were summarized with multiqc (36). The fastq-files from labs having adapter contamination or poor quality (identified by FastQC) were trimmed with Trimmomatic (37) using standard settings. A Perl script was used that downsampled the fastq-files by truncating them so that they corresponded to different coverages (10X, 20X, 30X etc.). (Coverage = how many times the raw data covers the genome on average = X). The coverage was calculated by dividing the number of bases in the fastq-file by the size of the corresponding reference genome. Assembly was generated with SPAdes (version 3.14.1) using the "-careful" option (38). Contigs with low coverage were filtered out with a Perl script that extracted the coverage of each SPAdes contig (which is part of the contig name) and dropped the contig if its coverage was less than five percent of the expected coverage. We used the taxonomic classification tool Kraken2 to identify possible contaminations (39). Kraken was run with standard settings using the database Minikraken2_v1_8GB which included refseq bacteria, refseq archaea and refseq viral sequences from April 2019. The Kraken data was visualized in Krona (40). Finally, we used BWA with standard settings (41) to align sequences of reads against reference sequencing genomes. The SAM files were converted to BAM (Binary Alignment Map) files and statistics were extracted with samtools (42). This allowed for quantification of unmapped reads and determination of the fragment size in the paired end sequencing. The number of contigs and the N50 values were counted by a custom Perl script.

Results

Overview of the labs

Sequencing data from the nine labs sequencing with Illumina technology was included in this study, whereas data from one lab (L01), using Ion Torrent sequencing, was excluded from the comparison. Each lab sequenced two *Campylobacter* strains in two independent replicates. One lab sequenced three replicates but on different instruments (L02) and one lab produced no replicates (L08). The instrument type, library preparation kit and amount of generated data is summarized in Table 1. The datasets include four different types of Illumina instruments (iSeq, MiSeq, NextSeq and NovaSeq) and six different library preparation kits (three variants of NEBNext, two variants of both Nextera and TruSeq). The coverage varied between 28 and 1080X. More information about the QC checks of the labs is found in Appendices.

Table1. *Overview of the technical implementations at the labs and the amount of data generated.*

Sample	**Instrument**	**Library kit**	**Campy1**		**Campy2**	
			Amount of data (Mb)	**Assembly coverage (X)**	**Amount of data (Mb)**	**Assembly coverage (X)**
L02A	iSeq100	Nextera DNA flex	106	66	164	96
L02B	MiSeq	Nextera DNA flex	80	50	113	66
L02C	NextSeq 500	Nextera DNA flex	419	259	766	446
L03A	iSeq100	Nextera DNA flex	106	66	48	28
L03B	iSeq100	Nextera DNA flex	63	39	142	83
L04A	MiSeq	Nextera DNA flex	156	96	223	130
L04B	MiSeq	Nextera DNA flex	232	143	162	95
L05A	MiSeq	NEBNext Ultra II DNA, NEBNext Multiplex Oligos	347	214	301	175
L05B	MiSeq	NEBNext Ultra II DNA, NEBNext Multiplex Oligos	319	197	287	167
L06A	MiSeq	Nextera XT DNA, Nextera XT Index Kit	176	109	147	86
L06B	MiSeq	Nextera XT DNA, Nextera XT Index Kit	161	100	194	113
L07A	MiSeq	TruSeq Nano DNA, TruSeq DNA Single Indexes Set A, and Set B	904	558	808	470
L07B	MiSeq	TruSeq Nano DNA, TruSeq DNA Single Indexes Set A, and Set B	567	350	858	499
L08A	MiSeq	Nextera XT DNA	182	113	468	273
L09A	NextSeq 500	NEBNext Ultra II FS DNA	350	216	285	166
L09B	NextSeq 500	NEBNext Ultra II FS DNA	330	204	376	219
L10A	NovaSeq 6000	NEBNext Ultra II FS DNA	1750	1080	1730	1006
L10B	NovaSeq 6000	NEBNext Ultra II FS DNA	1950	1204	1770	1029

Overview of assembly results

In assembly using SPAdes we tested different coverages, in order to identify which coverage was optimal, and number of contigs at different coverages. We found that raw data from different labs assembled very differently, despite originating from the same DNA sample. Labs 6 ,8 and 9 (L06, L08 and L09, respectively) had the highest number of contigs compared to the other labs for both strains Campy1 and Campy2. Lab5 (L05) showed the best assembly result (Figure 1).

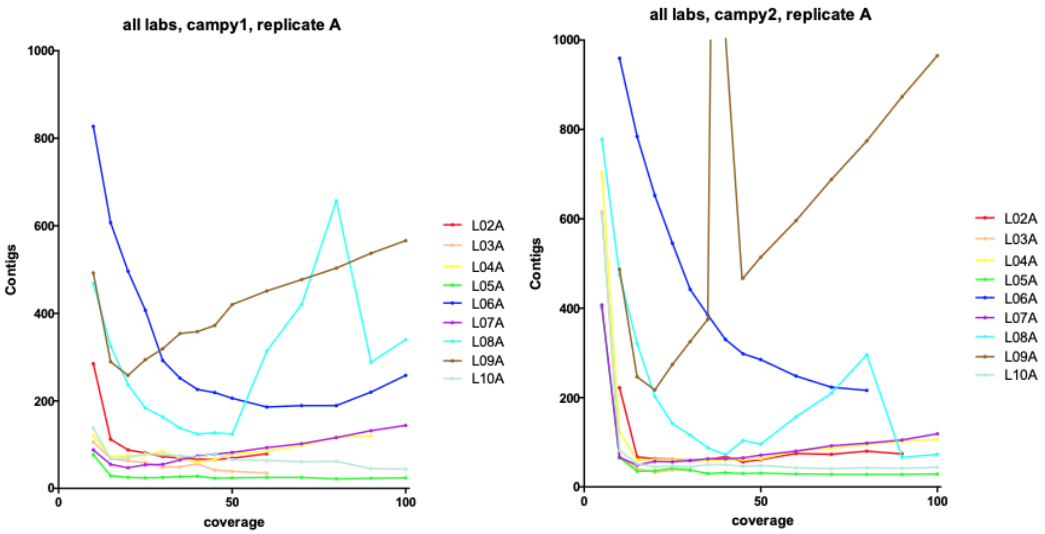


Figure 1. *Comparison of assembly results for all labs (strains Campy1 and Campy2) showed that sequencing data of the same isolate from different labs assembled very differently.*

Overview of FASTQC results

Errors and warnings in FASTQC

All FastQC-files were run through the FASTQC quality check tools. The error (“failed”) and warning flags are summarized in Figure 2 and Table 2. One lab (L06) had error flags in "Per tile sequence quality" in the R2 reads for the Campy2 strain. Eight labs (L02-L08 and one of the replicates of L09) had error flags in "Per base sequence content". Two labs (L02C, L10) had duplicate error flags and one lab (L09) had adaptor content error flags.

Table 2. *Overview of error flags of FASTQC.*

Lab ID	***Per tile sequence quality Campy1***	***Per tile sequence quality Campy2***	***Per base error Campy2***	***Duplicate error Campy1***	***Duplicate error Campy2***	***Adapter content Campy1***	***Adapter content Campy2***
L02A_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L02A_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L02B_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L02B_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L02C_1	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*
L02C_2	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*
L03A_1	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*PASSED*	*PASSED*	*PASSED*
L03A_2	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*PASSED*	*PASSED*	*PASSED*
L03B_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L03B_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L04A_1	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*WARNS*	*PASSED*	*PASSED*
L04A_2	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*WARNS*	*PASSED*	*PASSED*
L04B_1	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*WARNS*	*PASSED*	*PASSED*
L04B_2	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*WARNS*	*PASSED*	*PASSED*
L05A_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*

L05A_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L05B_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L05B_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L06A_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L06A_2	*PASSED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L06B_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L06B_2	*PASSED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L07A_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L07A_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L07B_1	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L07B_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*
L08A_1	*PASSED*	*PASSED*	*FAILED*	*WARNS*	*WARNS*	*PASSED*	*PASSED*
L08A_2	*PASSED*	*PASSED*	*FAILED*	*PASSED*	*WARNS*	*PASSED*	*PASSED*
L09A_1	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*
L09A_2	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*
L09B_1	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*
L09B_2	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*
L10A_1	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*WARNS*	*WARNS*
L10A_2	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*WARNS*	*WARNS*
L10B_1	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*
L10B_2	*PASSED*	*PASSED*	*PASSED*	*FAILED*	*FAILED*	*PASSED*	*PASSED*

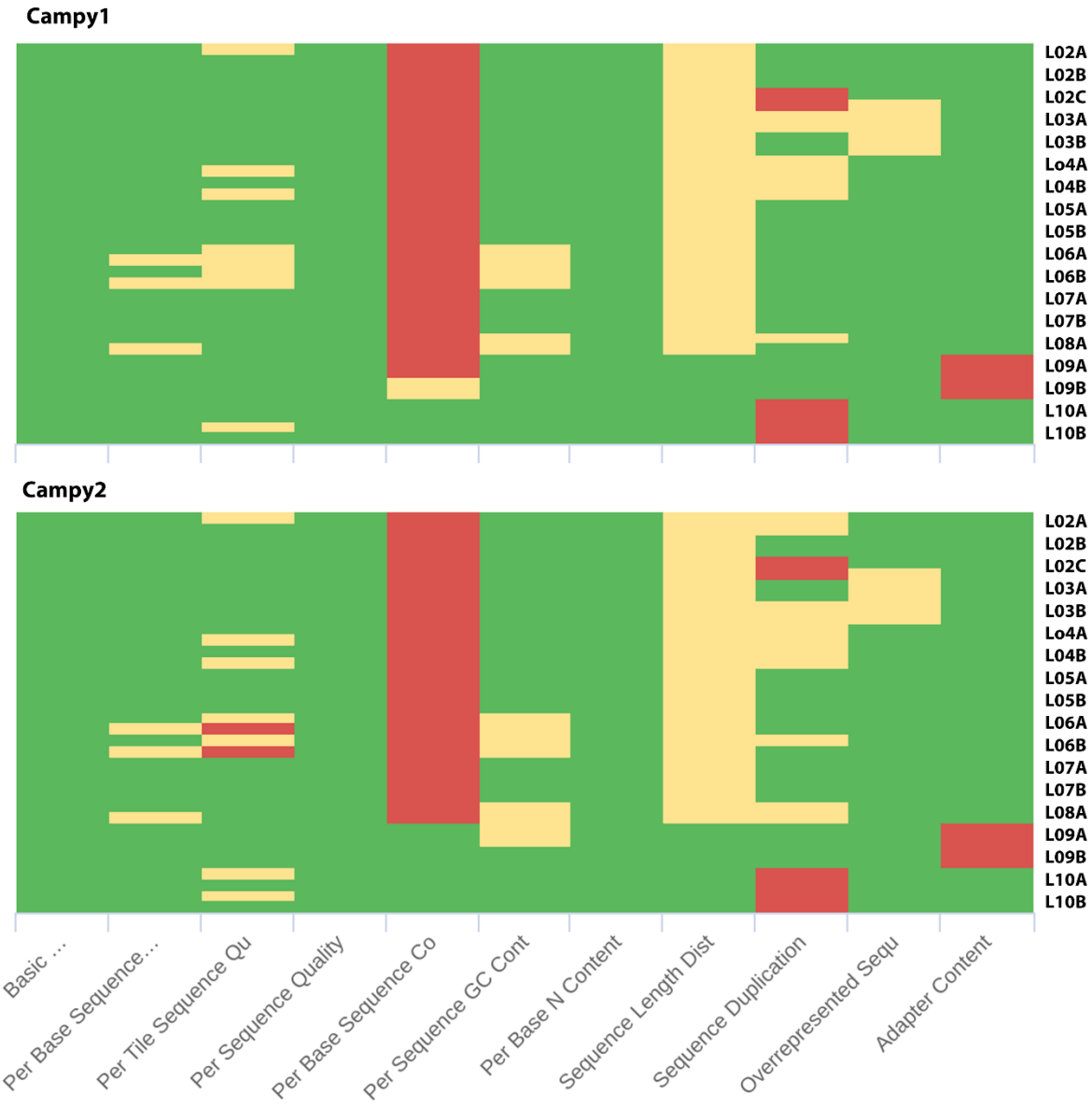


Figure 2. *An overview of the error flags (red) and warnings (yellow) generated by FASTQC analysis of the raw fastq-files. The heatmap was generated with MultiQC.*

FastQC error flags in a "Per tile sequence quality"

Lab6 (L06) had error flags in their reverse reads (R2 reads) for the Campy2 strain. This indicates that part of the sequencing lane has generated data of less quality. To evaluate if this disturbs the assembly process, the Campy1 (no tile error flag) and the Campy2 (tile error flag) data was compared. No major quality loss effect was seen for the tile error flagged data (Figure 3).

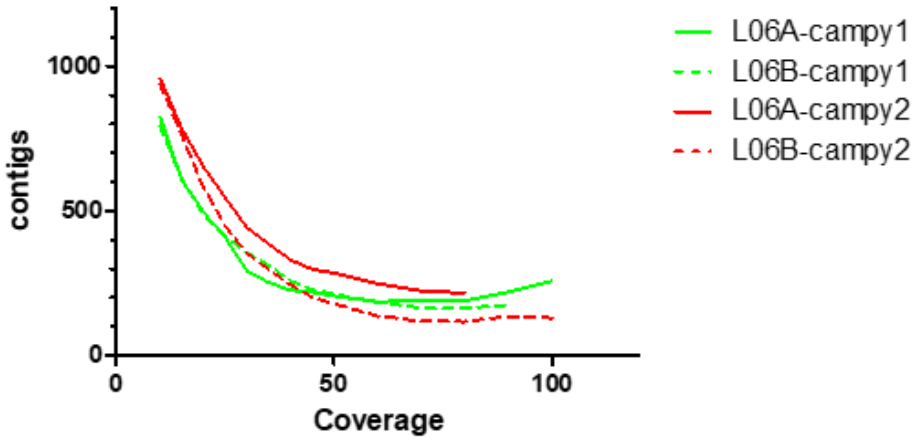
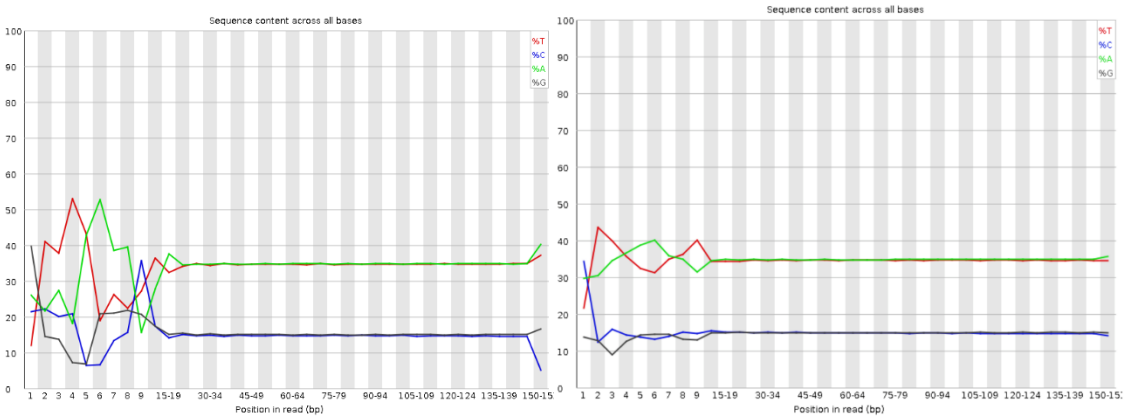


Figure 3. *Comparison of the number of contigs generated at different coverage depths for Lab6 (L06) sequencing of the Campy1 strain, which gave no error flag for "Per tile sequence quality", and the Campy2 strain, which gave this type of error flag.*

FastQC error flag in "Per base sequence content"

Most labs had a FastQC error flag in the "Per base sequence content" in some or all of their samples. Most had base composition deviation in the beginning of the reads, but sometimes also the very last base in the reads showed a deviating base composition. All the labs that had used Nextera library kit had a similar pattern of abnormal base composition in the first 9 bases of the reads in the "Per base sequence content" graph (Figure 3). The labs that used NEBNext library kit showed less abnormal base composition and the lab that used TruSeq showed almost no apparent base composition problem at all in the beginning of the reads. All labs that used TruSeq and Nextera showed a small base composition bias in their very last base (Figure 4). Two labs (L09 and L10) that used NEBNext library kit with the variant FS did not show the base composition bias in their very last base.

Nextera library kit (L02A) *NEBNext library kit (L09A)*



TruSeq library preparation kit (L07A).

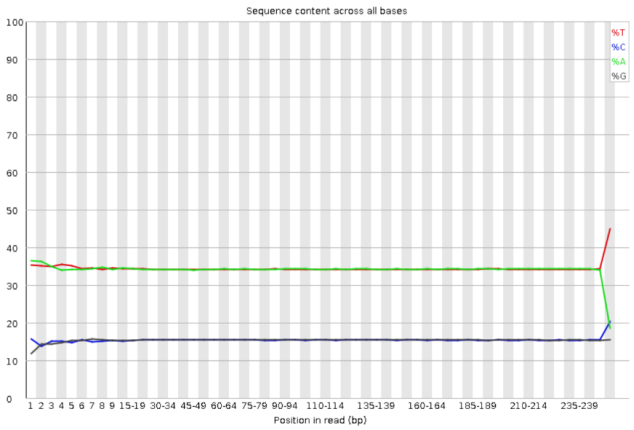


***Figure 4**. "Per base sequence content" plot for a lab using Nextera library preparation kit (L02A), NEBNext library preparation kit (L09A) and TruSeq library preparation kit (L07A).*

FastQC error flag in "Sequence duplicate levels"

Two labs failed "FastQC - Sequence duplicate levels" test for both strains Campy1 and Campy2, Lab2 (C replicate i.e. L02C)) and Lab10 (L10) (Figure 5). These were labs that had generated high amount of data (Table1). To test at which coverage these duplicate error flags started to appear for these two labs, their data were down sampled and re-analysed with FastQC until the error flags disappeared. For Lab2 (replicate C i.e. L02C)) we found that the error flag disappeared at 190X and was replaced by a warning flag. The warning flag disappeared at the coverage of 60X. For Lab10 (L10), the error flag disappeared at 700X and was replaced by a warning flag. The warning flag disappeared at the coverage 300X. The coverage that gave error flags could be influenced by how random the fragmentation in the library preparation kit is, and in comparison, of the two labs, L02 gave error flags at much lower coverage (Figure 6).

(Lab2 used Illumina NextSeq instrument and Nextera preparation kit)

(Lab10 used Illumina NovaSeq 6000 instrument and NEBNext preparation kit)

The strain Campy1 *The strain Campy2*

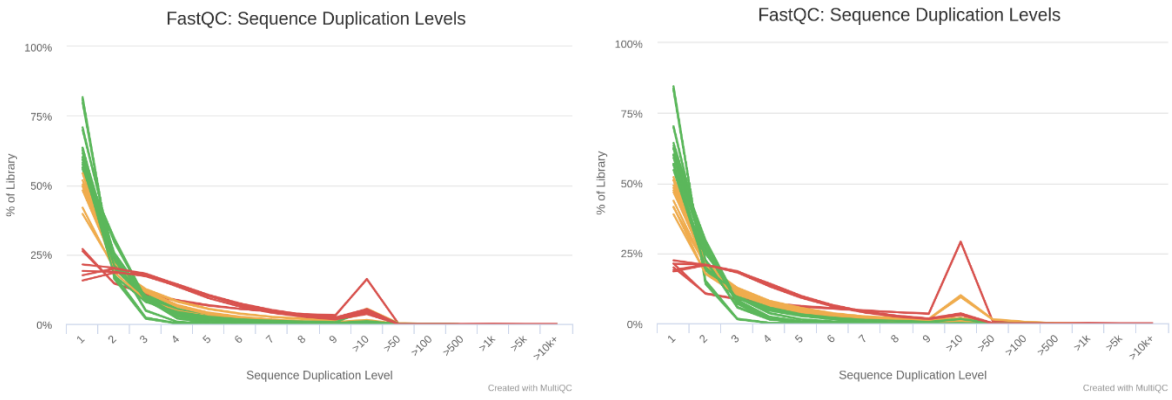


***Figure 5**. "Duplicate sequence level" for the strain Campy1 and Campy2, shows labs that have duplicate error marked in red (L02C, L10A, L10B).*

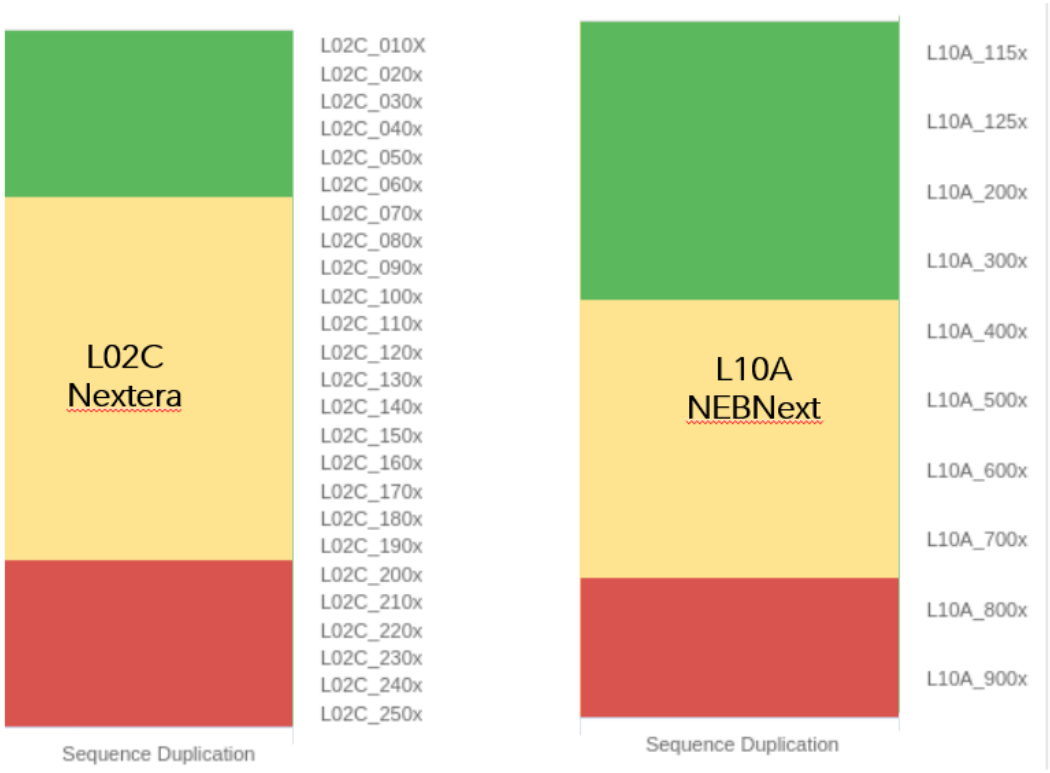


Figure 6. *Comparison of at which coverage Lab2 (using Nextera library kit) and Lab10 (using NEBNext library kit) gave error and warning flags for "Sequence duplicate level".*

FastQC error flag in the "Adapter sequence content":

Analysing of adapter sequence is a useful tool to determine whether one needs to trim for the adapter if it has a substantial amount of sequence or not. The plot itself exhibits a cumulative percentage count of the proportion of the library which had the adapter sequences at each position. When a sequence is displayed in the diagram, it is counted as present until the end of the reads, so the percentages you see will only increase as the read length increases. FastQC showed that 32 out of 36 passed and 4 out of 36 failed. Lab9 failed in both strain Campy1 and Campy2 (L09A_1, L09A_2, L09B_1, L09B_2). Lab10 has a little bit contamination of adapter content (Figure 7). However, Lab10 passed the adapter content analyses (Table 2). We removed the adapter for the Lab9 sequences using the Trimmomatic tool to improve the data for this lab. However, still Lab9 performed not so well after trimming the adapter indicating other remaining problems (Figure 8).

(Lab9 used Illumina NextSeq 500 instrument and NEBNext library preparation kit).

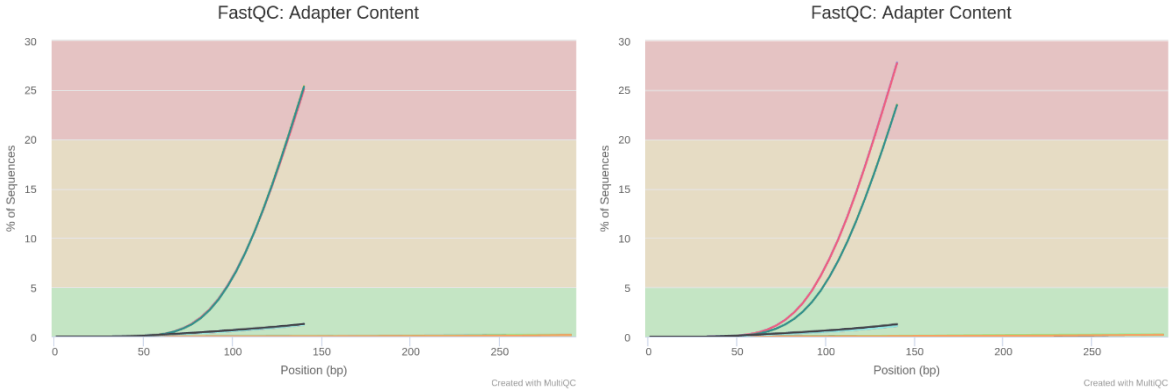


Figure 7. *The cumulative percentage count of the proportion of library which has seen each of the adapter sequences at each position for the strains Campy1 (left panel) and Campy2 (right panel). Lab9 (red and green) had significant presence of adapter sequence and Lab10 (Black) had also larger amount of adapter sequence but passed the adapter sequence analyses.*

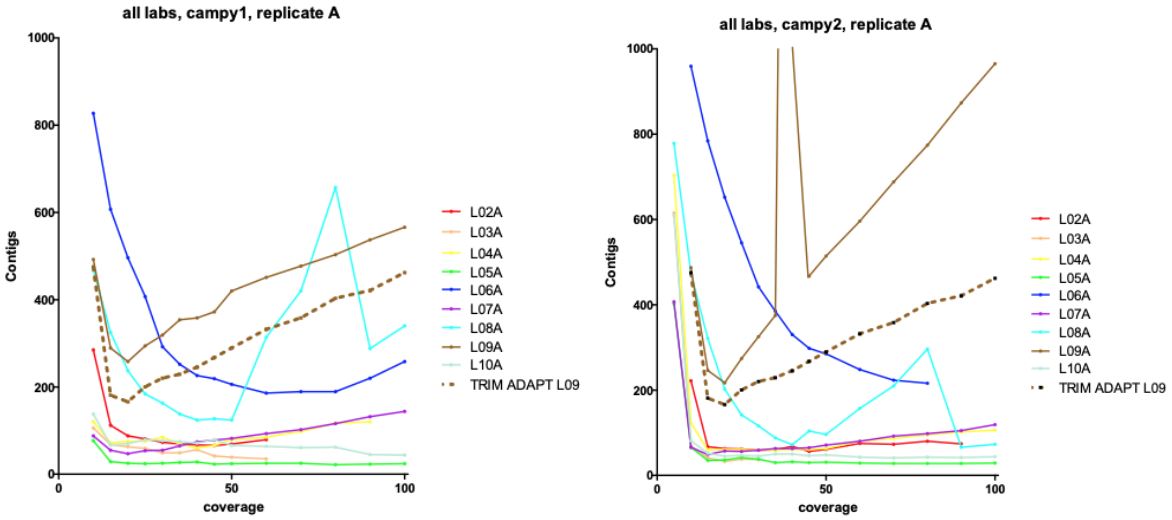


Figure 8. *Removing the adapter sequence content of the Lab9 for both strains Campy1 and Campy2. The bump at the coverage 40X for the strain Campy2 disappeared after adapter trimming.*

FastQC in "Per sequence quality score"

The "Per sequence quality score" shows if a subset of your sequences has low quality values. However, these represent a small proportion of the total sequences. The "Per sequence quality score" over all labs had high sequence quality for both strains Campy1 and Campy2 (Figure 9).

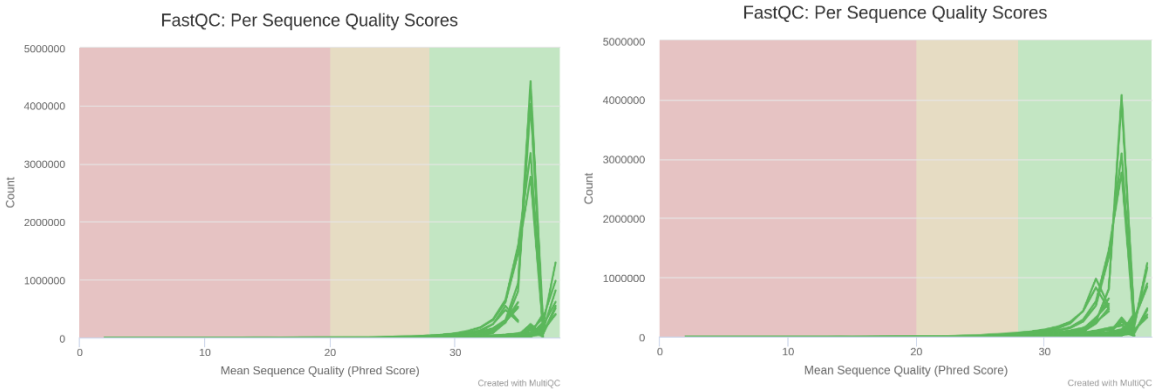


Figure 9. *"Per sequence quality score" for the strains Campy1 (left) and Campy2 (right). All labs had high score sequence quality.*

FastQC in "Per sequence GC content"

Two labs (Labs 6 and 8) had higher GC content than the other labs and these two labs used a different variant of Nextera library kit called Nextera XT whereas the other labs used Nextera DNA flex (Figure 10).

(Labs 6 and 8 used Illumina MiSeq instrument and Nextera XT library preparation kit).

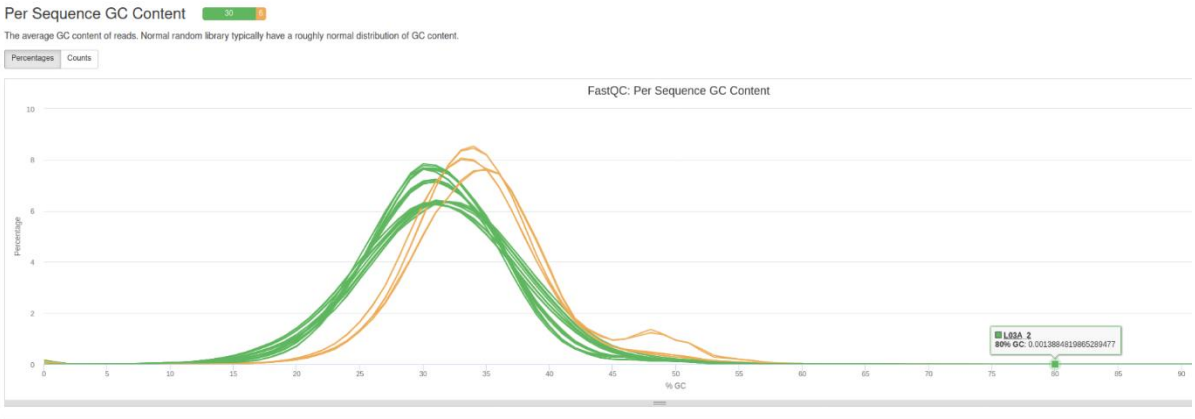


Figure 10. *The higher GC content for the Labs 6 and 8 compared to the other labs.*

Taxonomic classification by Kraken

Kraken is a tool for assigning taxonomic labels to DNA sequences. To perform this analysis, the percentage of each sample assigned to a particular classification is collected across all samples. Counts are then plotted with MultiQC for the top five classes for each of the nine different ranks. A visualization with Krona is shown in (Figure 11). The unclassified count is always displayed across all taxa ranks. According to Kraken almost all labs had very little contamination of other species and Lab5 had the highest correct species (*C. jejuni)* score (L05A 92,8%, L05B 92,6%) for the strain Campy1 and (L05A 90,4%, L05B 90,6%) for the strain Campy2 and Labs 2 and 6 (L02A 85,4% and L06A 85,9%) had the lowest score for the strain Campy1 and (L06B 82%) for the strain Campy2. In general, Lab6 had the lowest score and highest contamination of other species compared to other labs (Figures 12 &13 and Table 3). A small amount (approximately 1%) of *C.coli* was seen more in the strain Campy1 compared to the strain Campy2. Overall, no contamination problem was found (Figures 12 & 13).

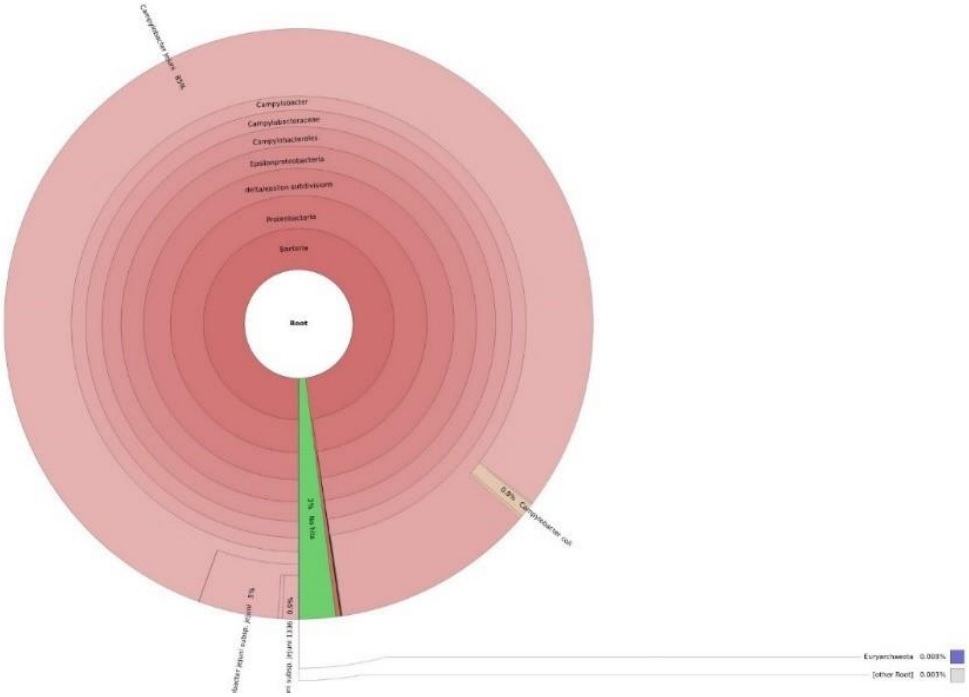


Figure 11. *An example of Krona plot of the L02A strain Campy1. No major contamination problem was observed.*

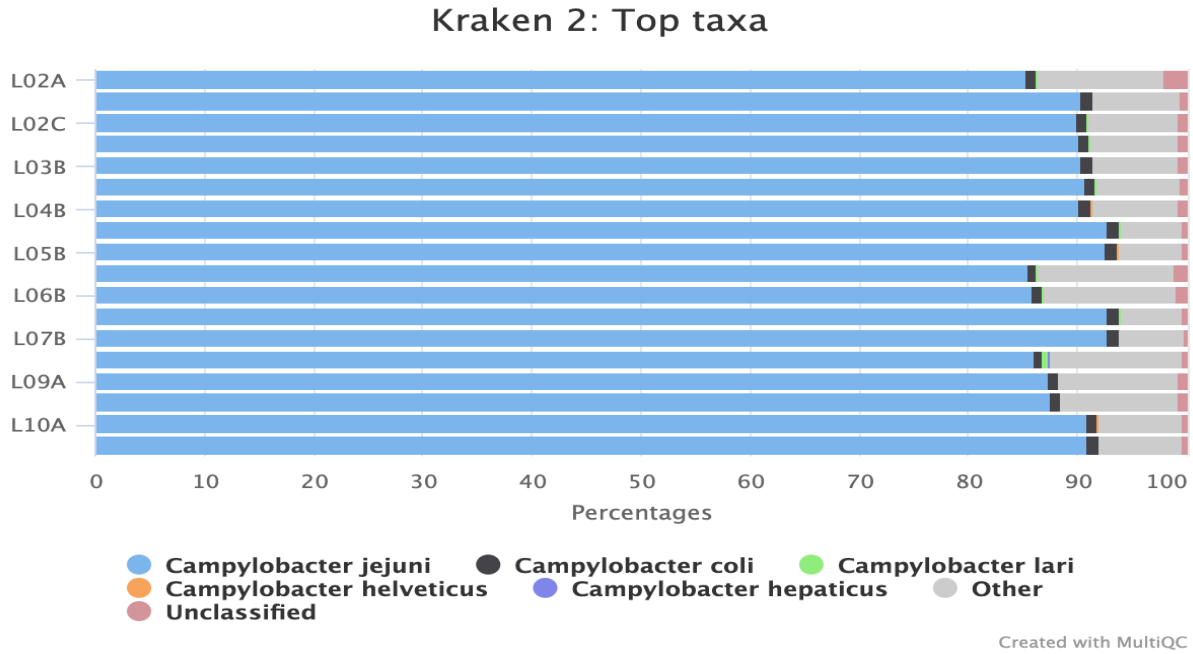


Figure 12. *Taxonomic classification by Kraken plot for the strain Campy1. No visible contamination over all labs was observed. A small amount of contamination with C.coli was detected.*

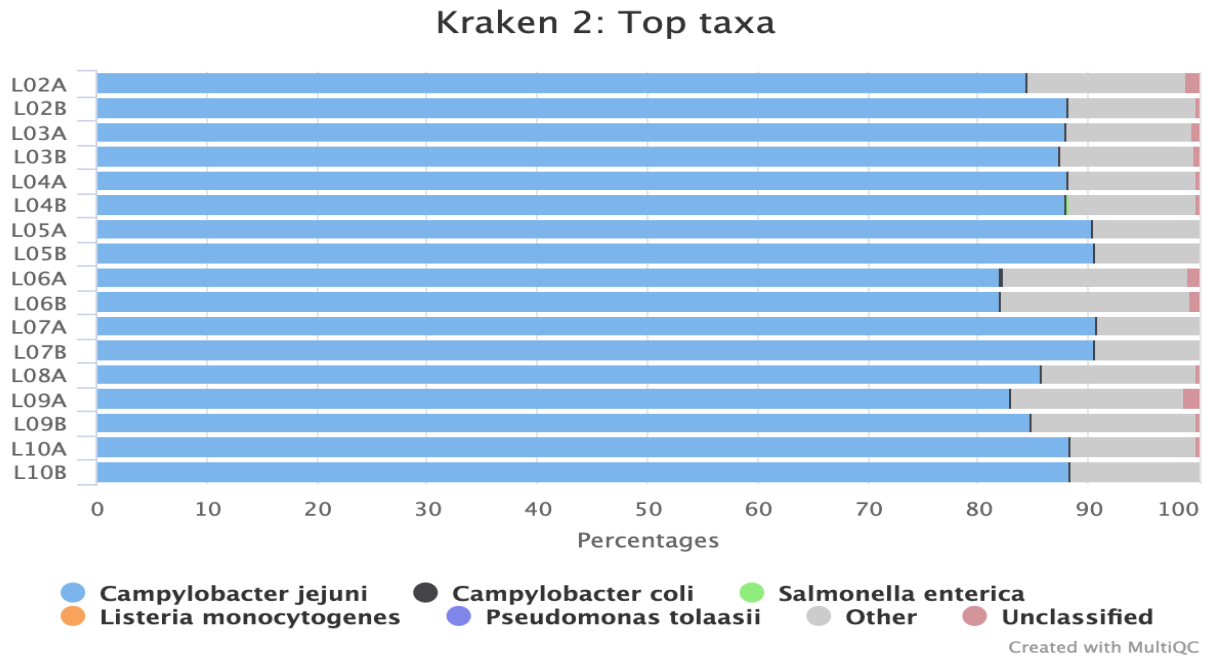


Figure 13. *Taxonomic classification by Kraken plot for the strain Campy2. Lower level of contamination with C.coli was detected compared to the strain Campy1.*

Table 3. *Percentage of correct species taxonomic classification by Kraken tool.*

Lab ID	***Kraken Campy1, Percentage (%)***	***Kraken Campy2, Percentage (%)***
L02A_1	*85.4%*	*84.5%*
L02A_2		
L02B_1	*90.3 %*	*88.1%*
L02B_2		
L02C_1	*89.9%*	*87.2%*
L02C_2		
L03A_1	*90.1%*	*88%*
L03A_2		
L03B_1	*90.4%*	*87.4%*
L03B_2		
L04A_1	*90.6%*	*88%*
L04A_2		
L04B_1	*90.2%*	*88%*
L04B_2		
L05A_1	*92.8%*	*90.4%*
L05A_2		
L05B_1	*92.6%*	*90.6%*
L05B_2		
L06A_1	*85.5%*	*82.1%*
L06A_2		
L06B_1	*85.9%*	*82%*
L06B_2		
L07A_1	*92.8%*	*90.6%*
L07A_2		
L07B_1	*92.8%*	*90.5%*
L07B_2		
L08A_1	*86%*	*85.6%*
L08A_2		

L09A_1	*87.4%*	*82.9%*
L09A_2		
L09B_1	*87.5%*	*84.8%*
LO9B_2		
L10A_1	*90.8%*	*88.3%*
L10A_2		
L10B_1	*91%*	*88.3%*
L10B_2		

Filtering low coverage

When we looked at short contigs in these assemblies they were often representing contaminating species (Figure 15) and therefore we composed a Perl script to filter out contigs with low coverage and thereby get better assembly results. Filtering indicated that the contamination of other species was the reason for the curves turning up again at higher coverages (Figure 14).

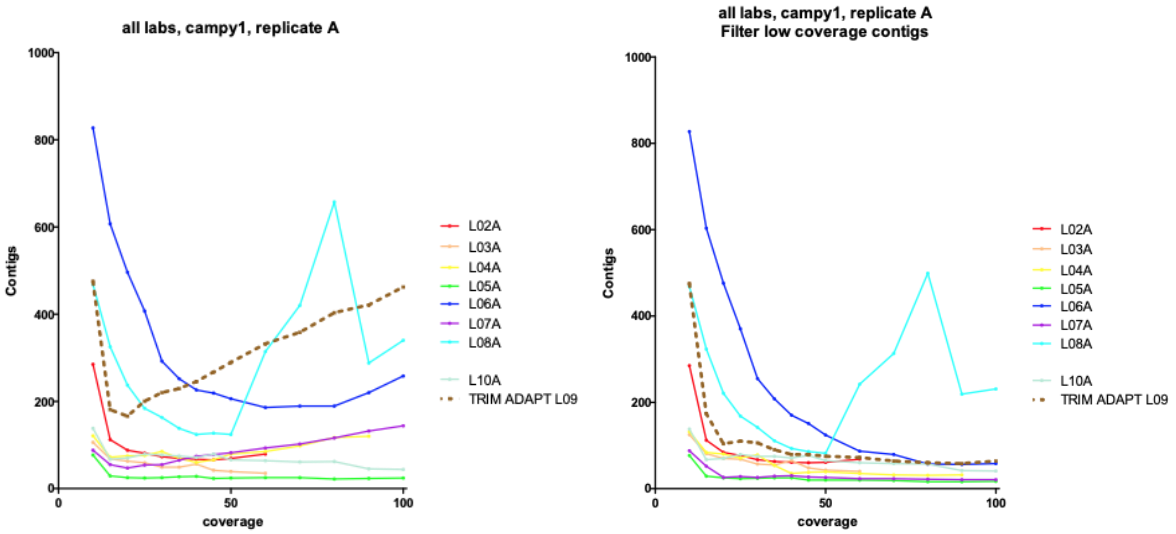


Figure 14. *Data filtering the low coverage contigs removed the problem that the initial decrease in the number of obtained contigs is followed by an increased number of contigs.*

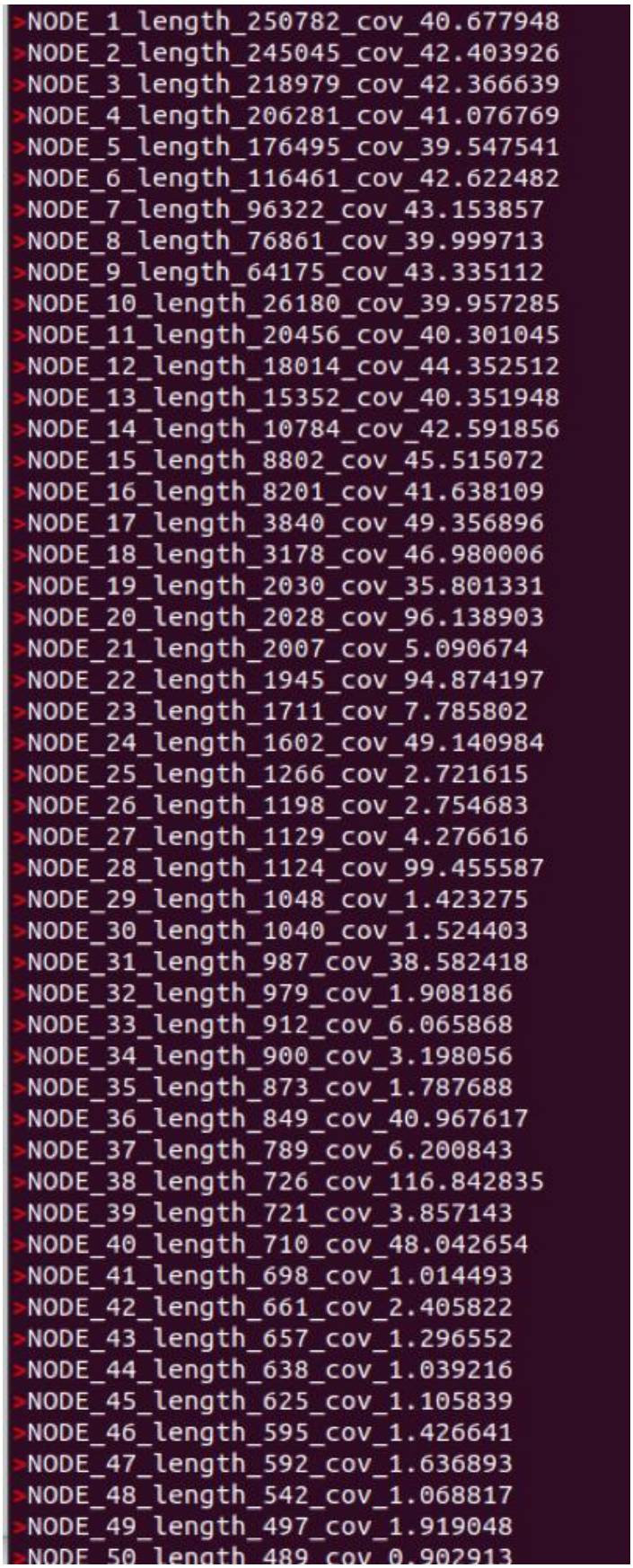

>NODE_1_length_250782_cov_40.677948
>NODE_2_length_245045_cov_42.403926
>NODE_3_length_218979_cov_42.366639
>NODE_4_length_206281_cov_41.076769
>NODE_5_length_176495_cov_39.547541
>NODE_6_length_116461_cov_42.622482
>NODE_7_length_96322_cov_43.153857
>NODE_8_length_76861_cov_39.999713
>NODE_9_length_64175_cov_43.335112
>NODE_10_length_26180_cov_39.957285
>NODE_11_length_20456_cov_40.301045
>NODE_12_length_18014_cov_44.352512
>NODE_13_length_15352_cov_40.351948
>NODE_14_length_10784_cov_42.591856
>NODE_15_length_8802_cov_45.515072
>NODE_16_length_8201_cov_41.638109
>NODE_17_length_3840_cov_49.356896
>NODE_18_length_3178_cov_46.980006
>NODE_19_length_2030_cov_35.801331
>NODE_20_length_2028_cov_96.138903
>NODE_21_length_2007_cov_5.090674
>NODE_22_length_1945_cov_94.874197
>NODE_23_length_1711_cov_7.785802
>NODE_24_length_1602_cov_49.140984
>NODE_25_length_1266_cov_2.721615
>NODE_26_length_1198_cov_2.754683
>NODE_27_length_1129_cov_4.276616
>NODE_28_length_1124_cov_99.455587
>NODE_29_length_1048_cov_1.423275
>NODE_30_length_1040_cov_1.524403
>NODE_31_length_987_cov_38.582418
>NODE_32_length_979_cov_1.908186
>NODE_33_length_912_cov_6.065868
>NODE_34_length_900_cov_3.198056
>NODE_35_length_873_cov_1.787688
>NODE_36_length_849_cov_40.967617
>NODE_37_length_789_cov_6.200843
>NODE_38_length_726_cov_116.842835
>NODE_39_length_721_cov_3.857143
>NODE_40_length_710_cov_48.042654
>NODE_41_length_698_cov_1.014493
>NODE_42_length_661_cov_2.405822
>NODE_43_length_657_cov_1.296552
>NODE_44_length_638_cov_1.039216
>NODE_45_length_625_cov_1.105839
>NODE_46_length_595_cov_1.426641
>NODE_47_length_592_cov_1.636893
>NODE_48_length_542_cov_1.068817
>NODE_49_length_497_cov_1.919048
>NODE_50_length_489_cov_0.902913

***Figure 15**. Example of the labels of the SPAdes contigs were from other species when they were Blasted against NR-database, such as NODE_25 and NODE_35.*

Mapping against reference genome using BWA

There were no major differences in alignment among all labs detected. Lower alignment score refers to contamination of samples (e.g., contamination of adapter sequence) or drop in sequence quality or other factors. Labs 6, 7, 9 and 10 had lower alignment rates for the strains Campy1 compared to the other labs. Labs 7 and 9 had lower alignment rates for the strain Campy2. In general, Lab10 for the strain Campy1 "stood out" as worse than the other labs (Figure 16).

(Lab7 used Illumina MiSeq instrument and TruSeq library preparation kit).

The strain Campy1 *The strain Campy2*

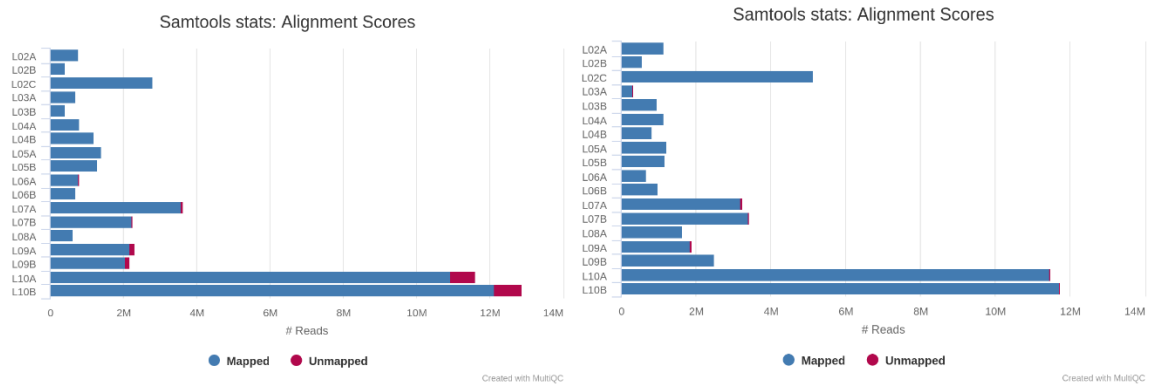


***Figure 16.** Samtools statistics: Alignment scores for both strains Campy1 and Campy2.*

Comparison between "Per library kit (Nextera, NEbNext and TruSeq)"

The labs that used Nextera were Labs 2-4, Labs 6 and 8. Nextera DNA flex was used by Labs 2-4. Nextera XT DNA was used by Labs 6 and 8. The labs that used Nextera with the index XT showed worse assembly and higher GC content than the other labs. Three labs used NEBNext library kit: Lab5 and Labs 9-10. NEBNext Ultra II FS DNA was used by Labs 9-10. NEBNext Ultra II DNA was used by Lab5. TruSeq was used by the Lab7 (Figure17).

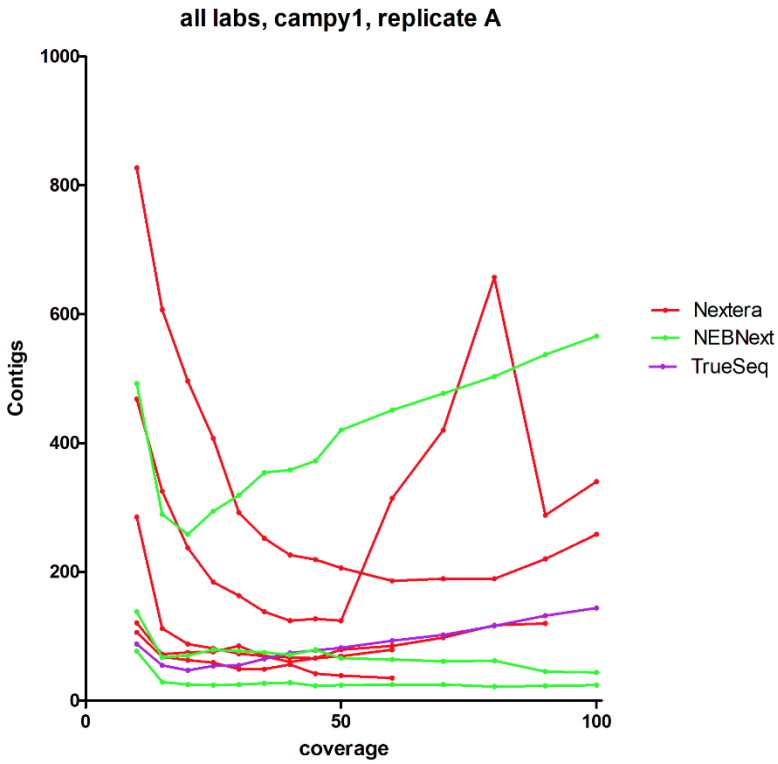


Figure 17. *The comparison using Nextera, NEBNext and TruSeq. Nextera gave the worst assembly and NEBNext gave the best assembly results.*

Discussion

The evolution of NGS (next-generation sequencing) techniques has facilitated researchers to study the world of microorganisms from deeper and broader perspectives. The development of DNA sequencing techniques has not only aided the accurate characterization of bacterial genomes, but also enabled a deeper understanding for the taxonomic structure of the microbiome associated with food. Thus, NGS analysis of microbial pathogens are becoming a conventional analysis and is making it possible to routinely sequence the whole genomes. This gives an overview of a complete map of genetic markers and enables to infer causation by direct testing of the genetic relationship and genetic origin (43).

It is somewhat surprising that different labs sequencing the same DNA get as different data as shown in this thesis. The reason behind this could refer to individual laboratories or contamination of other species despite contamination check using Kraken/Krona detected no contamination problem. Contaminations could have been so small that they were not visible in Krona/multiqc, but they still give rise to small contigs. This is best removed by filtering out low coverage contigs, thus filtering low coverage contigs improved the assemble curves and the investigation of the low coverage contigs confirmed their origin from other species. The contamination could come from the sequencing machine that the data was produced due to prior working with another species than *Campylobacter*. We concluded that the contaminations probably came from previous runs on the machine. Perhaps an extra quality check step could be added that always compares the data with what was sequenced in the previous run and thereby identifying contaminating reads. These low-level contaminations are probably a larger problem when metagenomic samples are sequenced. Metagenomic samples are expected to contain small low coverage contigs and low coverage contigs can therefore not be filtered away without removing contigs that are not contaminations.

Adapters are important to remove. The adapters for the Lab9 were removed using a Trimmomatic tool to improve the data for this lab. However still Lab9 performed not so well after trimming the adapter so other problems probably remained. The most surprising result here is that failing to remove adapters can for certain lead to specific coverages producing extremely bad assemblies as was shown by the coverage 40X for the strain Campy2 (Figure 7). The reason why the bump appears at specific coverage 40X is however, still unknown.

In most libraries the sequences only occur once in the final set. A small number of duplicates may demonstrate a high amount of target sequence coverage. However, a high number of duplicates is likely to demonstrate a kind of enrichment bias (such as during PCR amplification). The FastQC-graph indicates the degree of duplicates for each sequence in a library: the relative number of sequences with different degrees of duplicates. In other words, shotgun sequencing gives reads starting at random positions. But there are less than 2 million positions in a *Campylobacter* genome thus, if we sequence ten million reads, there will be reads starting at the same position generating “duplicate reads”. Basically, in low coverage and large genome the duplicate errors are a problem but in the short genome, such as *Campylobacter*, it is not a problem. The Nextera kit fragments the genome using less positions and therefore the duplicate error flags appear easier when using this library kit.

Comparing the quality of each library kit we found that the best performing NEBNext lab gives better assembly than the best Nextera lab. It is possible to see this by looking at the tables and figures above. Also, in comparison with different variants among the Nextera and NEBNext library preparation kits: the labs that used NEBNext with the variant FS showed better assembly results than the other labs and they were the only labs that did not show the last base composition bias of the very last sequence, whereas the

labs that used Nextera library preparation kit generated base composition bias in the beginning of the reads, generating error flags. We conclude that “Per base sequence content” error flags are largely coupled to the library prep kit used. The reason why Nextera gives sequence composition bias could be due to how the fragmentation is achieved in Nextera. This may reduce the assembly quality due to the fragmentation being less random.

***Fragmentation in Nextera*:** An engineered transposase is used in Nextera to concurrently fragment and tag ("tagment") input DNA. Unique adapter sequences are then added in the process. These adapter sequences are used to amplify the insert DNA in a limited-cycle PCR reaction (Figure 18). The transposase used for fragmentation has a sequence bias (it prefers certain sequences). It seems that Nextera prefers AT rich regions rather than GC rich regions. Fragmentation in NEBNext is also enzymatic, but not transposase based (NEBNext uses a nicking enzyme). TruSeq uses mechanical fragmentation (ultrasound) and is therefore less prone to produce sequence-specific bias.

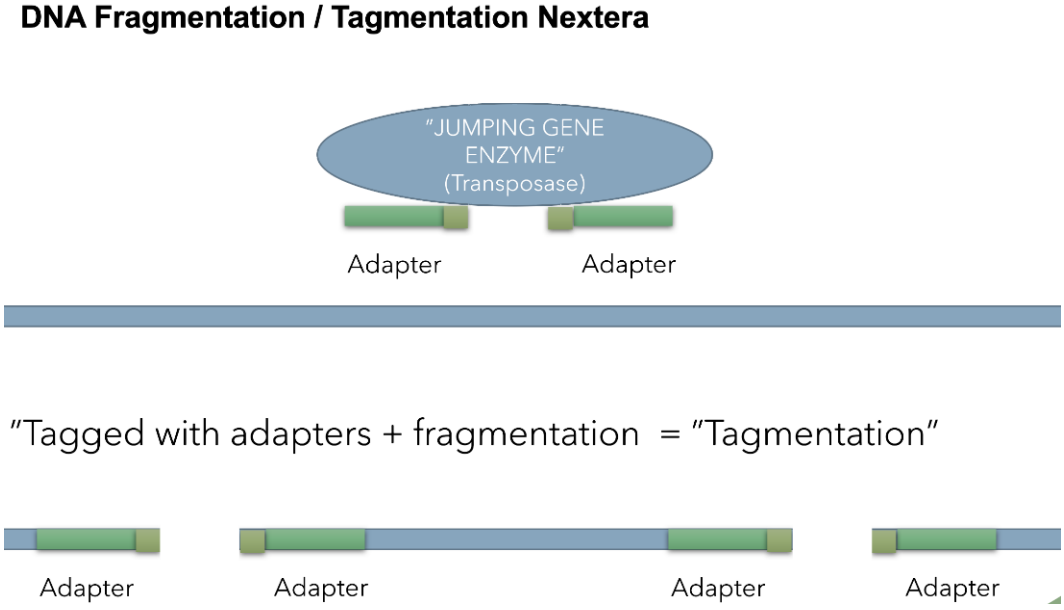


Figure 18. *Nextera uses transposase for fragmentation. (Made using PowerPoint).*

Since the problem is that some genomic regions are avoided by the transposase, and they will not come back by trimming the data, trimming this bias will likely not improve assembly. There will be regions in the genome that have lower coverage because there are less positions where the transposase cuts. Depending on how the base composition looks like in different regions of the genome sequence there might be regions which are not covered by sequenced fragments and then there will always be a gap in the WGS sequence coverage in this region. Some genome regions may generate less fragments than other regions but still enough to be able to assemble the region if enough data is produced. Transposase library preparation kits may need higher sequence coverage to generate good enough data. This makes sequencing more expensive, since more sequencing reads are needed for each sample. Mechanical fragmentation is probably the best way to do it, but it requires a special equipment. This may be expensive for a small lab and therefore enzymatic fragmentation is more popular in small labs.

After solving the problems that were detected in FastQC, two labs had still bad assembly which were probably connected to Nextera XT variant. Labs 6 and 8 had higher GC content in their sequences than the

other labs and these two labs used Nextera XT library kit. It has also been seen before that Nextera XT gives a strong sequence bias (44).

It is not strange that Read2 showed poorer quality compared to Read1. The Illumina paired end sequencing involves a temporal separation and a mechanistic process. The "first in pairs" are sequenced first then the fragments "bent" over, reattached, and thereafter the reads in the second pair sequenced on the same flow cell surface. Depending on the protocol, this second step may occur days later. Hence it is not unexpected to observe lower qualities for Read2.

All labs except two (used NEBNext non-FC variant, variant FS) have a last base composition bias. The reason could refer to the library preparation kit and may be related to the adapter trimming – by trimming more sequence than necessary. For example, if the adapter started with nucleotide A, the machine trimmed the adapter as well as sequences that started with nucleotide A considering it is an adapter.

The Nextera library preparation kit with the variant XT gave worst assembly results. The labs that used NEBNext gave better assembly results with the variant FS. Nextera XT is probably the worst library kit and NEB next the best among the ones analysed in this study.

Lab2 did not produce identical replicates, but instead sequenced the same sample on three different Illumina machines: iSeq, MiSeq and NextSeq. This gave an opportunity for a comparison between these three machines. The “Per base sequence quality” showed high quality along the whole reads for iSeq and MiSeq, while the NextSeq data showed a slight drop in sequence quality towards the end. Despite this quality drop, the NextSeq (L02C) gave fewer contigs and larger N50 (i.e., better assembly) than the others for both strains Campy1 and Campy2. The reason for this is unknown, and somewhat surprising, since the “Per base quality” was worse in the NextSeq data compared to the others.

In summary this study shows that the final assembly quality can be affected by a large number of factors, and it is important for each lab to learn how to control their data for these factors. Optimizing WGS performance in a lab would include selecting a well performing library preparation kit. The raw data should be checked for adapter content and poor-quality bases in the ends of the reads. If needed, these should be trimmed off. Trimming adapters is very important but trimming based on quality is probably only needed if very bad quality bases are present. Checking for contamination will identify if a major part of the sample is contaminated, but smaller levels of contaminations are best handled by removing small and low coverage contigs. This study was performed using data from *Campylobacter jejuni* which has a low GC content compared to other species and it is possible that some of the conclusions drawn here could be different for other species. Other species may have different average GC content and therefore perform differently with different library preparation kits.

Acknowledgment

All through the composing of this thesis I have gotten an awesome bargain of back and help. I want to begin with thanking my supervisor, Bo Segerman, (Department of Medical Biochemistry and Microbiology, Uppsala University), whose ability in defining the investigation questions and strategies was important for the project. Your shrewd input pushed me to hone my considering and brought my work to a better level.

Also, I would like to thank my family for their shrewd guide and thoughtful ear.

References

1.Tauxe RV, Doyle MP, Kuchenmüller T, Schlundt J, Stein CE. Evolving public health approaches to the global challenge of foodborne infections. Int J Food Microbiol. 2010 May 30;139 Suppl 1: S16-28. doi: 10.1016/j.ijfoodmicro.2009.10.014. Epub 2009 Oct 29. PMID: 19931203.

2. McInnes, Colin , "The Many Meanings of Health Security" , in *Routledge Handbook of Global Health Security* ed. Simon Rushton and Jeremy Youde (Abingdon: Routledge, 26 Aug 2014), accessed 27 Jun 2021, Routledge Handbooks Online.

3.Scallan E, Hoekstra RM, Angulo FJ, Tauxe RV, Widdowson MA, Roy SL, Jones JL, Griffin PM. Foodborne illness acquired in the United States--major pathogens. Emerg Infect Dis. 2011 Jan;17(1):7-15. doi: 10.3201/eid1701.p11101. PMID: 21192848; PMCID: PMC3375761.

4.Epps SV, Harvey RB, Hume ME, Phillips TD, Anderson RC, Nisbet DJ. Foodborne Campylobacter: infections, metabolism, pathogenesis and reservoirs. Int J Environ Res Public Health. 2013 Nov 26;10(12):6292-304. doi: 10.3390/ijerph10126292. PMID: 24287853; PMCID: PMC3881114.

5.Altekruse SF, Swerdlow DL, Stern NJ. Microbial food borne pathogens. Campylobacter jejuni. Vet Clin North Am Food Anim Pract. 1998 Mar;14(1):31-40. PMID: 9532665.

6.Swaminathan S, Ellis HM, Waters LS, Yu D, Lee EC, Court DL, Sharan SK. Rapid engineering of bacterial artificial chromosomes using oligonucleotides. Genesis. 2001 Jan;29(1):14-21. doi: 10.1002/1526-968x(200101)29:1<14::aid-gene1001>3.0.co;2-x. PMID: 11135458.

7.Aarestrup FM. The livestock reservoir for antimicrobial resistance: a personal view on changing patterns of risks, effects of interventions and the way forward. Philos Trans R Soc Lond B Biol Sci. 2015 Jun 5;370(1670):20140085. doi: 10.1098/rstb.2014.0085. PMID: 25918442; PMCID: PMC4424434.

8.Allard, Opalko H E, et al. Stable Pom1 clusters form a glucose modulated concentration gradient that regulates mitotic entry .2018, eLife 2019;8: e46003 DOI: 10.7554/eLife.46003

9.Allard JP, Keller H, Jeejeebhoy KN, Laporte M, Duerksen DR, Gramlich L, Payette H, Bernier P, Vesnaver E, Davidson B, Teterina A, Lou W. Malnutrition at Hospital Admission-Contributors and Effect on Length of Stay: A Prospective Cohort Study from the Canadian Malnutrition Task Force. JPEN J Parenter Enteral Nutr. 2016 May;40(4):487-97. doi: 10.1177/0148607114567902. Epub 2015 Jan 26. PMID: 25623481.

10.Jackson BR, Tarr C, Strain E, Jackson KA, Conrad A, Carleton H, Katz LS, Stroika S, Gould LH, Mody RK, Silk BJ, Beal J, Chen Y, Timme R, Doyle M, Fields A, Wise M, Tillman G, Defibaugh-Chavez S, Kucerova Z, Sabol A, Roache K, Trees E, Simmons M, Wasilenko J, Kubota K, Pouseele H, Klimke W, Besser J, Brown E, Allard M, Gerner-Smidt P. Implementation of Nationwide Real-time Whole-genome Sequencing to Enhance Listeriosis Outbreak Detection and Investigation. Clin Infect Dis. 2016 Aug 1;63(3):380-6. doi: 10.1093/cid/ciw242. Epub 2016 Apr 18. PMID: 27090985; PMCID: PMC4946012.

11.Eric Brown, Uday Dessai, Sherri McGarry, and Peter Gerner-Smidt.Foodborne Pathogens and Disease.Jul 2019.441-450.http://doi.org/10.1089/fpd.2019.2662

12.Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977 Dec;74(12):5463-7. doi: 10.1073/pnas.74.12.5463. PMID: 271968; PMCID: PMC431765.

13.Margulies, M., Egholm, M., Altman, W. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437, 376–380 (2005). https://doi.org/10.1038/nature03959

14.Vincent, W.J.B., Harvie, E.A., Sauer, J.D., Huttenlocher, A. (2017) Neutrophil derived LTB4 induces macrophage aggregation in response to encapsulated Streptococcus iniae infection. PLoS One. 12: e0179574

15.https://www.sciencedirect.com/topics/immunology-and-microbiology/illumina-dye-sequencing

16.Ansorge WJ. Next-generation DNA sequencing techniques. N Biotechnol. 2009 Apr;25(4):195-203. doi: 10.1016/j.nbt.2008.12.009. Epub 2009 Feb 3. PMID: 19429539.

17.Merriman B; Ion Torrent R&D Team, Rothberg JM. Progress in ion torrent semiconductor chip-based sequencing. Electrophoresis. 2012 Dec;33(23):3397-417. doi: 10.1002/elps.201200424. Erratum in: Electrophoresis. 2013 Feb;34(4):619. PMID: 23208921.

18.Rothberg JM, Leamon JH. The development and impact of 454 sequencing. Nat Biotechnol. 2008 Oct;26(10):1117-24. doi: 10.1038/nbt1485. PMID: 18846085.

19.Zeng F, Jiang R, Chen T. PyroHMMsnp: an SNP caller for Ion Torrent and 454 sequencing data. Nucleic Acids Res. 2013 Jul;41(13): e136. doi: 10.1093/nar/gkt372. Epub 2013 May 21. PMID: 23700313; PMCID: PMC3711422.

20.Lysholm F, Andersson B, Persson B. FAAST: Flow-space Assisted Alignment Search Tool. BMC Bioinformatics. 2011 Jul 19; 12:293. doi: 10.1186/1471-2105-12-293. PMID: 21771335; PMCID: PMC3228549.

21.Feng, W., Zhao, S., Xue, D. *et al.* Improving alignment accuracy on homopolymer regions for semiconductor-based sequencing technologies. *BMC Genomics* 17, 521 (2016). https://doi.org/10.1186/s12864-016-2894-9

22.James M. Heather, Benjamin Chain,The sequence of sequencers: The history of sequencing DNA,Genomics,Volume 107, Issue 1,2016, Pages 1-8, ISSN 0888-7543,

https://doi.org/10.1016/j.ygeno.2015.11.003.

23.Adewale BA. Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? Afr J Lab Med. 2020 Nov 26;9(1):1340. doi: 10.4102/ajlm. v9i1.1340. PMID: 33354530; PMCID: PMC7736650.

24.Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, Leon A, Pullambhatla M, Temple-Smolkin RL, Voelkerding KV, Wang C, Carter AB. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn. 2018 Jan;20(1):4-27. doi: 10.1016/j.jmoldx.2017.11.003. Epub 2017 Nov 21. PMID: 29154853.

25.Oakeson KF, Wagner JM, Mendenhall M, Rohrwasser A, Atkinson-Dunn R. Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerg Infect Dis. 2017 Sep;23(9):1441-1445. doi: 10.3201/eid2309.170416. PMID: 28820135; PMCID: PMC5572866.

26.Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

27.Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16. PMID: 22506599; PMCID: PMC3342519.

28.Nikolenko SI, Korobeynikov AI, Alekseyev MA. BayesHammer: Bayesian clustering for error correction in single-cell sequencing. BMC Genomics. 2013;14 Suppl 1(Suppl 1): S7. doi: 10.1186/1471-2164-14-S1-S7. Epub 2013 Jan 21. PMID: 23368723; PMCID: PMC3549815.

29.Oakeson KF, Wagner JM, Mendenhall M, Rohrwasser A, Atkinson-Dunn R. Bioinformatic Analyses of Whole-Genome Sequence Data in a Public Health Laboratory. Emerg Infect Dis. 2017 Sep;23(9):1441-1445. doi: 10.3201/eid2309.170416. PMID: 28820135; PMCID: PMC5572866.

30.Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

31.Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257. doi: 10.1186/s13059-019-1891-0. PMID: 31779668; PMCID: PMC6883579.

32.Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12:385. doi: 10.1186/1471-2105-12-385. PMID: 21961884; PMCID: PMC3190407.

33.Uelze L, Borowiak M, Bönn M, et al. German-Wide Interlaboratory Study Compares Consistency, Accuracy and Reproducibility of Whole-Genome Short Read Sequencing. Frontiers in Microbiology. 2020; 11:573972. DOI: 10.3389/fmicb.2020.573972.

34.Laura Uelze, Maria Borowiak, Erik Brinks, Carlus Deneke, Kerstin Stingl, Sylvia Kleta, Simon H. Tausch, Kathrin Szabo, Anne Wöhlke, Burkhard Malorny

bioRxiv 2020.04.22.054759; doi: https://doi.org/10.1101/2020.04.22.054759.

35.https://www.bioinformatics.babraham.ac.uk/projects/FastQC/.

36.Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016 Oct 1;32(19):3047-8. doi: 10.1093/bioinformatics/btw354. Epub 2016 Jun 16. PMID: 27312411; PMCID: PMC5039924.

37.Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114-20. doi: 10.1093/bioinformatics/btu170. Epub 2014 Apr 1. PMID: 24695404; PMCID: PMC4103590.

38.Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77. doi: 10.1089/cmb.2012.0021. Epub 2012 Apr 16. PMID: 22506599; PMCID: PMC3342519.

39.Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. Genome Biol. 2019 Nov 28;20(1):257. doi: 10.1186/s13059-019-1891-0. PMID: 31779668; PMCID: PMC6883579.

40.Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011 Sep 30; 12:385. doi: 10.1186/1471-2105-12-385. PMID: 21961884; PMCID: PMC3190407.

41.Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009 Jul 15;25(14):1754-60. doi: 10.1093/bioinformatics/btp324. Epub 2009 May 18. PMID: 19451168; PMCID: PMC2705234.

42.Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078-9. doi: 10.1093/bioinformatics/btp352. Epub 2009 Jun 8. PMID: 19505943; PMCID: PMC2723002.

43.Pereira R, Oliveira J, Sousa M. Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics. *Journal of Clinical Medicine*. 2020; 9(1):132. https://doi.org/10.3390/jcm9010132 .

44. Sato MP, Ogura Y, Nakamura K, Nishida R, Gotoh Y, Hayashi M, Hisatsune J, Sugai M, Takehiko I, Hayashi T. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. 2019 Oct 1;26(5):391-398. doi: 10.1093/dnares/dsz017. PMID: 31364694; PMCID: PMC6796507.

Appendices

Descriptions of the labs

For Lab2:

Lab2 sequenced the two strains with three replicates using three different instruments, Illumina iSeq 100 (L02A), Illumina MiSeq (L02B), and NextSeq 500 (L02C). Lab2 used the Illumina Nextera Flex library kit for all sequencing replicates. The iSeq and the NextSeq data was 2x151 bp and the MiSeq data was 2x201 bp. The iSeq run gave 106 and 164 Mbases (66 and 96X coverage, respectively), the MiSeq 80 and 113 Mbases (50 and 66X coverage, respectively), and the NextSeq 419 and 766 Mbases (262 and 451X coverage, respectively).

FastQC showed an error flag in "Per base sequence content" for all fastq-files from Lab2. All the following graphs showed a similar pattern of abnormal base composition in the first 9 bases of the reads (Figure 1).

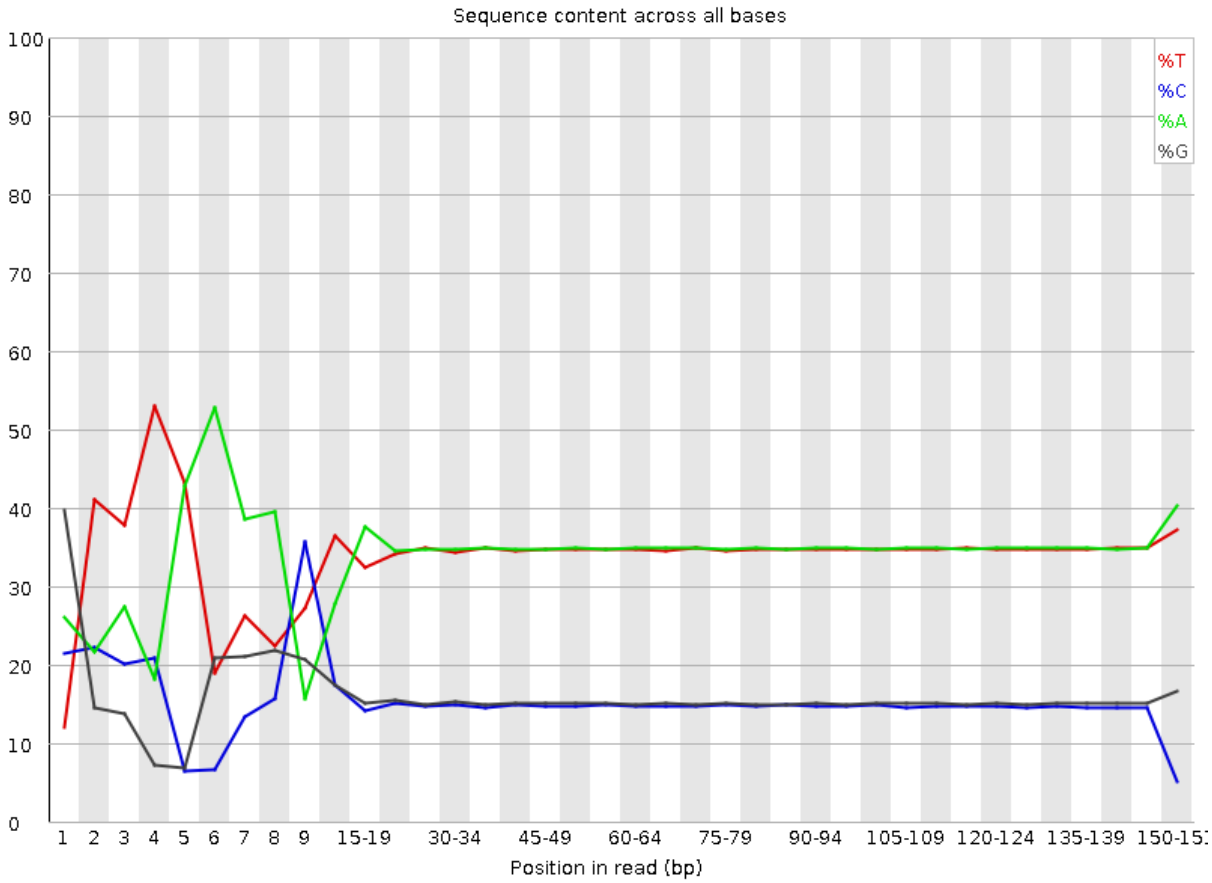
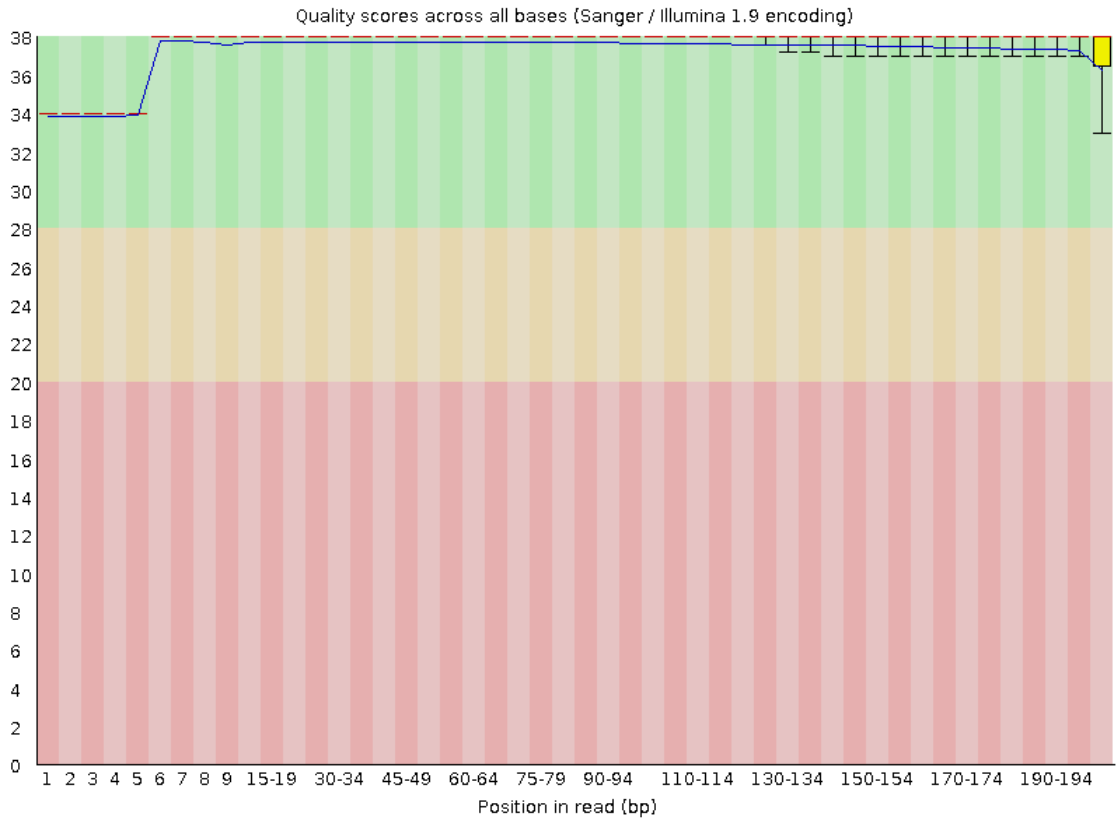
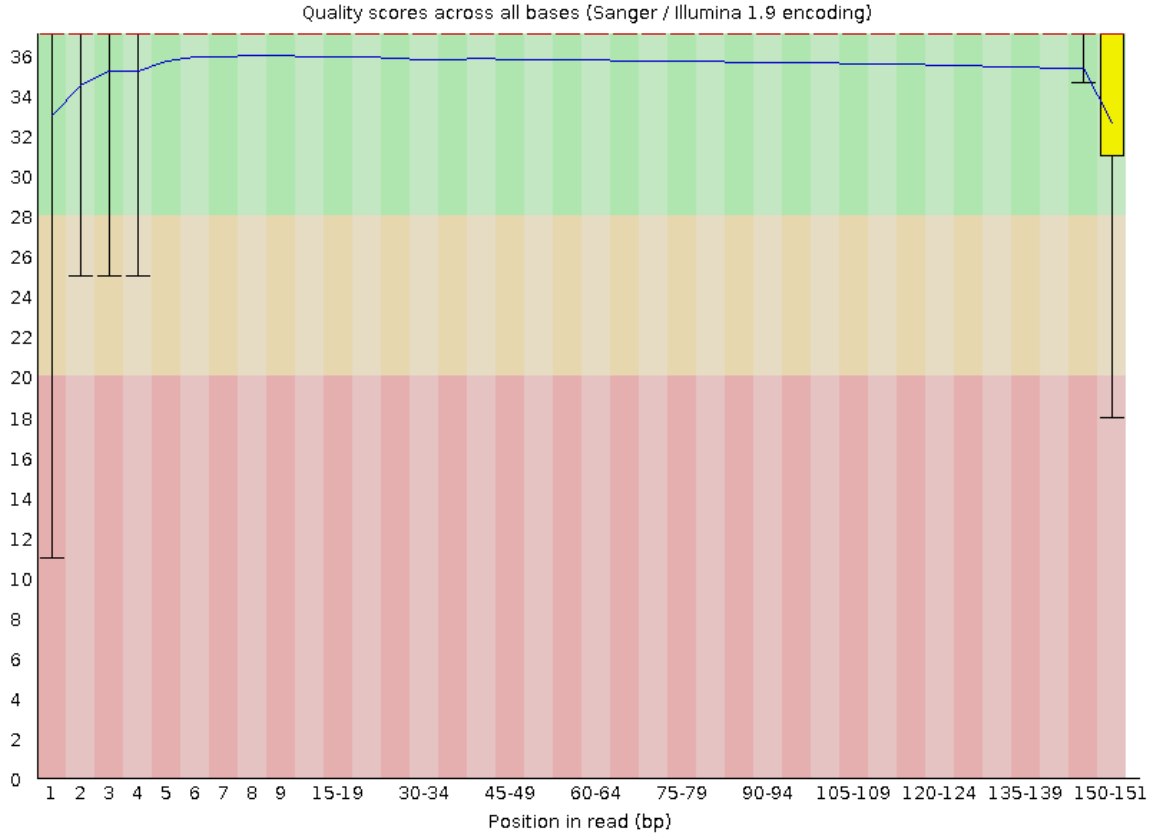


***Figure 1**. The "Per base sequence content" showed an abnormal base composition in the first 9 bases. The pattern seen is likely linked to the transposase-based fragmentation. Also, the very last base showed abnormal base composition.*

The "Per base sequence quality" showed high quality along the whole reads. However, the NextSeq data showed a slight drop in sequence quality towards the end (Figure 2).


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 110-114 130-134 150-154 170-174 190-194
Position in read (bp)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150-151
Position in read (bp)

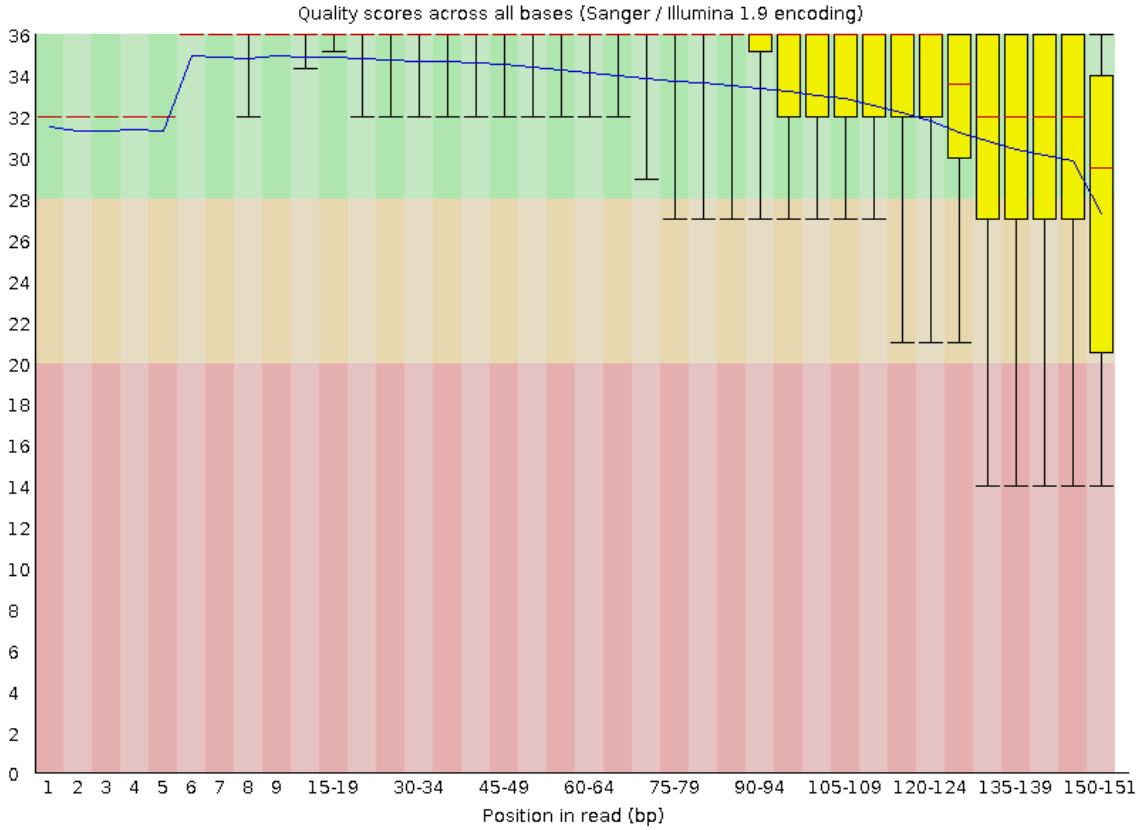


Figure 2. *The "Per base sequence quality" of the L02A (iSeq), L02B (MiSeq) and L02C (NextSeq) data for strain Campy1.*

FastQC-file for the L02C showed an error flag in sequence duplication levels. The reason could be a consequence of high total coverage. It could be that the reads have a bias towards starting at a certain position (transposon selectivity) (Figure 3).

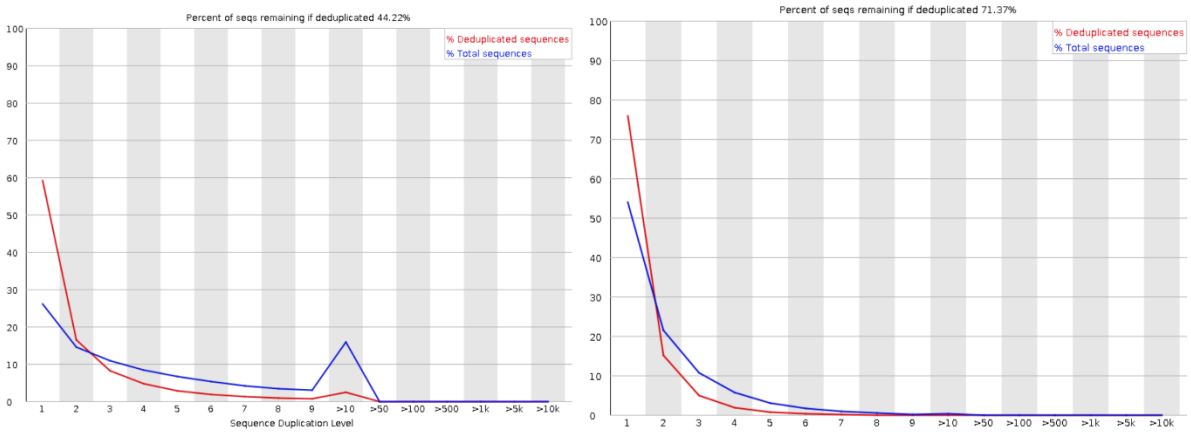


Figure 3. *The error flag in "Sequence Duplication Levels" for the L02C (strain Campy1) at the coverage 262X on the left versus no error flag at the coverage 60X on the right.*

The fastq-files were then taxonomically classified with Kraken2, in order to quantify any putative contaminations of other species. In a Krona plot of the Kraken classification, no contaminations were visible (Figure 4).

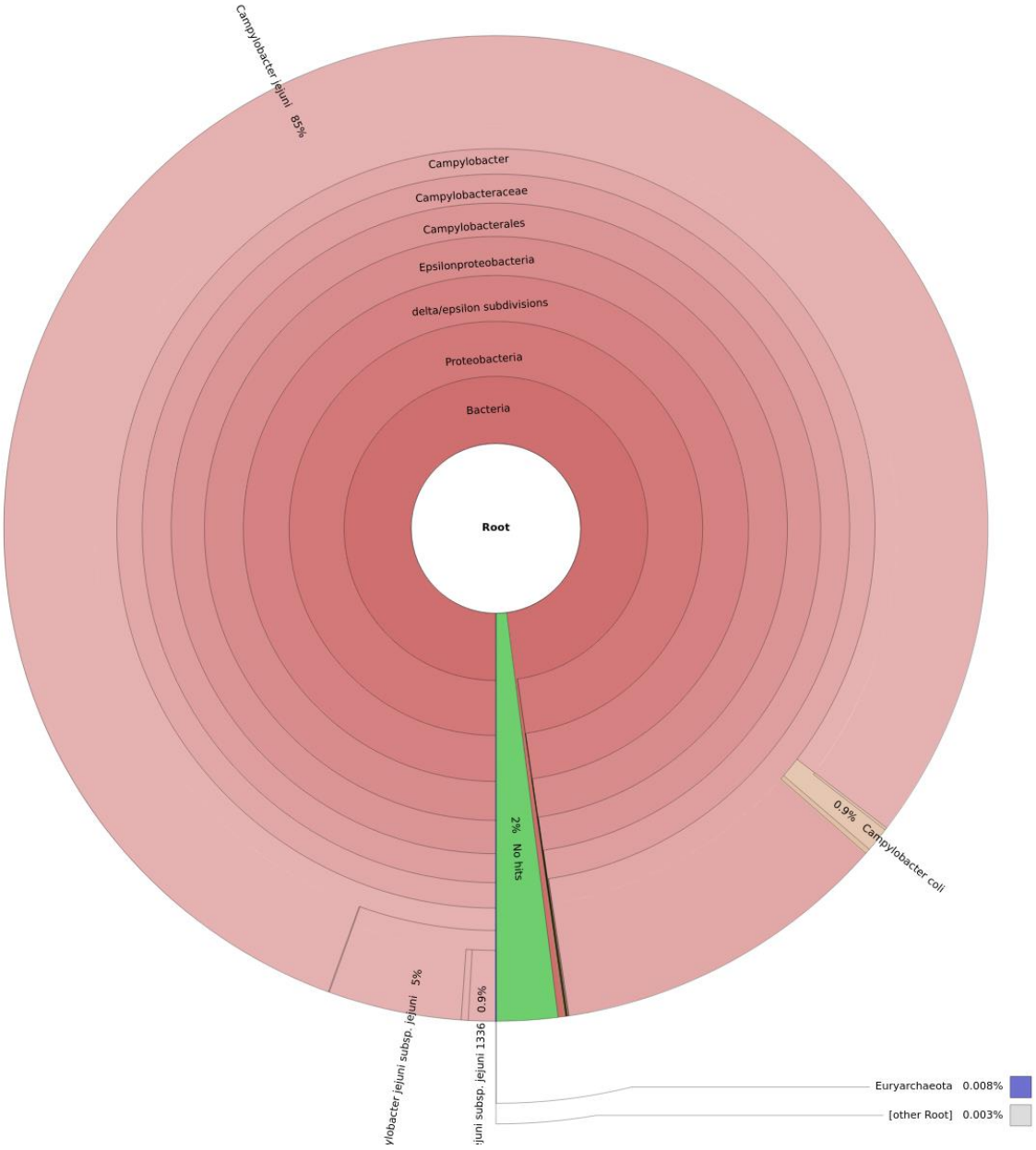


Figure 4. *Example (Strain Campy1, sample L02A) of a Krona plot of the Kraken2 taxonomic classification using the Minikraken database. The Krona plots were all very similar but the "No Hit" category varied whilst always being less than 2%. Very minor proportions of the reads were classified as other taxa than Campylobacter indicating no major contamination problem was present for Lab2.*

The data from Lab2 was then assembled with SPAdes using a titration of the amount of data used (coverage). The NextSeq data (L02C) produced fewer contigs and larger N50 (i.e., a better assembly) than the others for both strains Campy1 and Campy2. The reason for this is unknown, and somewhat surprising, since the "Per base sequence quality" was worse in the NextSeq data as compared to the others. There was often a trend that the amount of contigs first rapidly decreased and the slowly increased again as the amount of data used was increased. This indicates that there was an optimal coverage, but the optimum was not

always the same. Notably, the curves sometimes jumped up and down for the strain Campy1 indicating the assembly program sometimes performing irregularly along the “coverage axis” when the amount of data is titrated. The N50 curve for the strain Campy2 was steady (Figure 5).

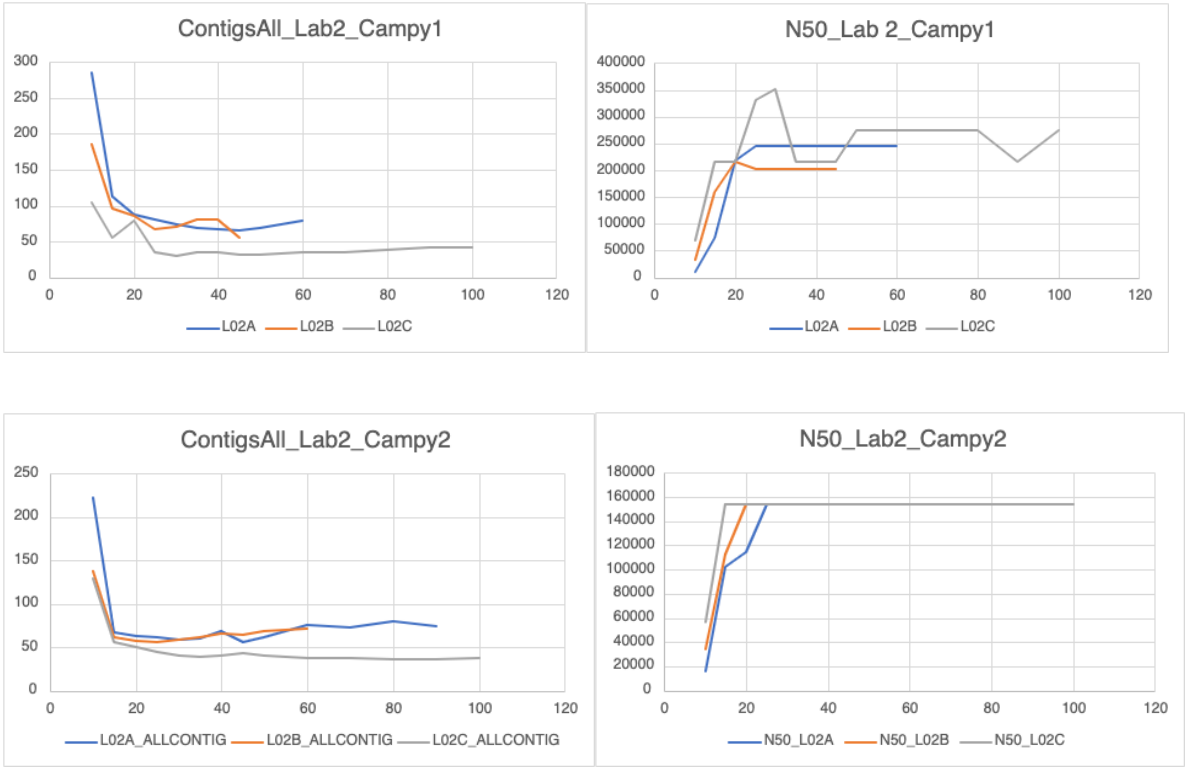


Figure 5*. Change in total number of contigs and N50 values when amount of data used (coverage) was varied.*

For Lab3:

Lab3 sequenced two replicates for the two strains with the same instrument, Illumina iSeq100. Lab3 used the Illumina Nextera Flex library kit for both sequencing replicates. The read length was 2x151 bp and the replicate runs gave 106 and 63 Mbases (65 and 39X coverage, respectively) for the Campy1 strain. The Campy2 strain run gave 48 and 142 Mbases (28 and 83X coverage, respectively). FastQC showed the same error flag as Lab2 in “Per base sequence content” for all fastq-files originating from Lab3. Both these graphs showed a similar Nextera related pattern of abnormal base composition in the first 9 bases of the reads. Also, for this lab, the very last base in the reads showed a deviating base composition (Figure 6).

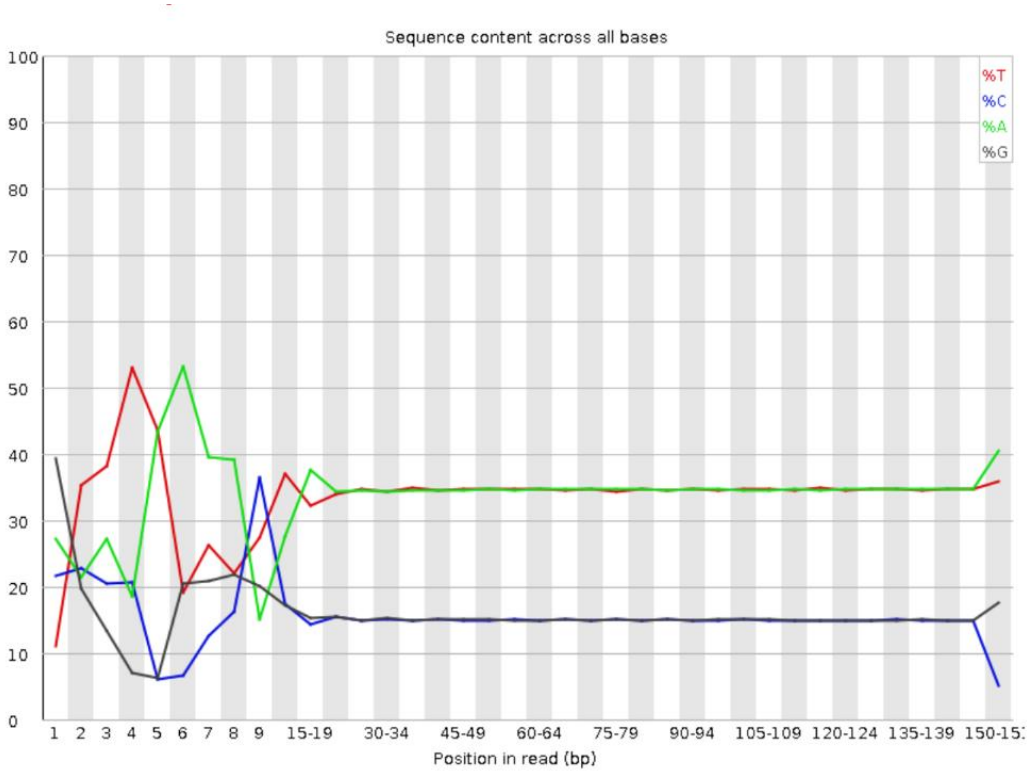
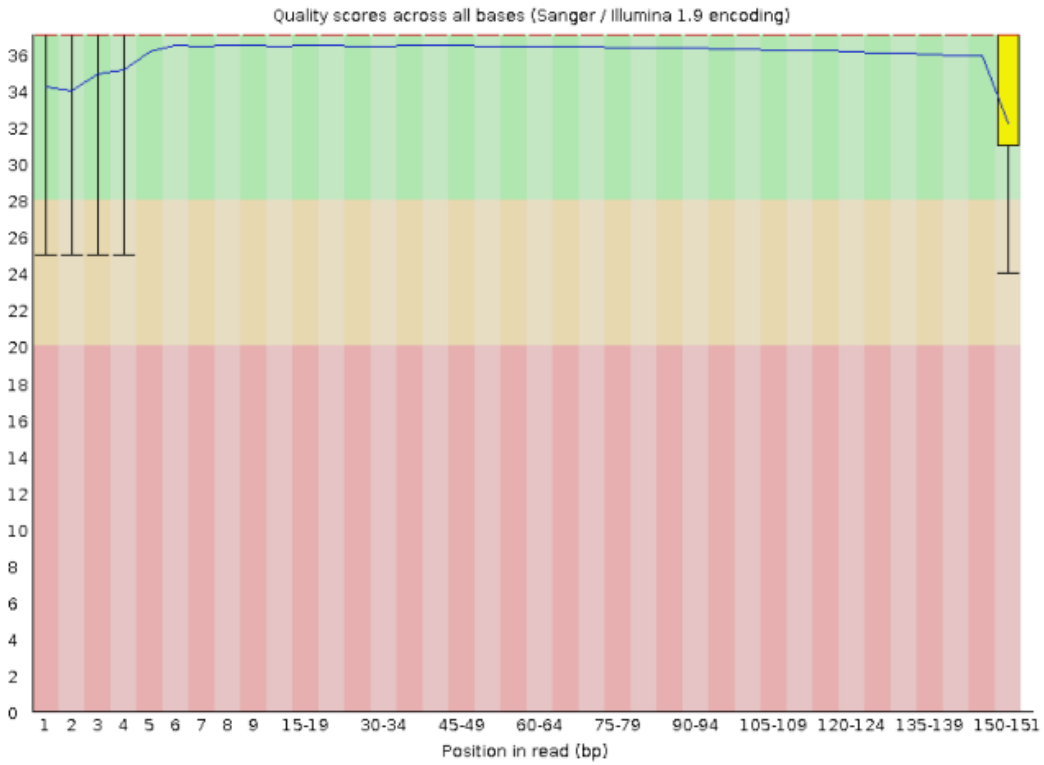


***Figure 6**. The "Per base sequence content" for Lab3.*

The "Per base sequence quality" showed high quality along the whole reads and even better quality compared to Lab2 (Figure 7).



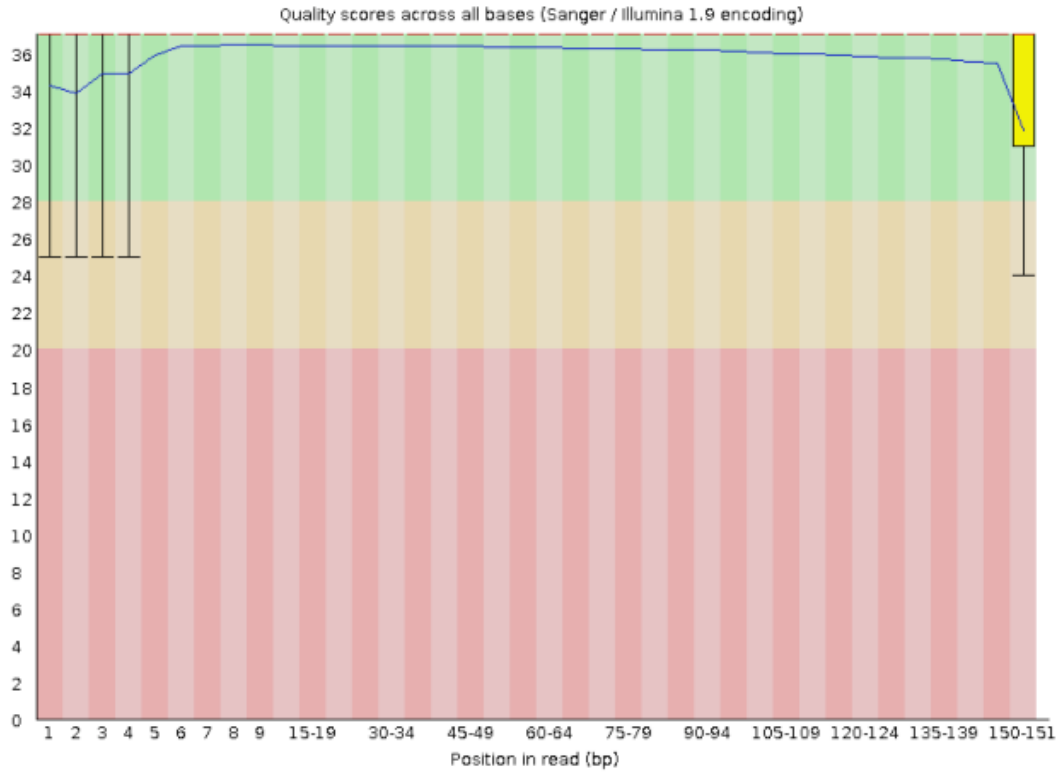


Figure 7*. The "Per base sequence quality" of the L03A (iSeq), L03b (iSeq), data for strain Campy1.*

The fastq-files were taxonomically classified with Kraken2 as was done for Lab2. The Krona plot of the Kraken classification did not show any major contamination problems (Figure 8).

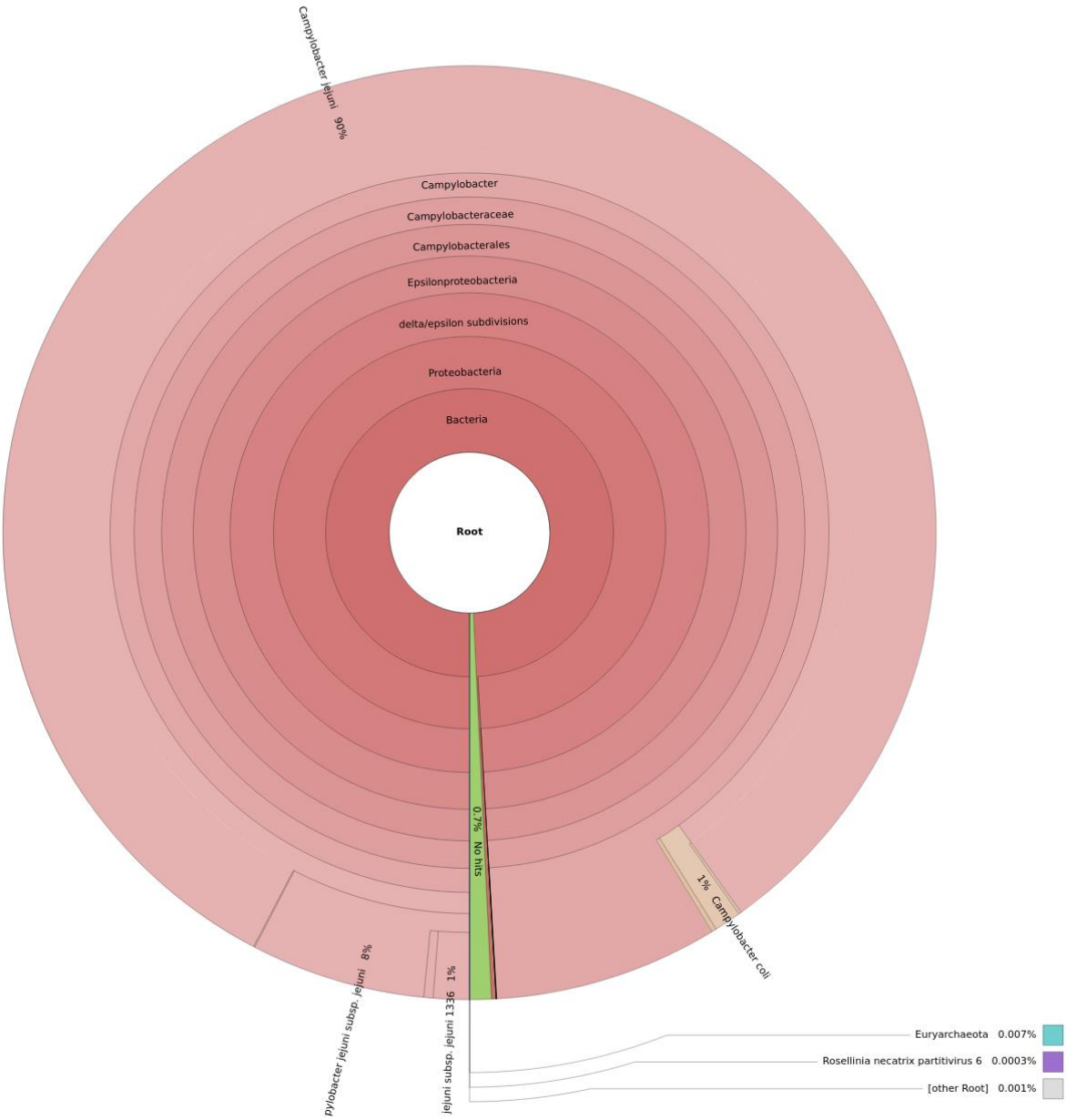


***Figure 8**. Example (Strain Campy1, sample L03A) of a Krona plot of the Kraken2 taxonomic classification using the Minikraken database. In all Krona plots for Lab3, only a very minor proportion of the reads were classified as other taxa than Campylobacter indicating no major contamination problem.*

The Lab3 data was then assembled with SPAdes using a titration of the amount of data used (coverage) (Figure 9). There were slightly fewer contigs and larger N50 for L03A than the iSeq from L03B and Lab2 (L02A, L02B) for both strains Campy1 and Campy2. This indicates that there was an optimal coverage, but the optimum was not always the same. As was seen for Lab2, the assembly program sometimes made irregular "jumps" in output quality along the "coverage axis" when the amount of data was titrated, but it was steady for the strain Campy2.

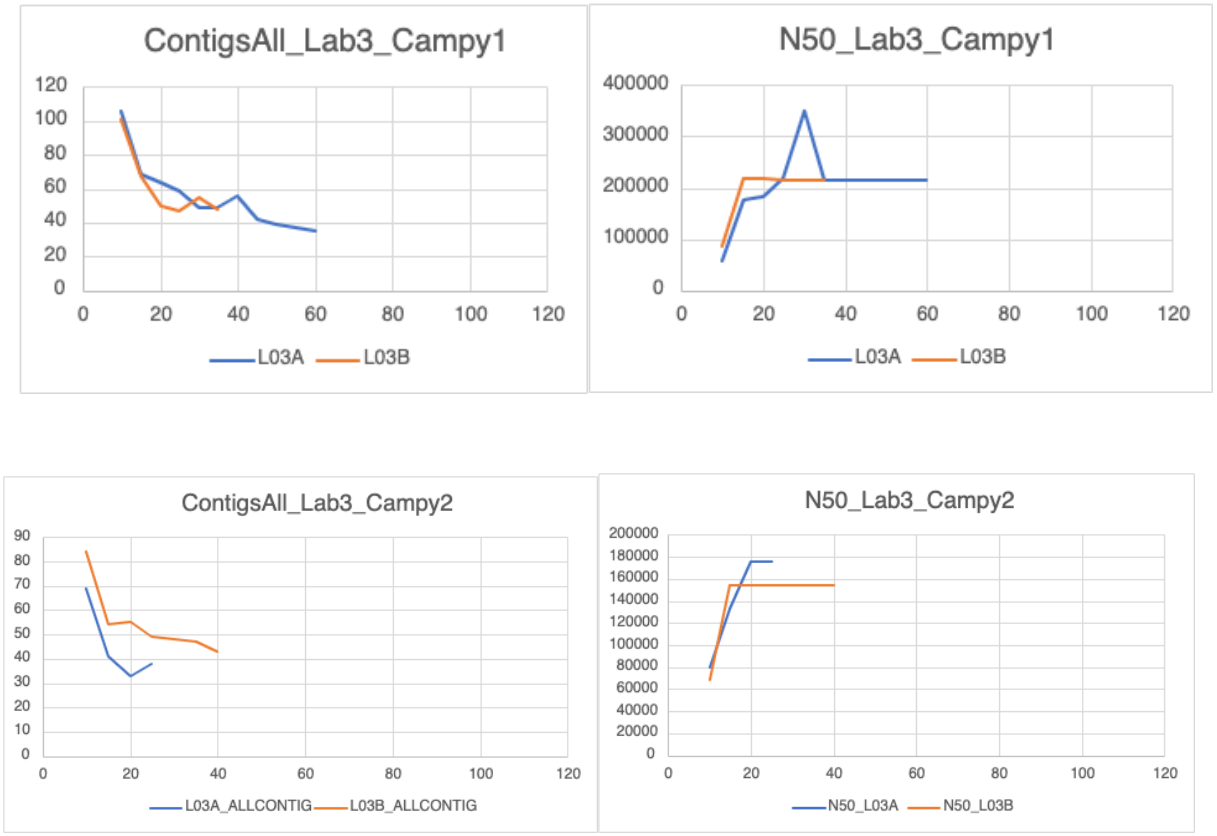


Figure 9*. Change in total number of contigs and N50 values when amount of data used (coverage) was varied.*

For Lab4:

Lab4 sequenced also using the same instrument, Illumina MiSeq, for two replicates as the pervious labs. Lab4 used the Illumina Nextera Flex library kit for both sequencing replicates. The data length was 2x201 bp and the replicates run for Campy1 gave 156 and 232 Mbases (96 and143X coverage, respectively) for Campy2 223 and 162 Mbases (130 and 95X coverage, respectively). We could still see the error flag in "Per base sequence content" for all fastq-files coming from Lab4. And thus, Nextera pattern of abnormal base composition in the first 9 bases of the reads was shown in these graphs besides the deviation composition of the very last base (Figure 10). The "Per base sequence quality" showed high quality along the whole reads (Figure 11).

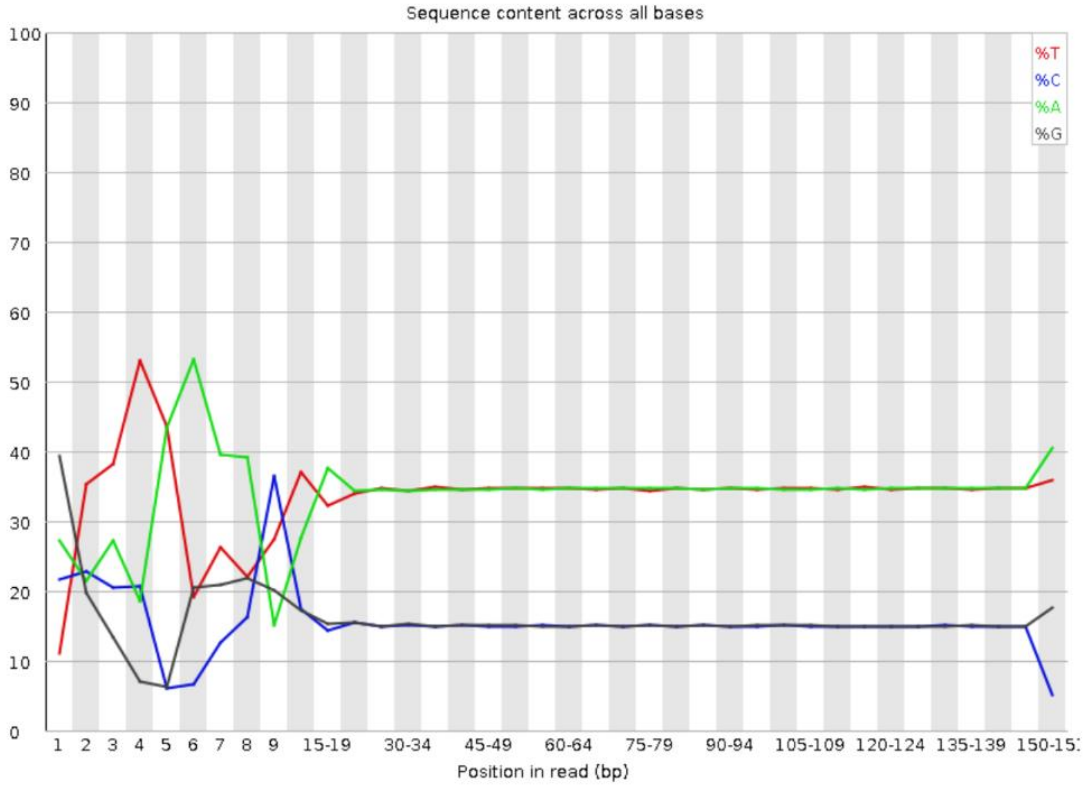
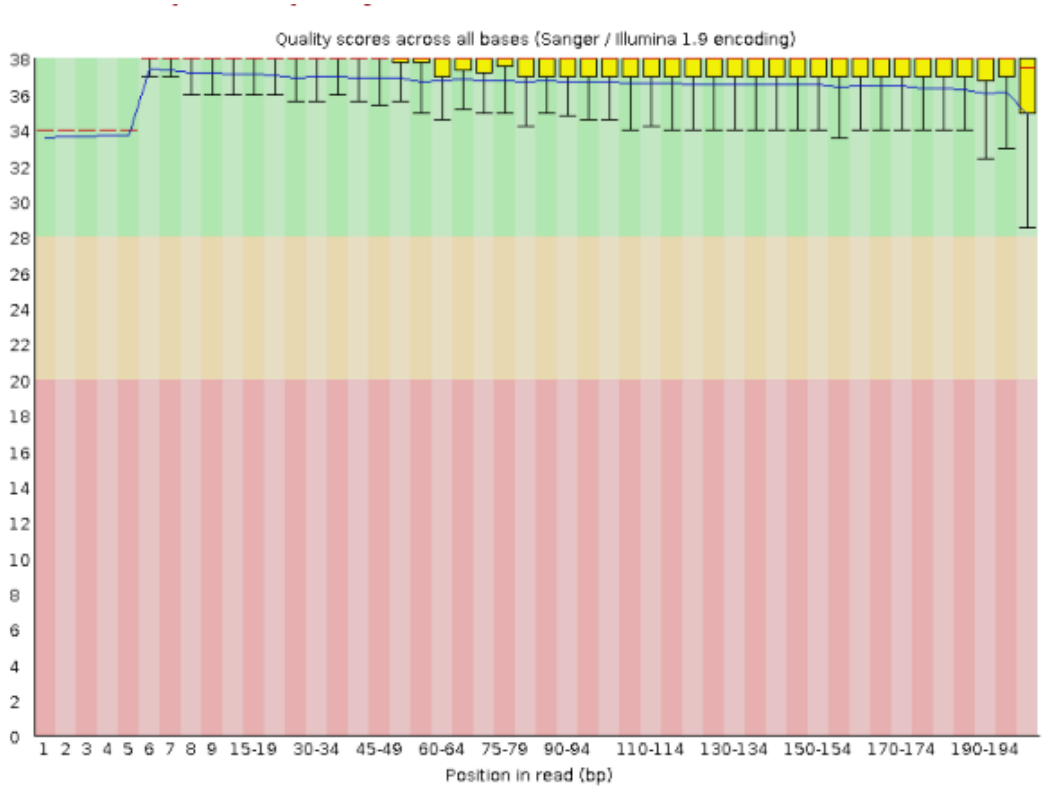


***Figure 10**. The "Per base sequence content" for Lab4.*



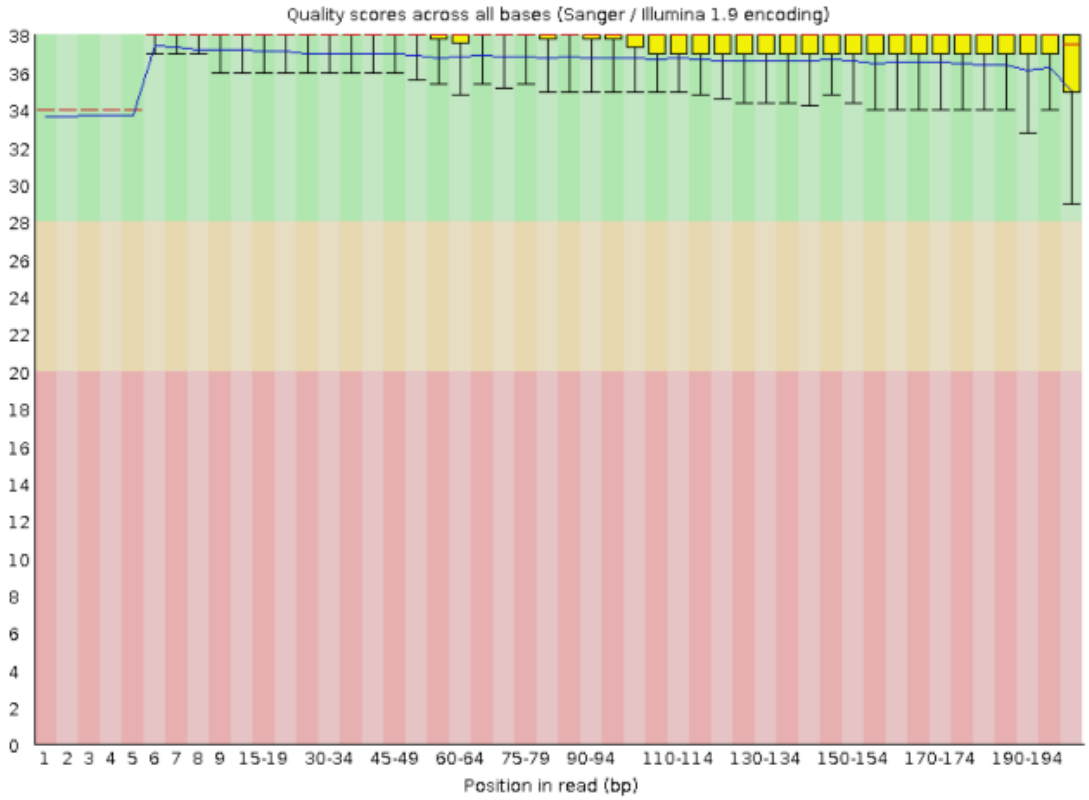


Figure 11. *The "Per base sequence quality" of the L04A, L04B (MiSeq) data for strain Campy1.*

As the pervious labs the taxonomic classification with Kraken2 showed no visible contaminations (Figure 12).

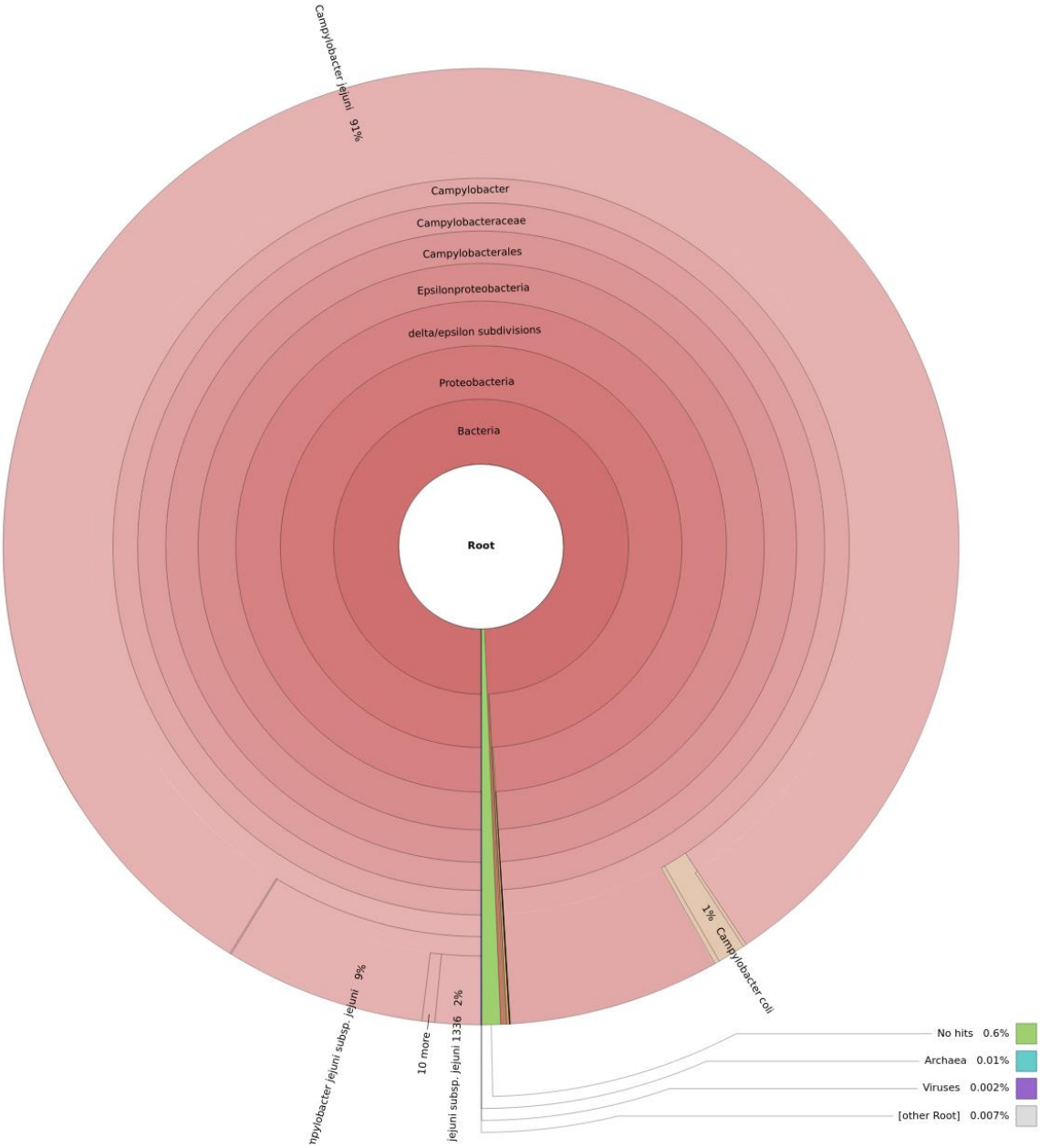


Figure 12. *Example (Strain Campy1, sample L04A) of a Krona plot of the Kraken2 taxonomic classification using the Minikraken database. As the pervious labs, very minor proportions of the reads (less than 2%.) were classified as other taxa than Campylobacter.*

The assembled data with SPAdes from the Lab4 using a titration of the amount of data used (coverage), (Figure 13). For both strains Campy1 and Campy2 the MiSeq gave almost the same number of contigs and the same N50 for both replicates. However, there was a sign of bias from the coverage 60 to 100. The reason could be that it is due to contamination of other species. There was often a trend that the amount of contigs first rapidly decreased and then increased radically again as the amount of data used was increased, whereas the optimum amount of contigs was not always the same. As the pervious labs, the titration of the data showed a “jump” in the curves along the “coverage axis”, however, for the Campy2 the curve was steady as the other labs.

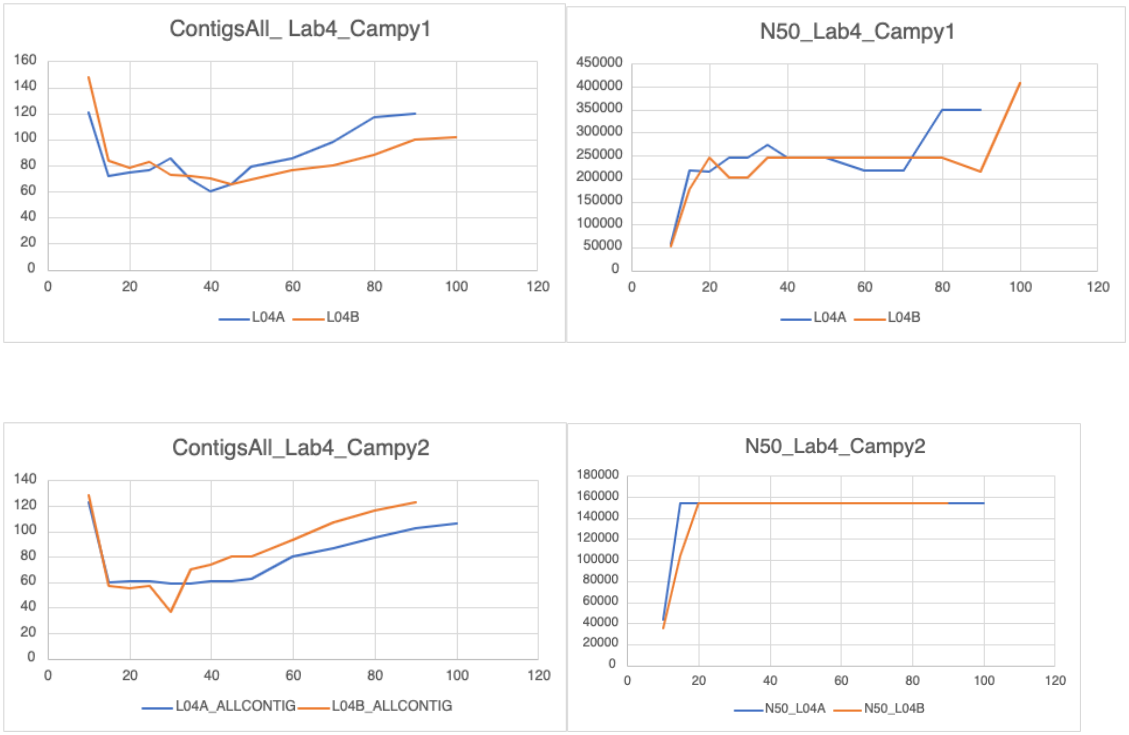


***Figure 13**. The variation of total number of contigs and N50 values for different amount of data (coverage).*

For Lab5:

Lab5 used the same instrument Illumina MiSeq as the Lab4 for both replicates but changed the kit. The Illumina NEBNext Ultra II library kit and NEBNext Multiplex Oligos library kit was used by Lab5. The data was 2x251 bp and the run gave 347 and 319 Mbases (214 and 197X coverage, respectively) for Campy1 and 301 and 287 Mbases (175 and 167X coverage, respectively) for Campy2 duplicates. Unlike the pervious labs, a different kit was used in Lab5 and the error flag in “Per base sequence content” did not show as it did for the pervious labs. But the very last base in the reads still showed a deviating base composition as the pervious labs (Figure 14), indicating that the error flag was caused by the Nextera kit.

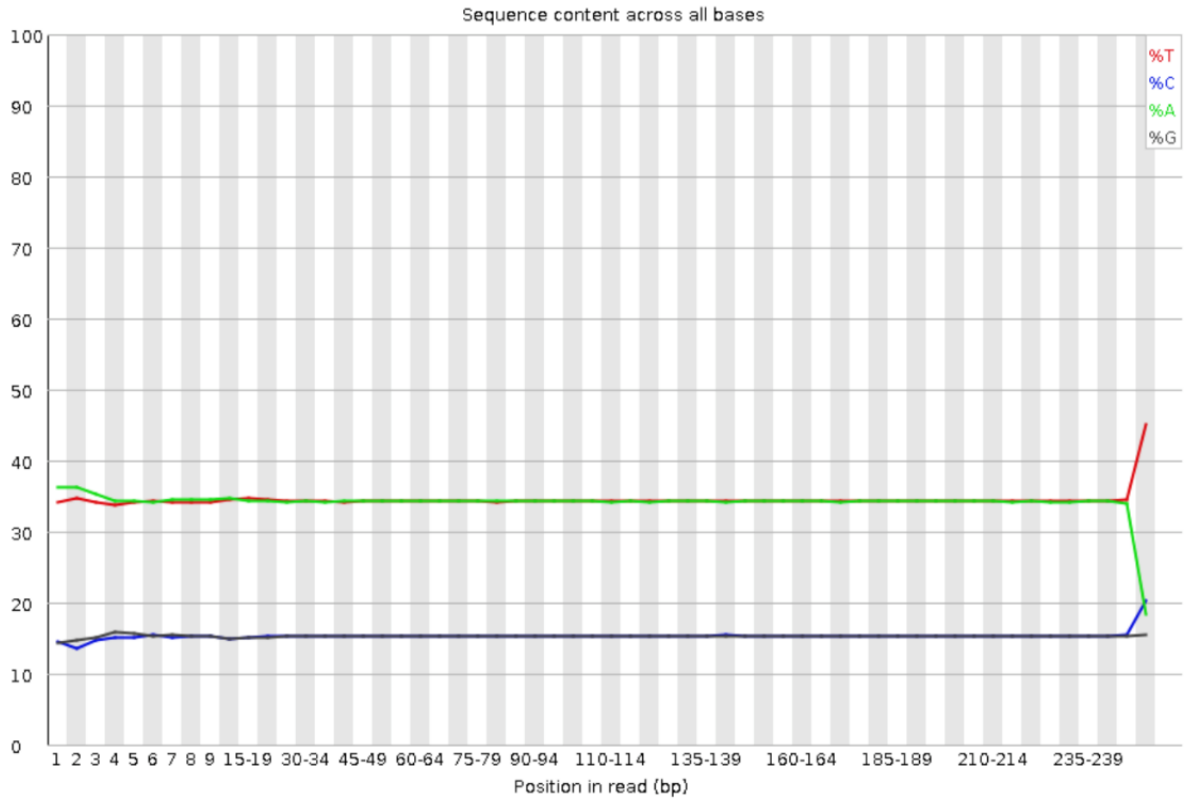


***Figure 14**. The "Per base sequence content" showed an abnormal base composition for the very last base.*

As the pervious labs, high quality along the whole reads was seen in the "Per base sequence quality" (Figure 15).

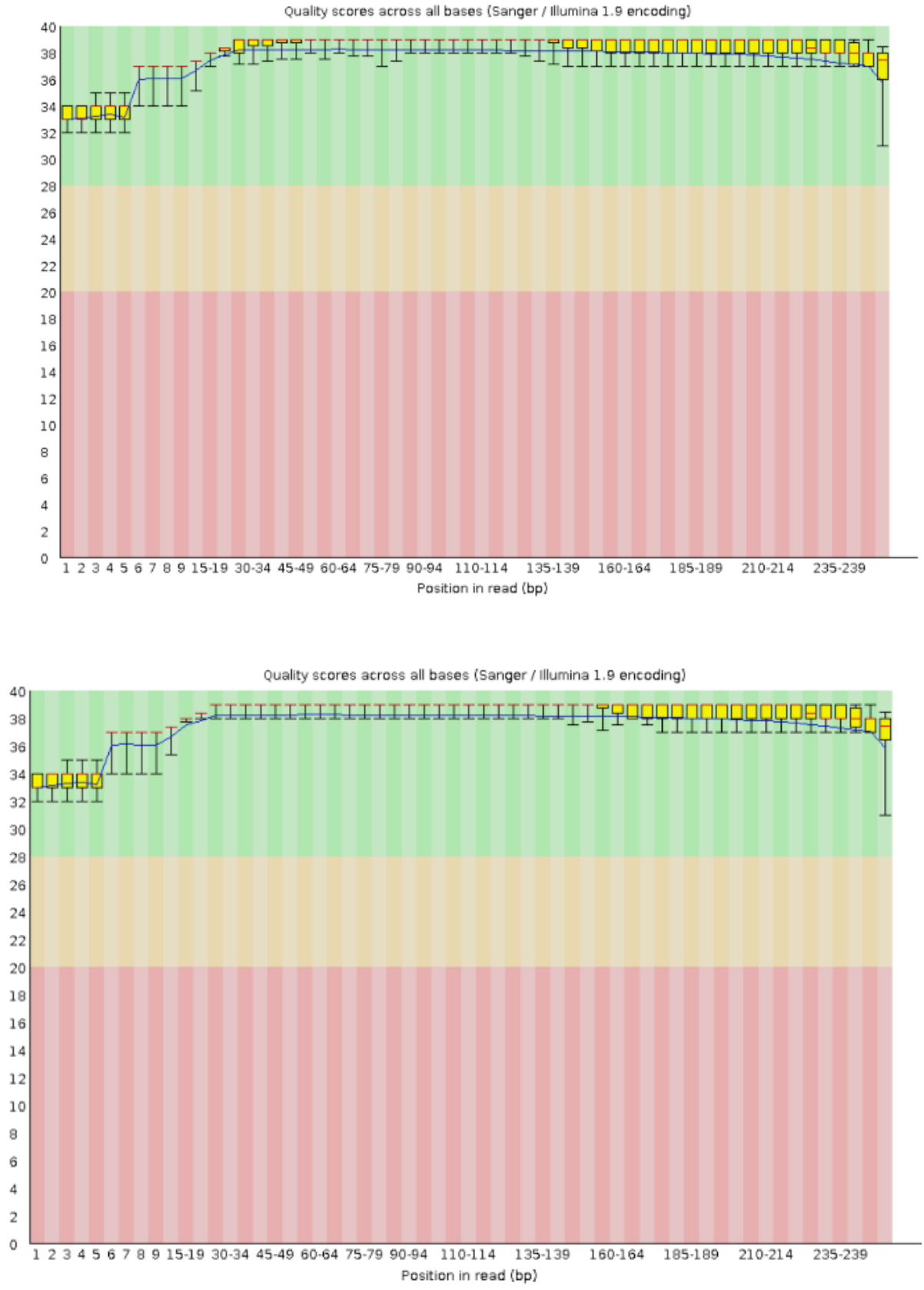


***Figure 15**. The "Per base sequence quality" of the L05A, L05B (MiSeq) data for strain Campy1.*

In Krona plot no contamination was visible as the pervious labs in the taxonomic classification with Kraken2. L05 had the highest classification with *C.jejuni* compared to the other labs which means less contamination with other species (Figure 16).

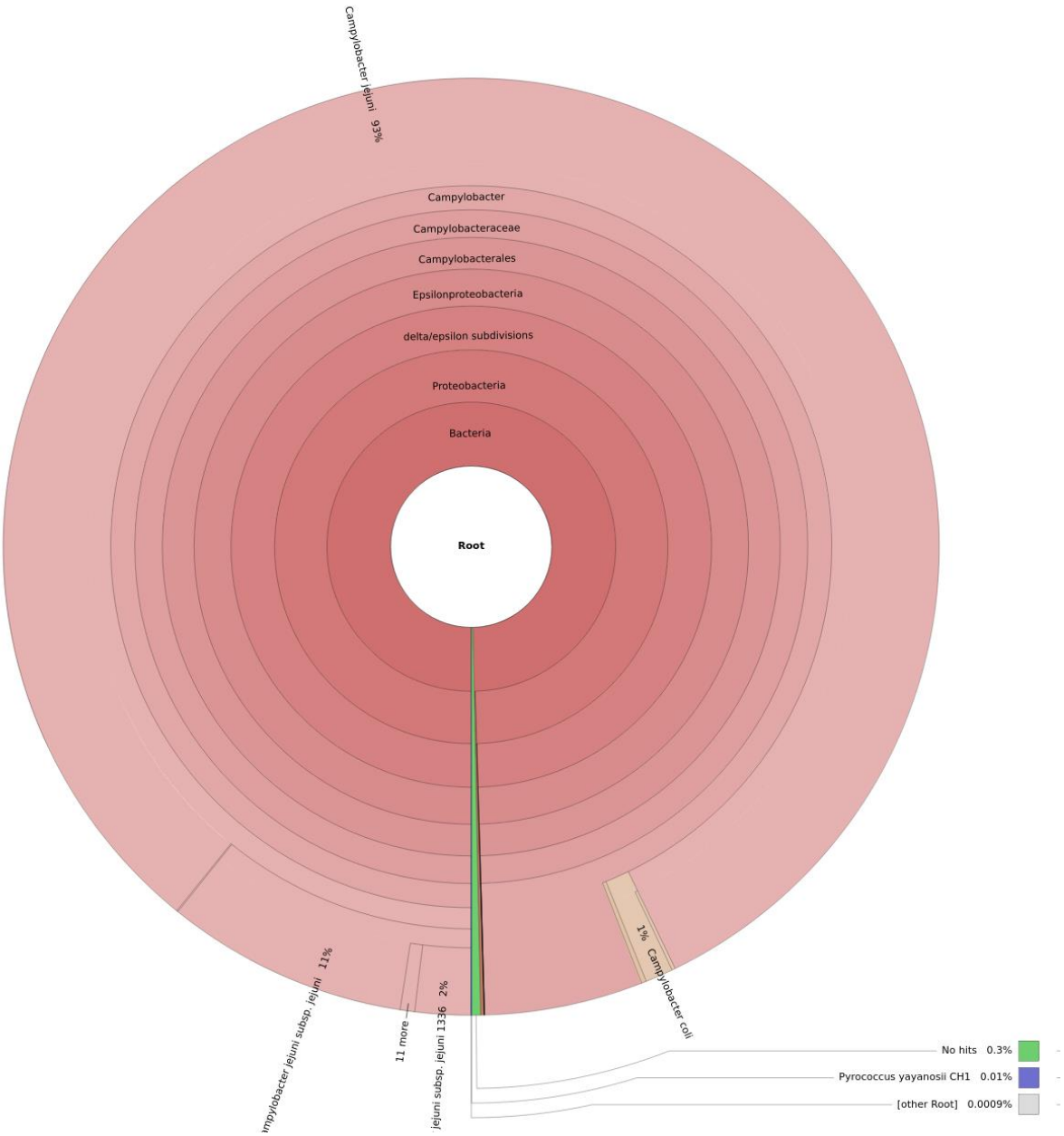


Figure 16*. Example (Strain Campy1, sample L05A) of a Krona plot of the Kraken2 taxonomic classification using the Minikraken database. The Krona plot showed very minor proportions of the reads (less than 2%) being classified as other taxa than Campylobacter.*

The data from Lab5 was then assembled with SPAdes using a titration of the amount of data used (coverage) (Figure 17). The MiSeq produced fewer contigs for L05A and larger N50 for L05A than L05B. The reason for this is unknown. There was a difference in the L05A and L05B: the curve of the L05A jumped up and down along the "coverage axis", while the curve of the L05B increased rapidly, then was steady and increased slightly again for the strain Campy1. For the strain Campy2 both replicates showed

"up and down"-trend, thereafter becoming steady until the end. This indicates that there was an optimal coverage, but the optimum was not always the same.

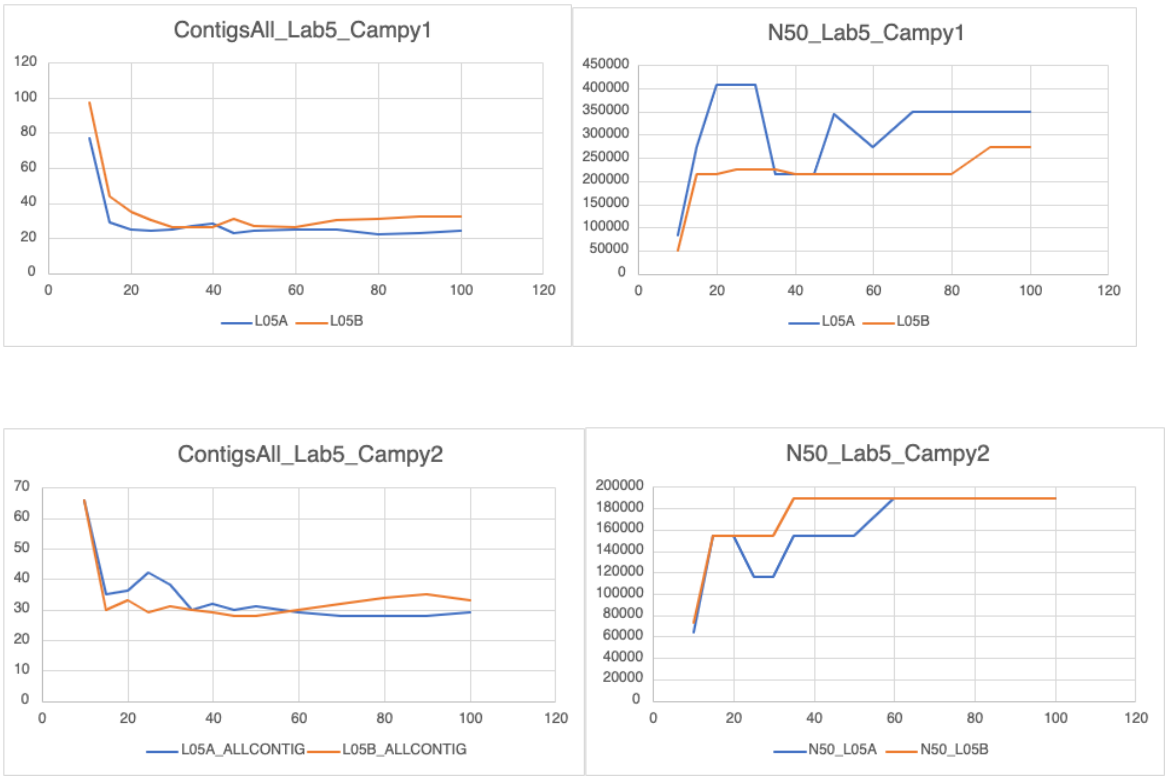


***Figure 17**. Change in total number of contigs and N50 values when amount of data used (coverage) was varied.*

For Lab6:

For Lab6, the instrument that was used was MiSeq as Labs 4 and 5 for both replicates and the kit was Nextera DNA flex, Nextera XT Index. The resulted data was 2x301 bp and the run gave 904 Mbases (558X coverage) for Campy1 and 808 Mbases (470X coverage) for Campy2. Both the error flag in "Per base sequence content" and the deviation base composition of the very last reads were shown in the Lab6 (Figure 18).

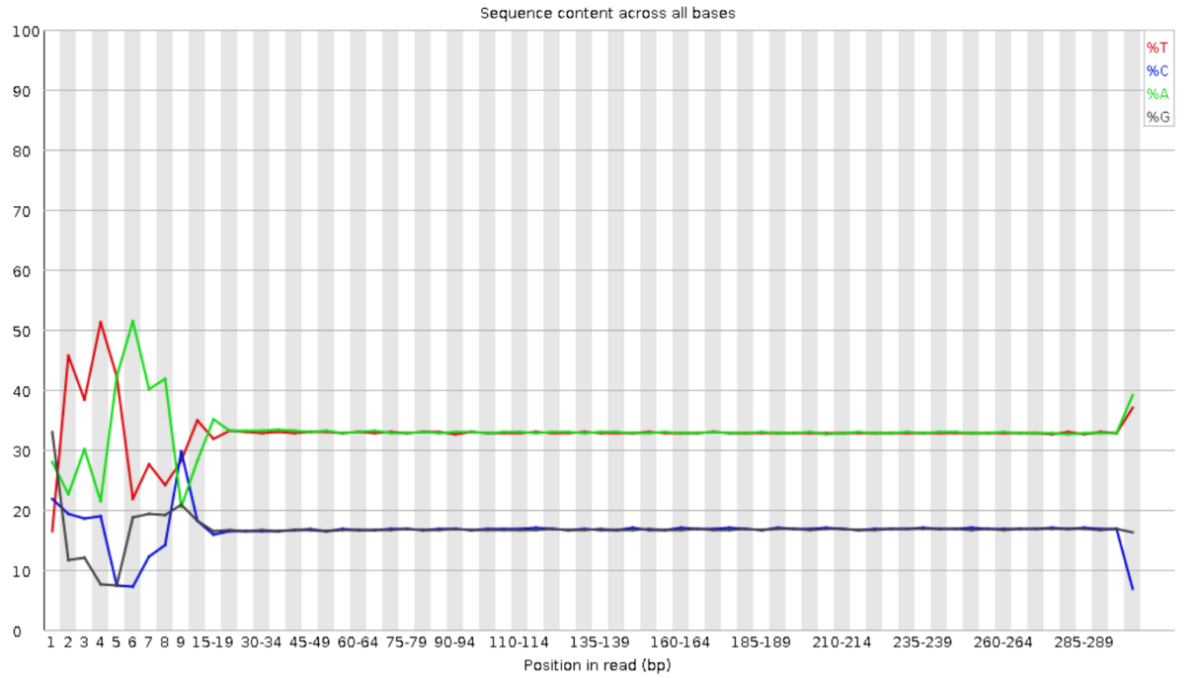


***Figure 18**. The "Per base sequence content". Similar to Labs 2, 3 and 4.*

There was a slightly reduced sequence quality towards the ends of the sequences (Figure 19). We removed the low-quality sequence ends using Trimmomatic tool (Figure 20).

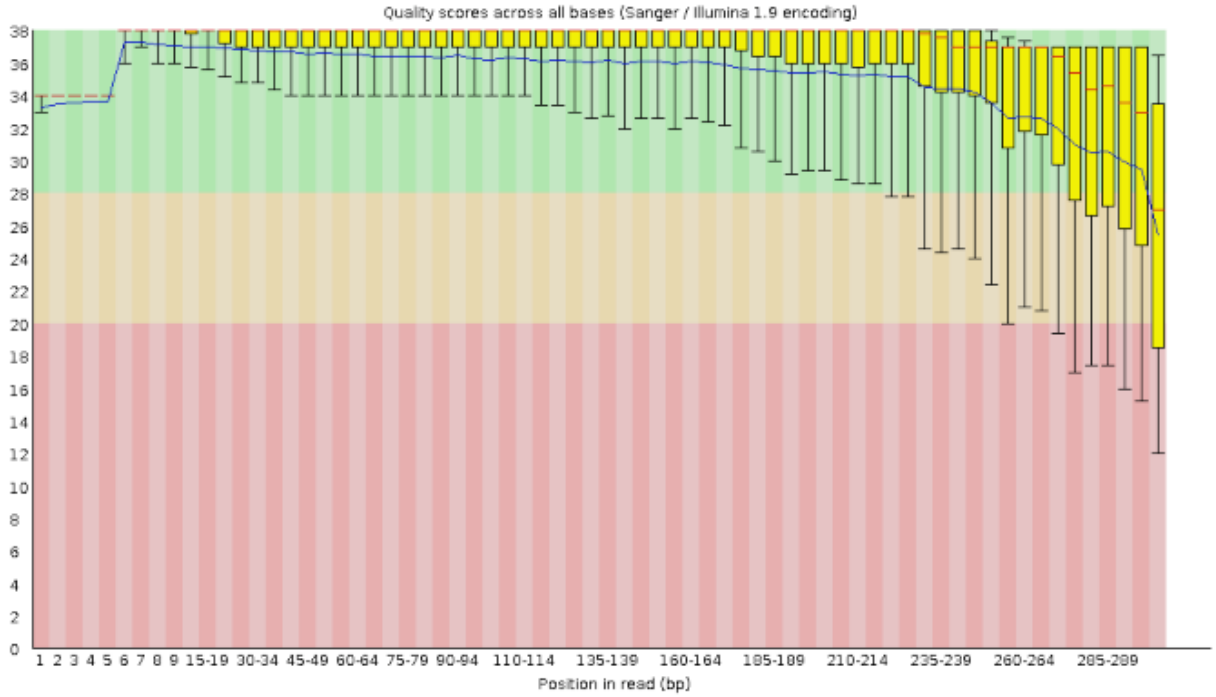


Figure 19. *The "Per base sequence quality" of the L06A, L06B (MiSeq) data for strain Campy1 before trimming. Drop in the quality towards the end.*

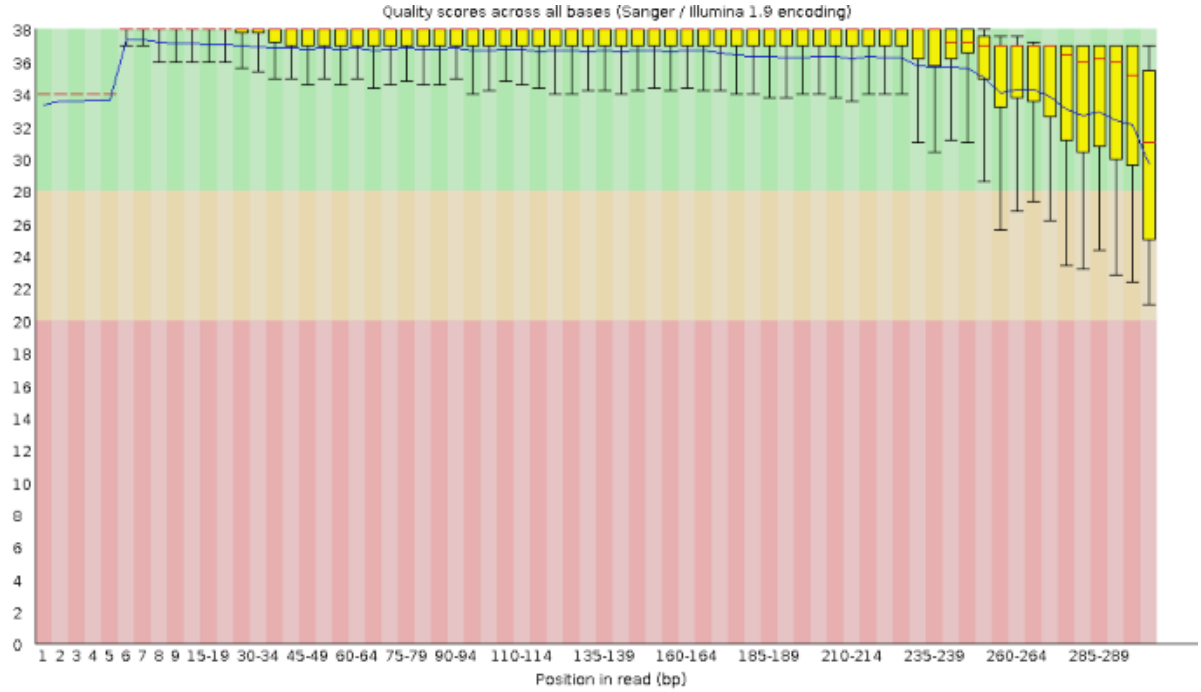


Figure 20*. The "Per base sequence quality" of the L06A, L06B (MiSeq) data for strain Campy1 after trimming.*

Unlike the pervious labs the "Per tile sequence quality" showed worse qualities than other tiles for the base compared to previous labs (Figure 21).

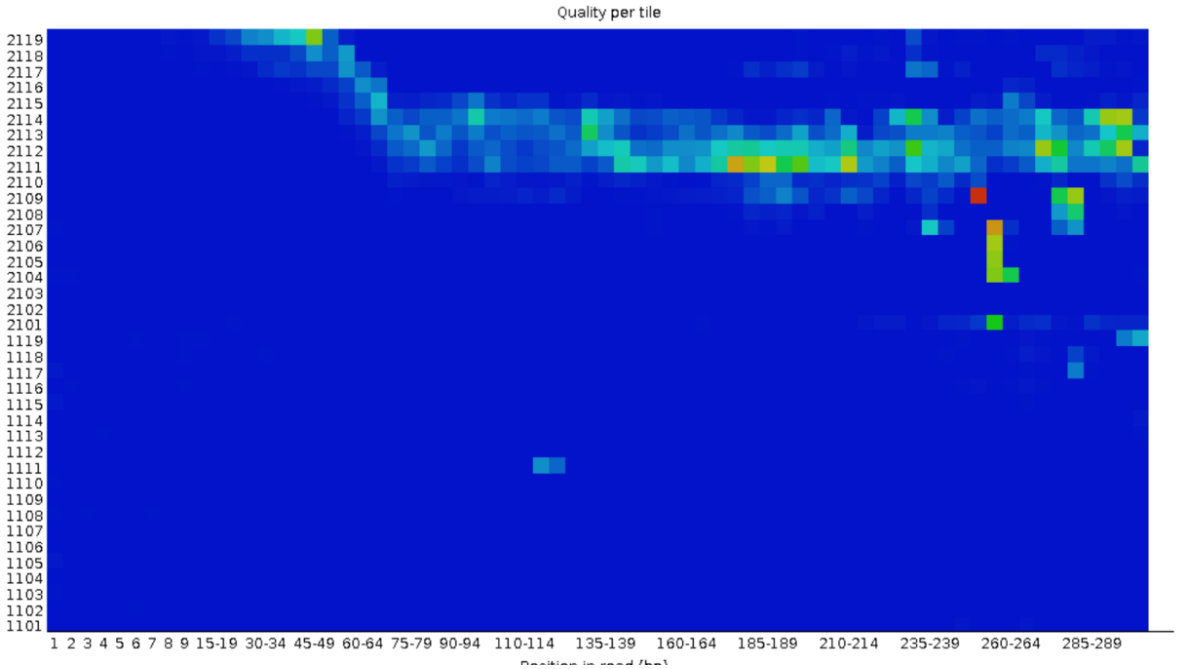


Figure 21*. The "Per tile sequence quality" of the L06A data for strain Campy1.*

No contamination was seen in Krona plot as the other labs. L06 had the least reads that classified as *C.jejuni* compared to the other labs and thus highest contamination of other species than the other labs (Figure 21).

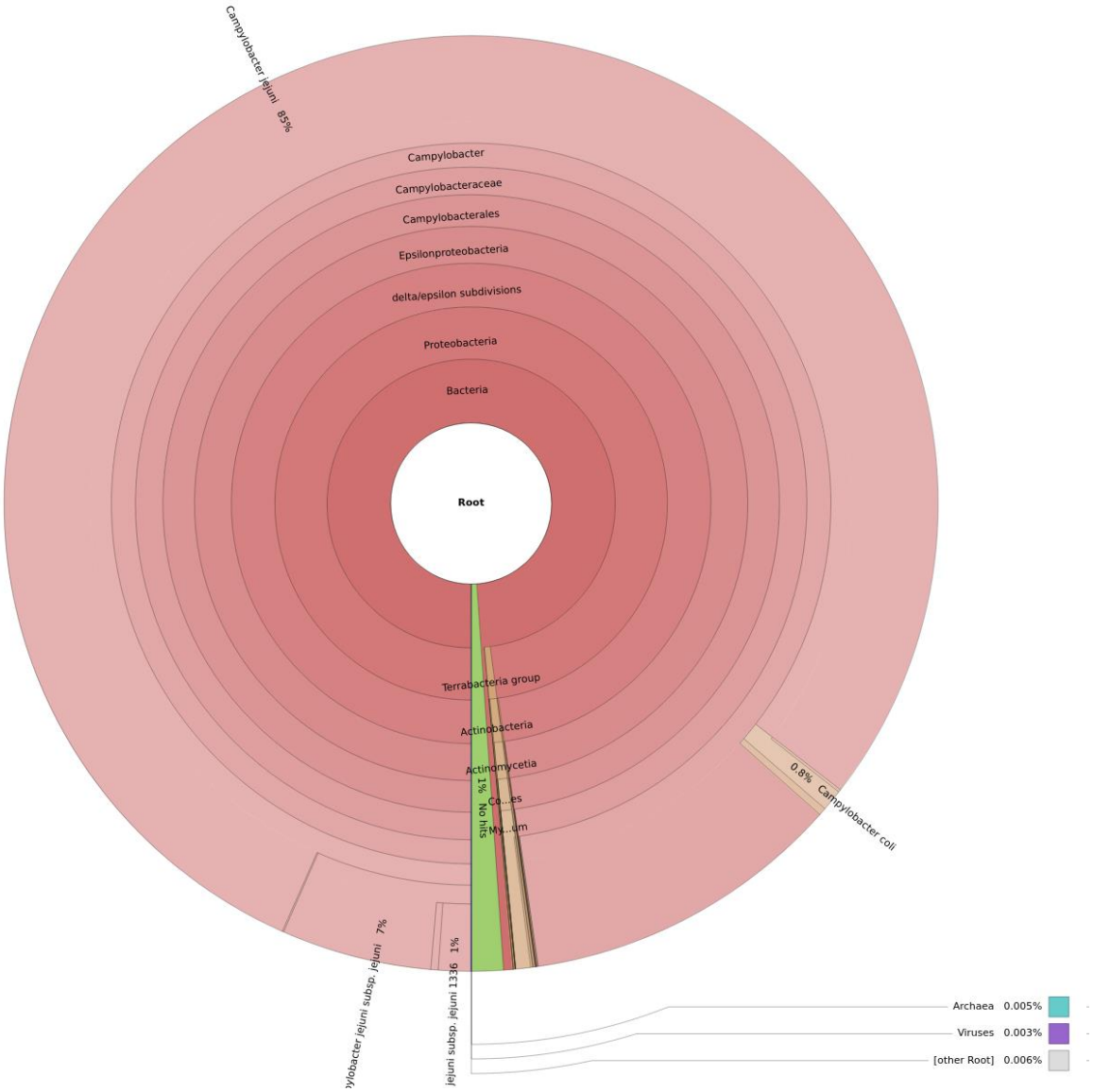


Figure 21. *(Strain Campy1, sample L06A). Very minor proportions of the reads were classified as other taxa than Campylobacter.*

For both strains Campy1 and Campy2, nearly the same amount of contigs and the same N50 for both replicates were given from the assembled data with SPAdes. The number of contigs were almost optimal but raised slightly at the highest coverage used. The N50 kept increasing towards the end (Figure 22). However, the assembly of this lab was worse than the previous labs.

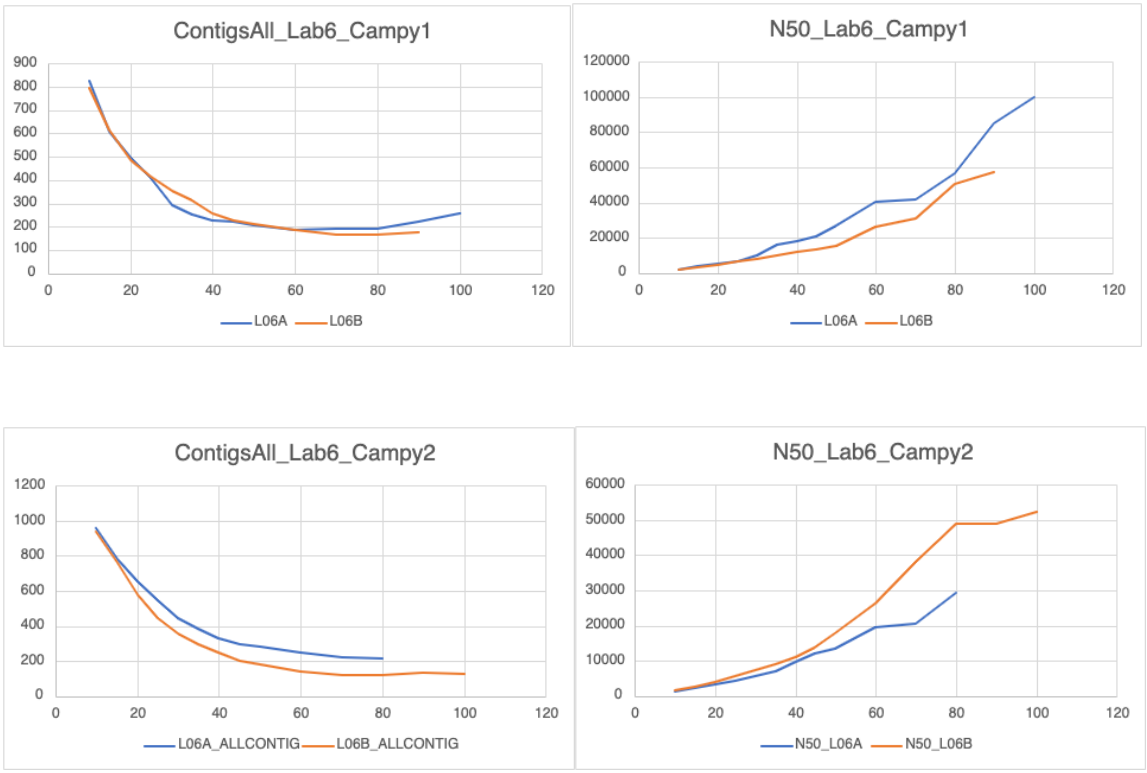


Figure 22*. The contigs and N50 values with different coverage for the Lab6 (L06A and L06B).*

For Lab7:

MiSeq instrument and TruSeq Nano library kit were used for both replicates (i.e., the only lab that used TruSeq library preparation kit). The data was 2x251 bp and the run gave 904 and 567 Mbases (558 and 350X coverage, respectively) for Campy1 and 808 and 858 Mbases (470 and 499X coverage, respectively) for Campy2 duplicates. No error flag was seen in the Lab7, but the deviation in base composition of the very last base was seen in the Lab7 (Figure 23).

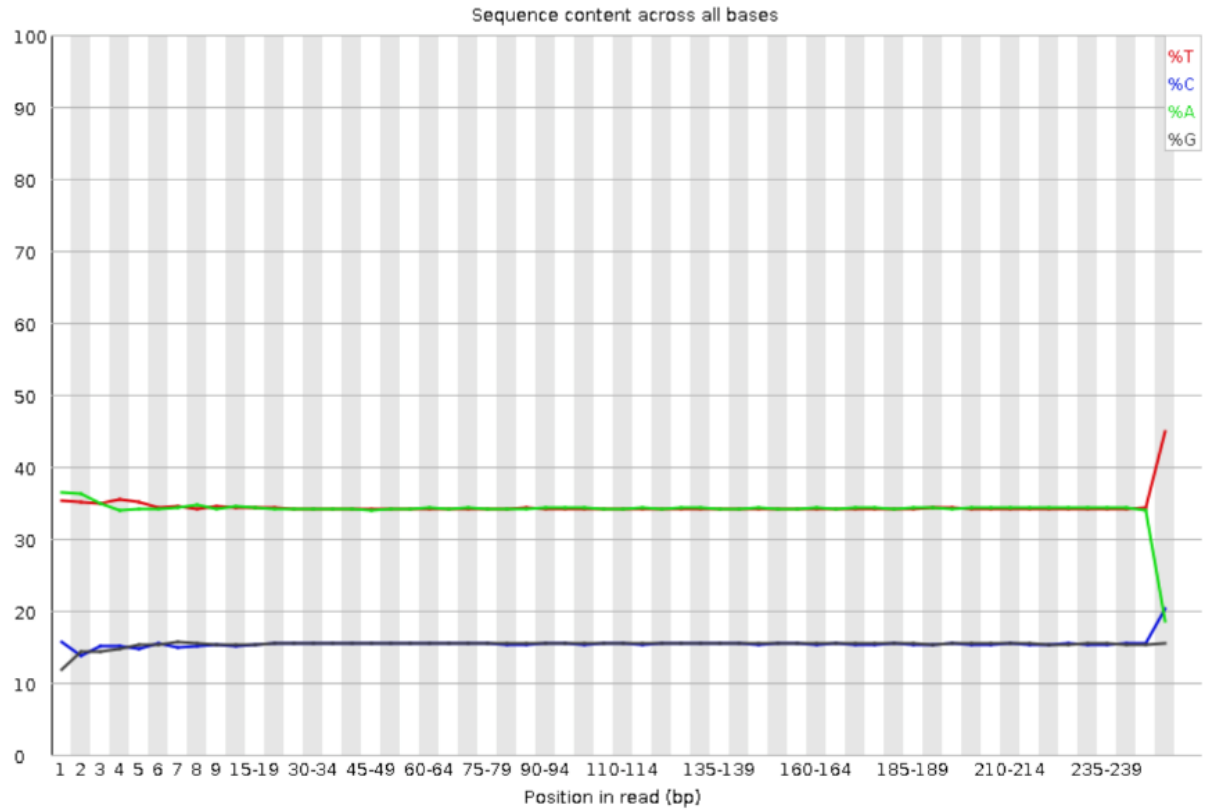
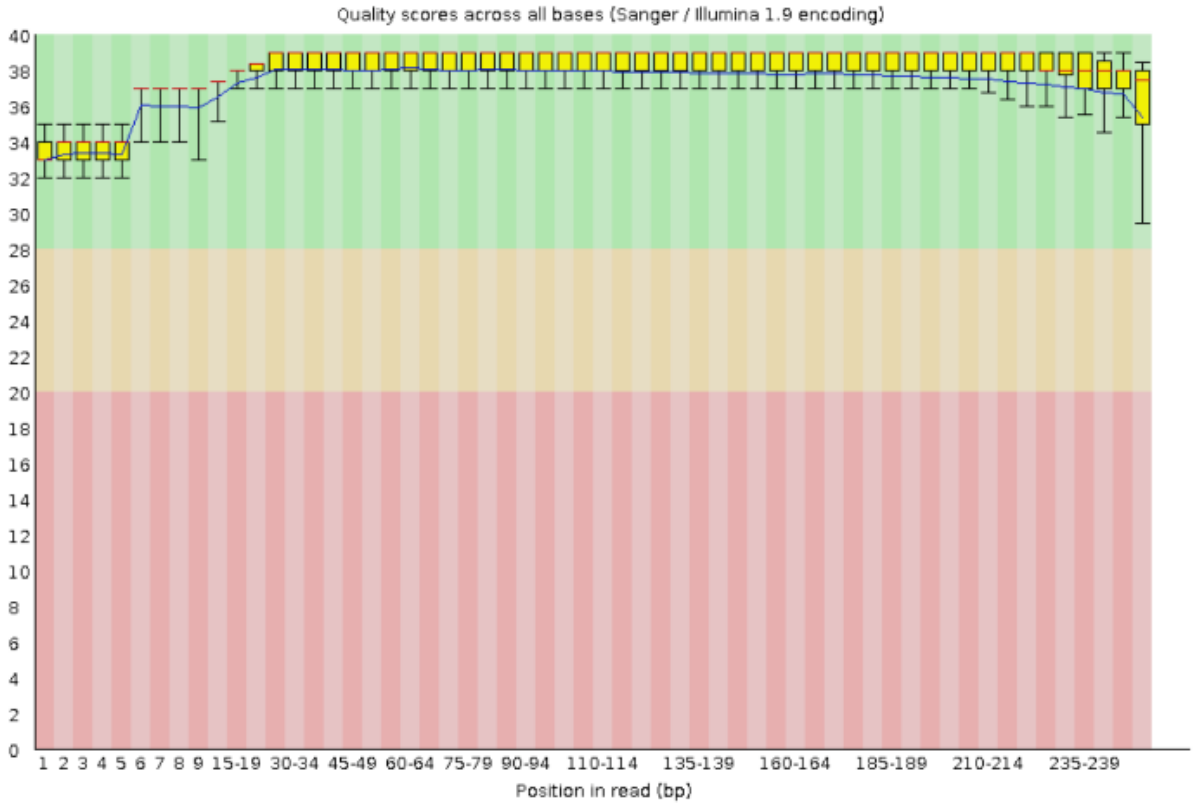


***Figure 23**. The "Per base sequence content" similar to the Lab5.*

“Per base sequence quality” showed high quality as the other labs and similar to Lab5 (Figure 24).



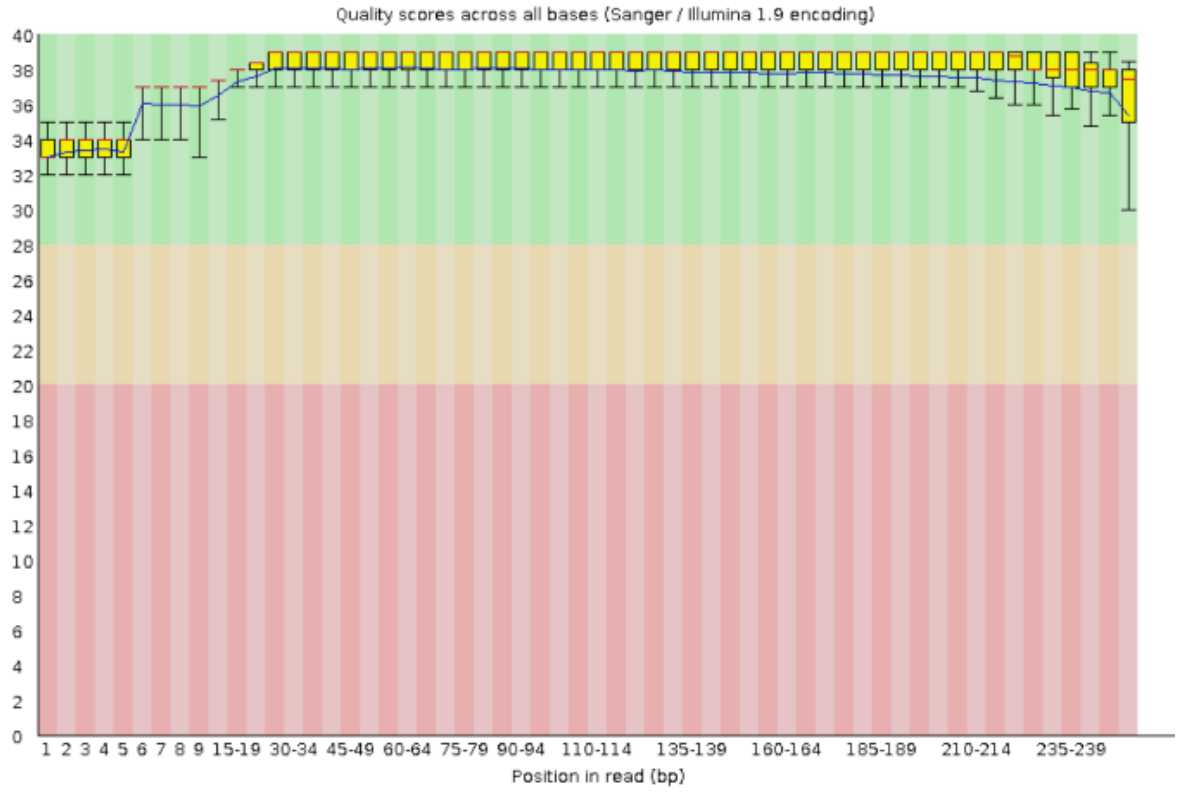


Figure 24. *The "Per base sequence quality" of the L07A, L07B (MiSeq) data for strain Campy1.*

Krona plot showed no contamination as pervious labs (Figure 25).

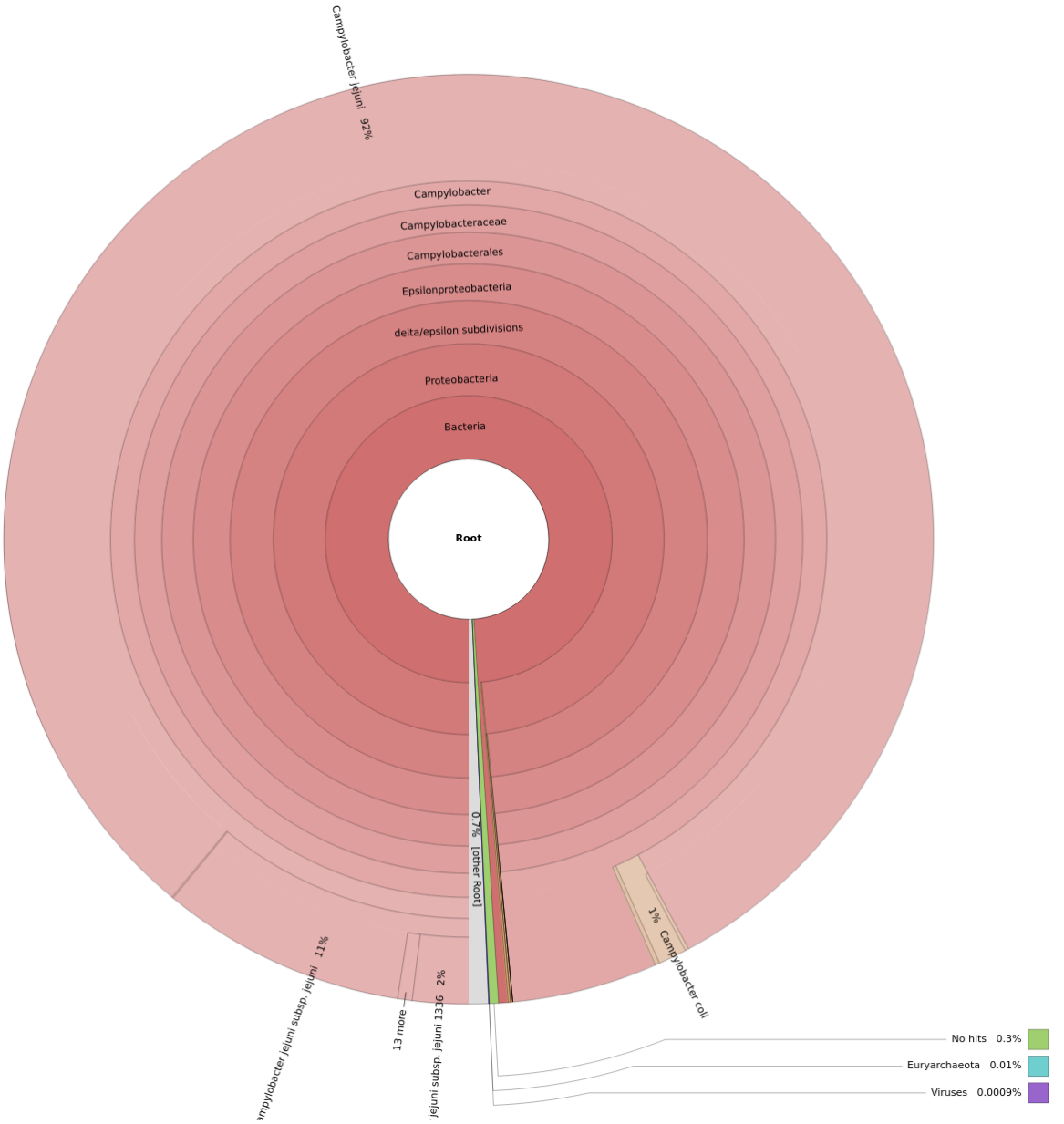


Figure 25. *No more than 2% of the reads were classified as other taxa for the Lab7. Example L07A strain Campy1.*

Assembled data using SPAdes for both strains Campy1 and Campy2 showed that the contigs started with a slight drop at the beginning to the coverage limit of 20, then increased radically until the end. N50 showed a significant increase at first and then stabilized with a slight decrease for the L07A. N50 for the L07B was the same case as for L07A except that it showed "up and down"-pattern along the curve (Figure 26).

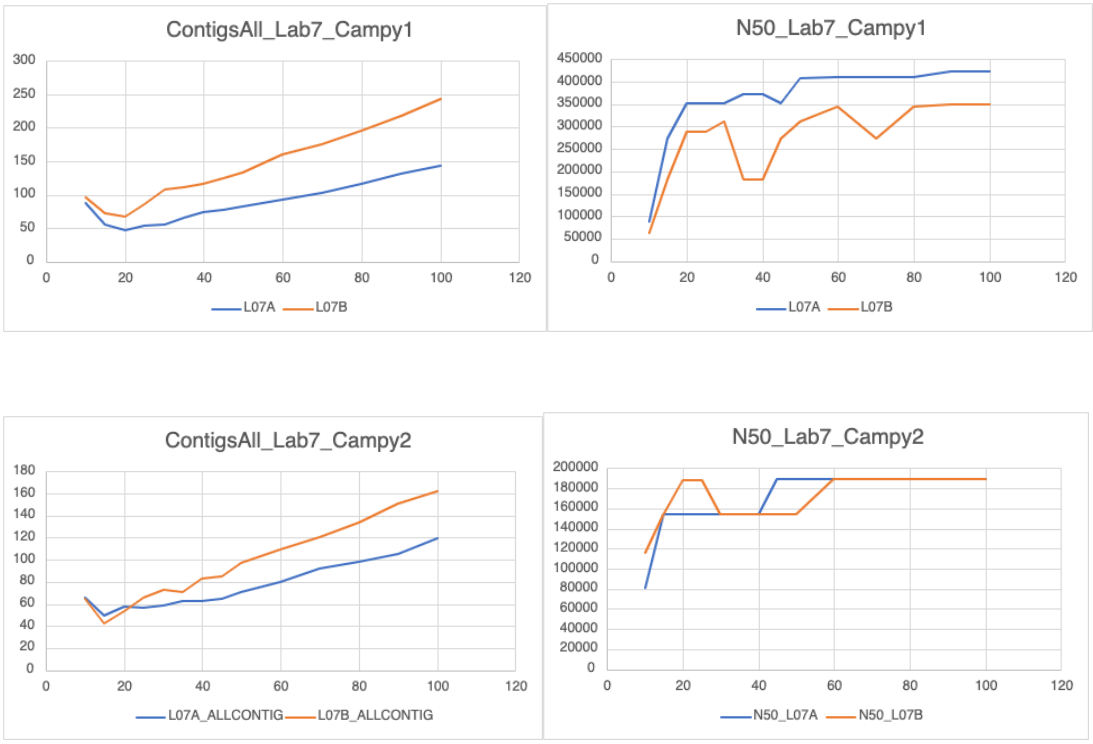


Figure 26*. Shows the difference in data of contigs and N50 values in different coverages for Lab7 (L07A, L07B).*

For Lab8:

No replicate was used in the Lab8 with the instrument MiSeq and Nextera XT library kit. The data was 2x301 bp and gave 182 Mbases (113X coverage) for Campy1 and 468 Mbases (273X coverage) for Campy2. As for the pervious labs when using the Nextera kit, both deviation of base composition and the error flag were seen in the Lab8 (Figure 27).

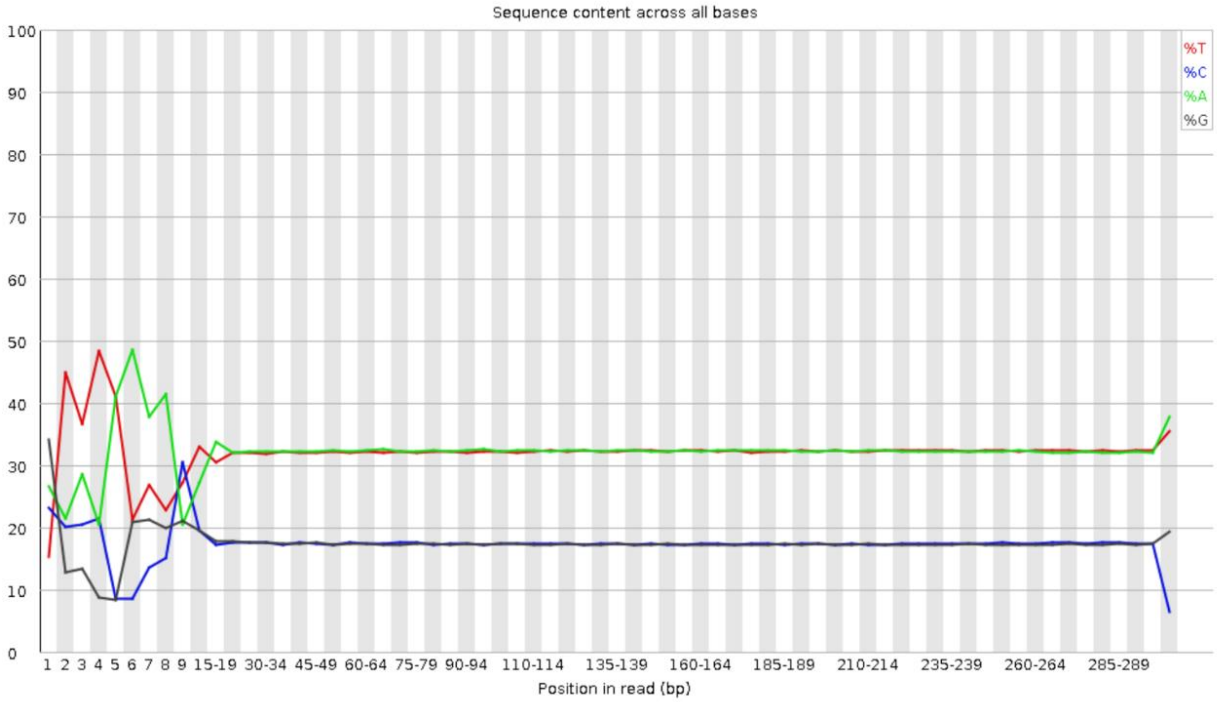


***Figure 27**. The "Per base sequence content" similar as the pervious labs with the use of Nextera kit.*

The sequence quality was determined to reduce towards the ends of the sequences (Figure 28).

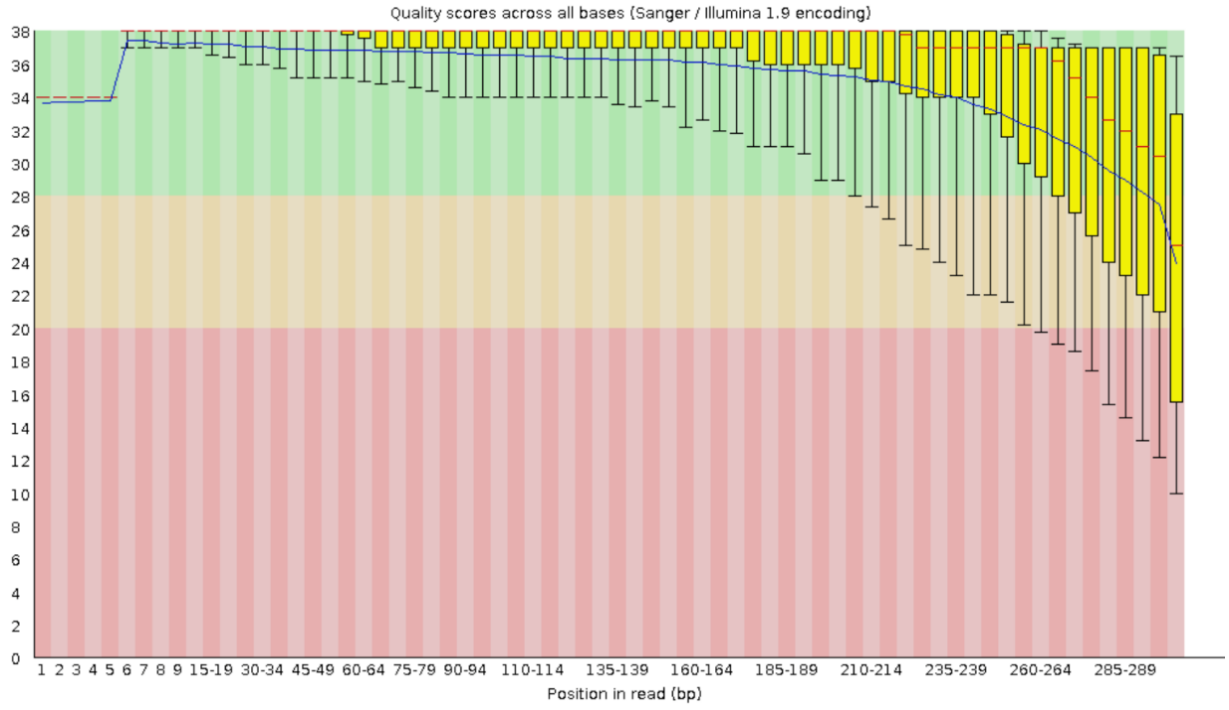


***Figure 28**. The "Per base sequence quality" for the Lab8 for strain Campy1.*

SPAdes data for both strains Campy1 and Campy2 showed that the contigs had an “up and down” jump along the coverage. N50 kept increasing and then stabilized at the coverage of 80 (Figure 29). Lab8 produced the worst assembly compared to the other labs, except for Lab6.

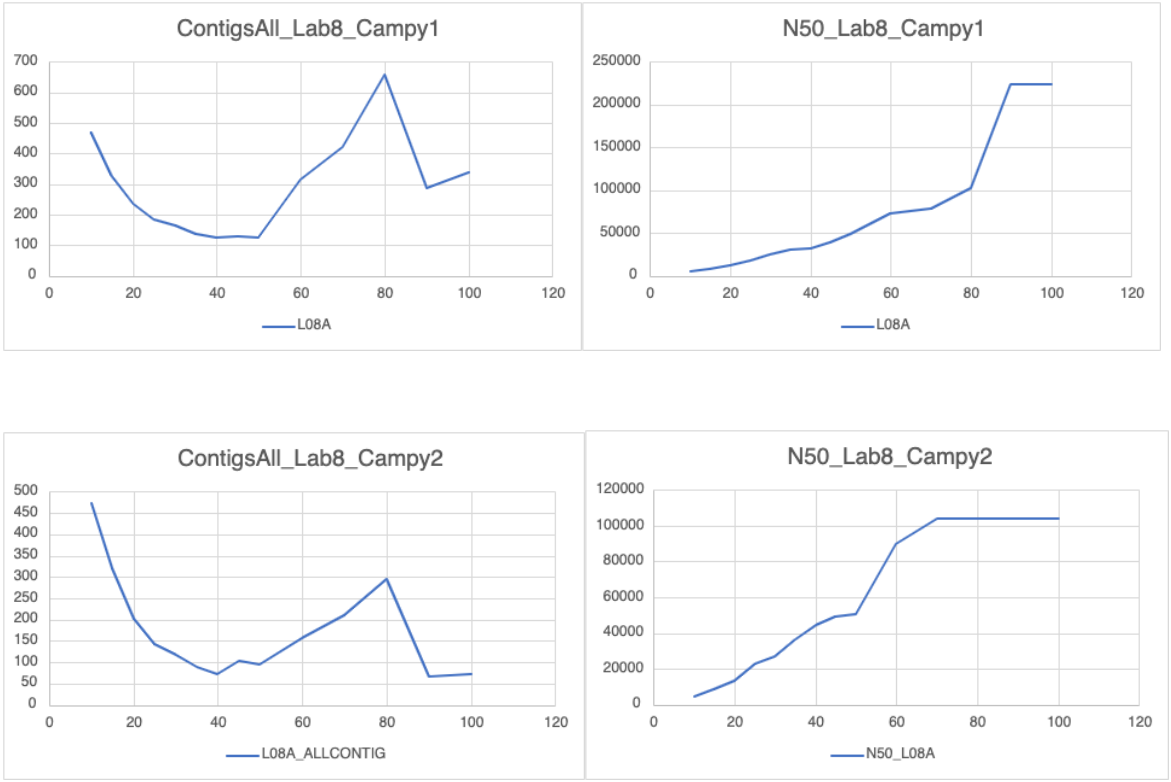


Figure 29*. The variants in data of contigs and N50 values in different coverages for Lab8.*

As for the pervious labs the taxonomic classification with Kraken2 showed no contaminations (Figure 30).

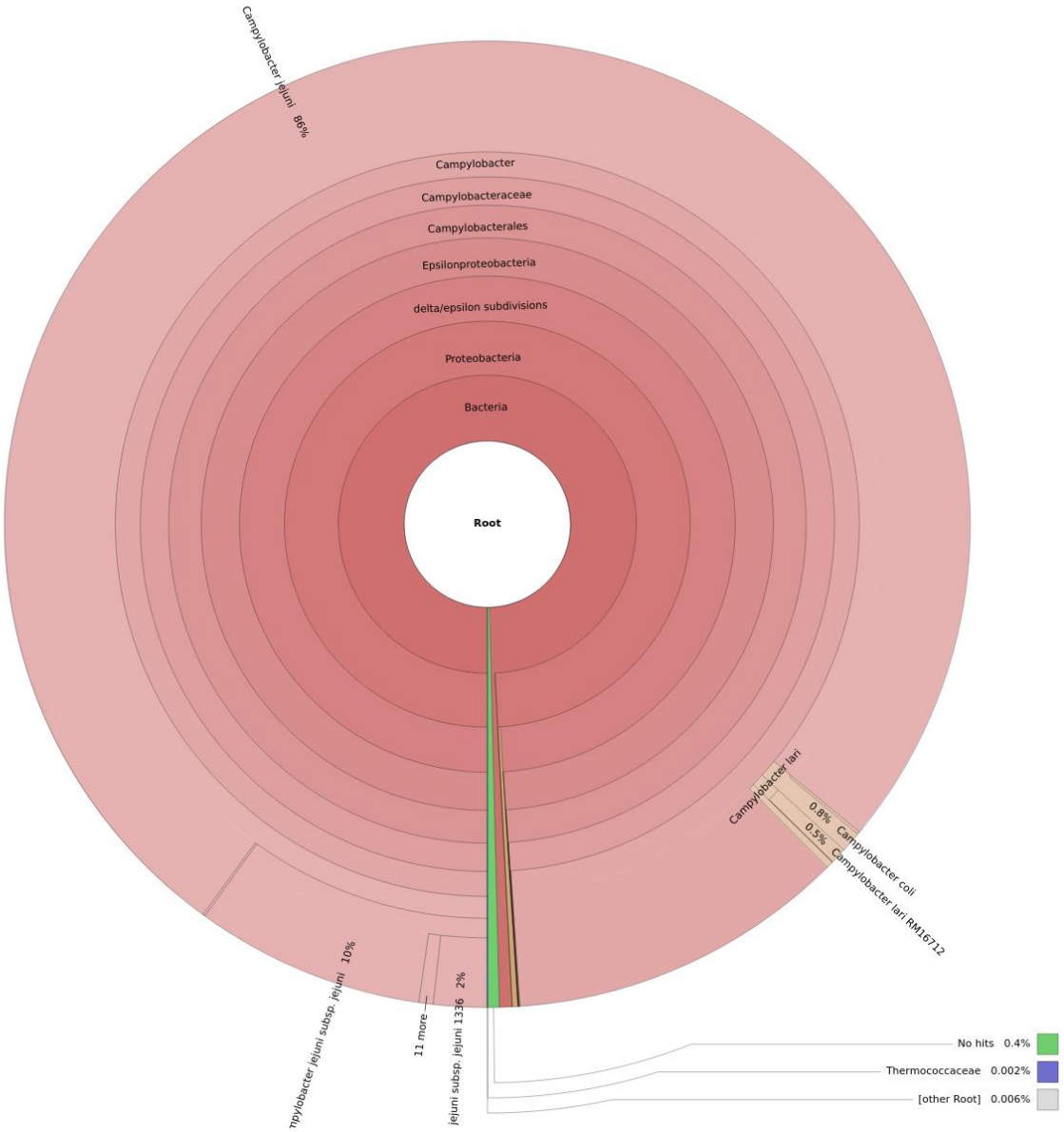


Figure 30*. (Strain Campy1, sample L08A). About 2% of the reads were classified as other taxa than Campylobacter.*

For Lab9:

Lab9 sequenced two replicates with the same instruments, Illumina NextSeq 500 and NEBNext Ultra II FS Library kit for both replicates. The read length was 2x151 bp. For the Campy1 the run gave 350 and 285 Mbases (216 and 166X coverage, respectively). For the Campy2 the run gave 330 and 376 Mbases (204 and 219X coverage, respectively). FastQC showed the error flags in “Per base sequence quality” for all fastq-files coming from L09 (Campy1 and Campy2), however L09 passed the “Per base sequence content” for the strain Campy2 despite the error flags. There was a smaller proportion of reads that showed deviation in base composition of the very last base compared to the other labs for the Campy1, but not for the strain Campy2. L09 lost its deviation after adapter trimming (Figure 31).

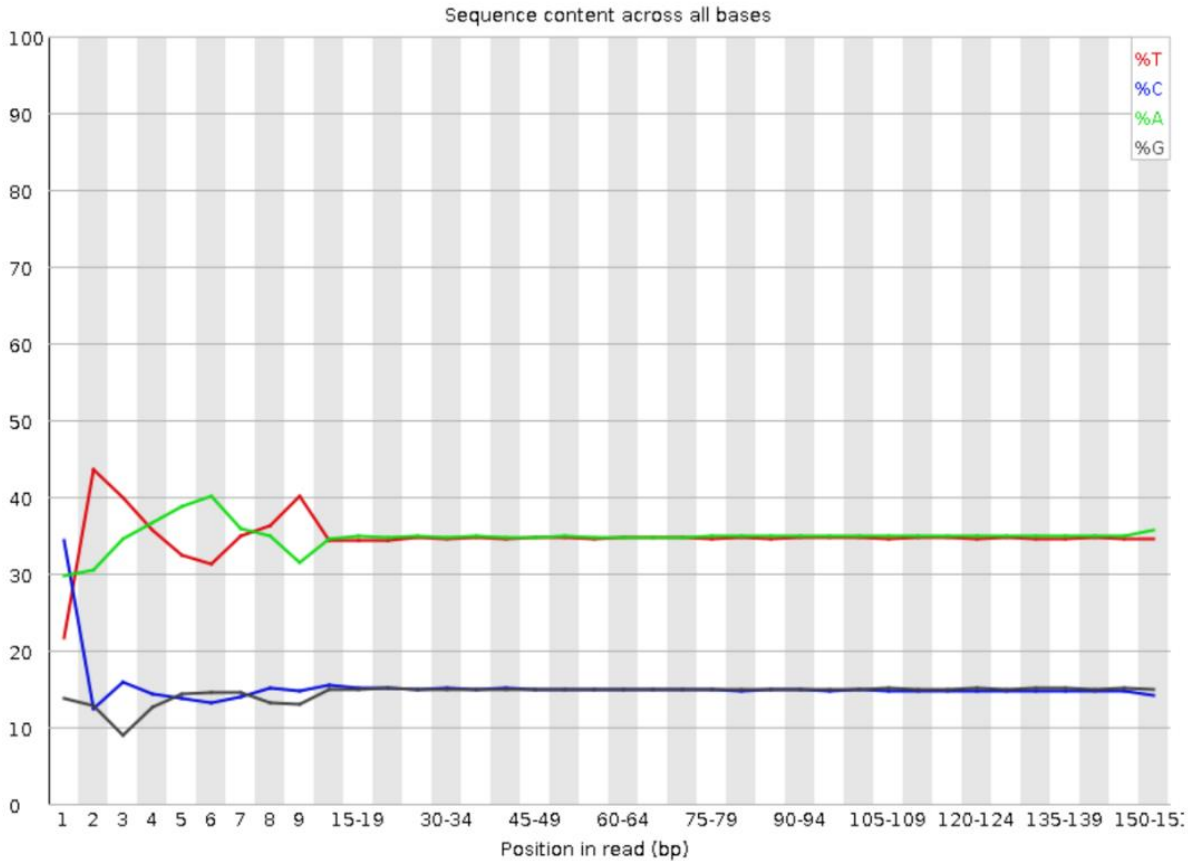


Figure 31. *The "Per base sequence content" showed the error flags at the first read sequences. The very last base did not show abnormal base composition as the other labs.*

High quality along the whole reads was seen in "Per base sequence quality "as the other labs (Figure 32).

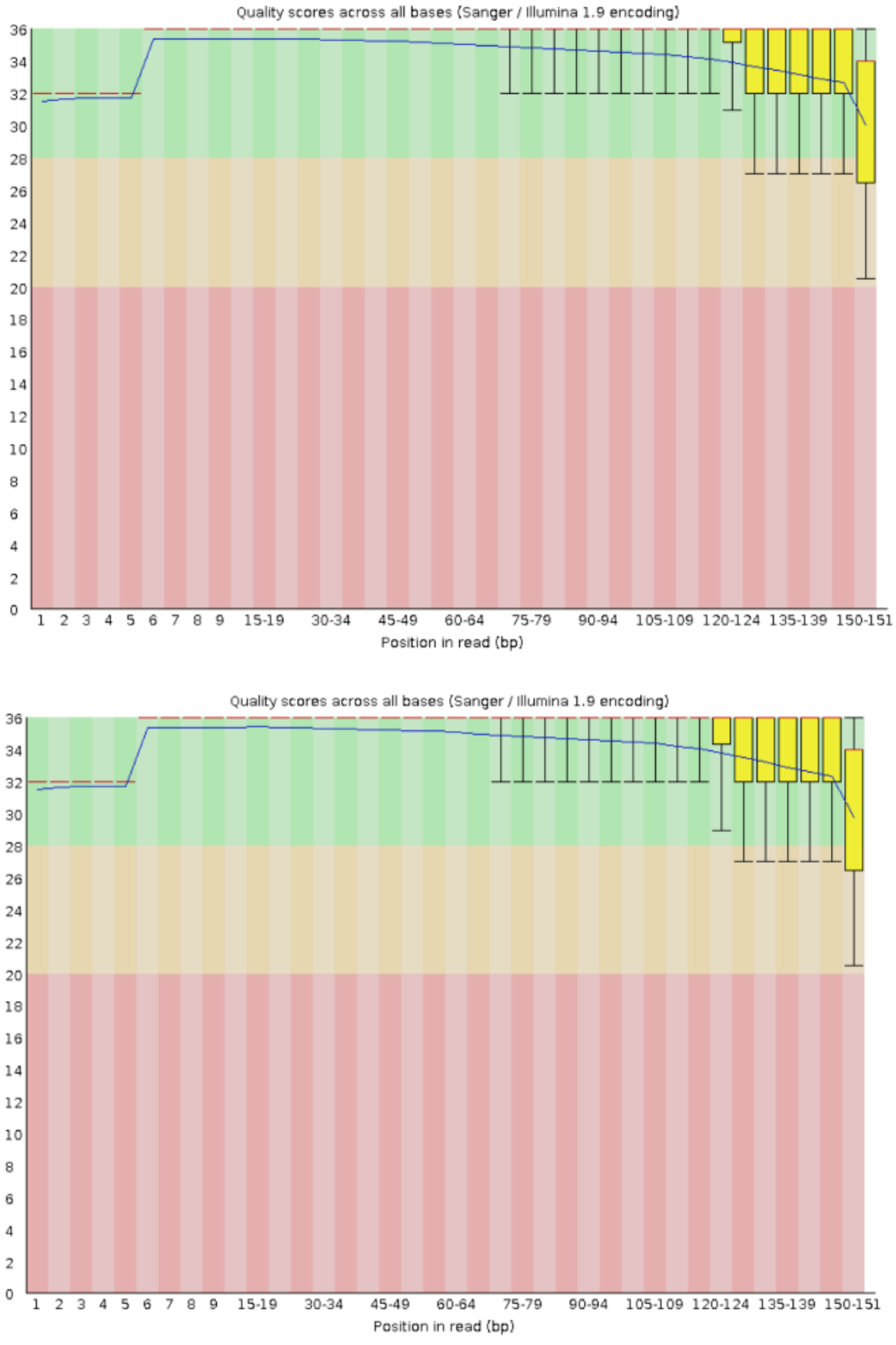
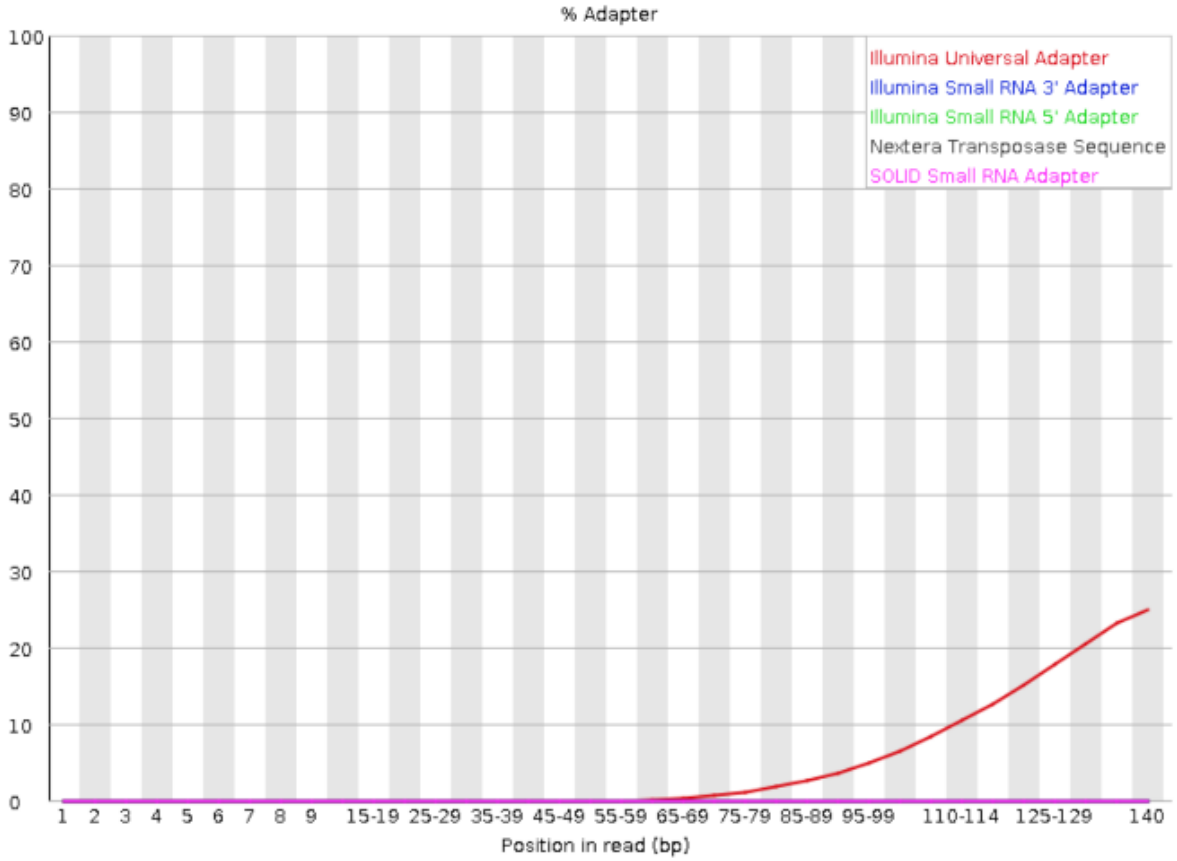


***Figure 32**. The "Per base sequence quality" for the L09A and L09B for strain Campy1.*

There was adapter sequence problematic with the L09 sample. Thus, we removed the adapter sequences to improve the quality of the sequence reads (Figure 33).



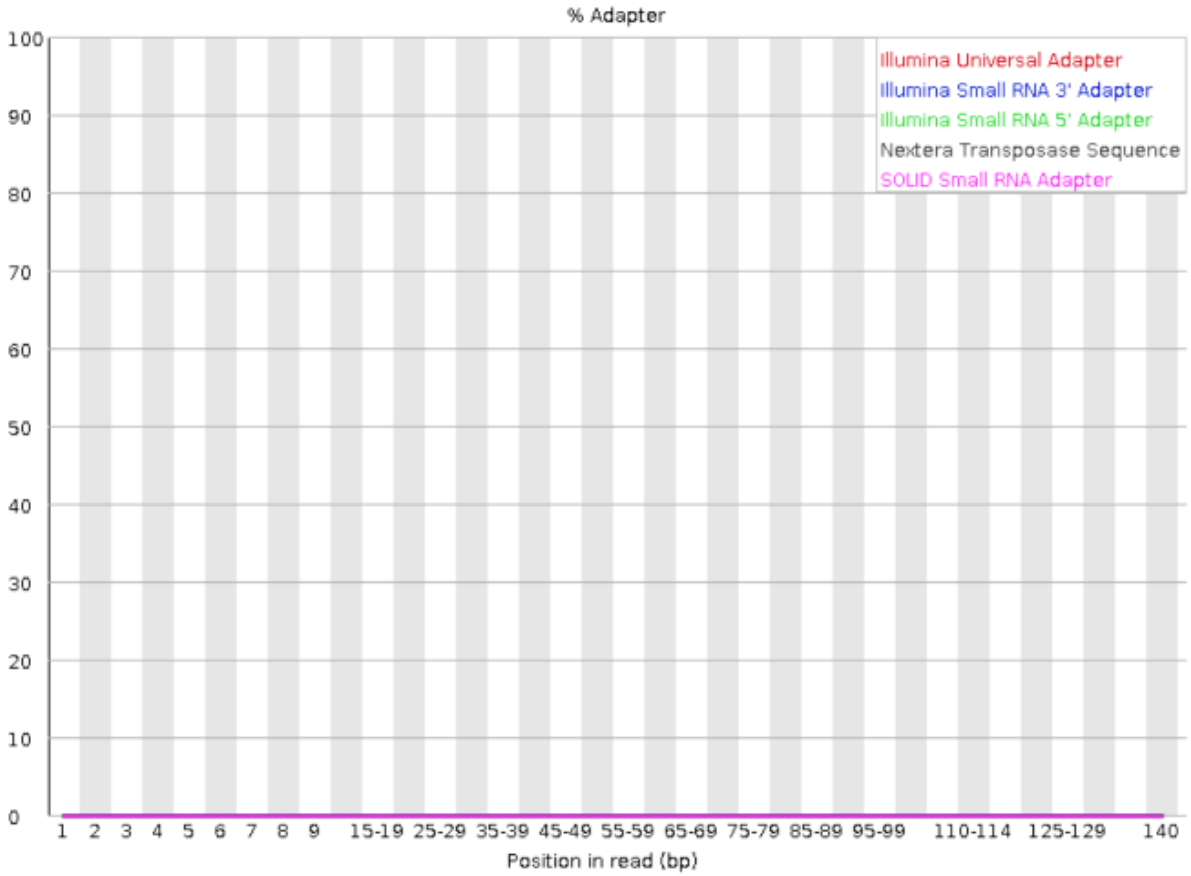


***Figure 33**. An example of how adapter sequences looked like before and after trimming in sample L09A.*

The taxonomic classification with Kraken2 looked similar as the other labs. No contamination of other species was detected (Figure 34).

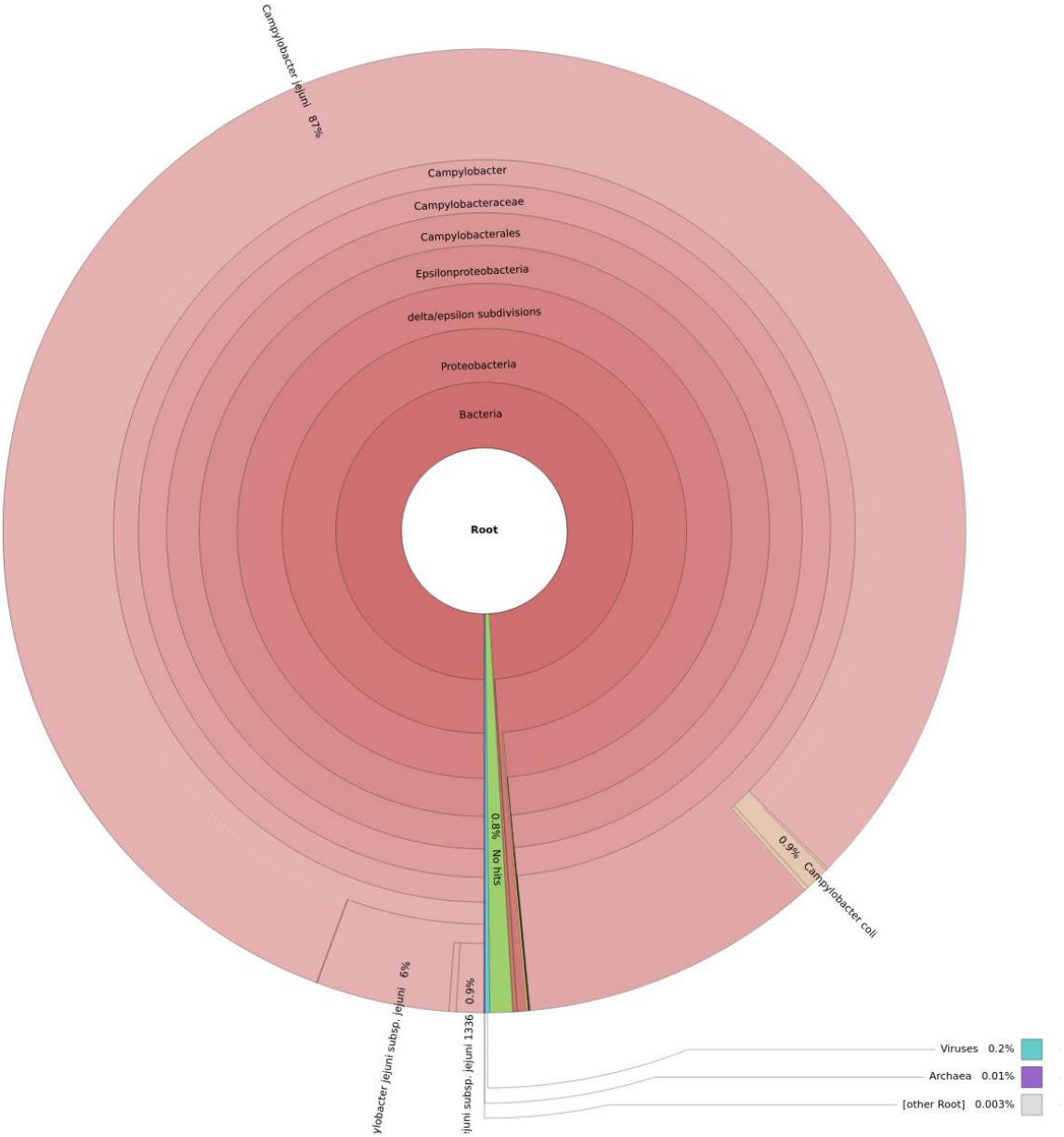


Figure 34*. (Strain Campy1, sample L09A). Very minor proportions of the reads were classified as other taxa than Campylobacter.*

Assembled data using SPAdes showed significant difference between the strain Campy1 and the strain Campy2. For the strain Campy1 the contigs showed dropping of the number of contigs at the beginning, then increased radically until the end for both replicates L09A and L09B. N50 started with a sharp increase, followed by an “up and down”-pattern along the curve. For the strain Campy2, there was an outlier in the middle of the curve for both contigs curve and N50 (Figure 35).

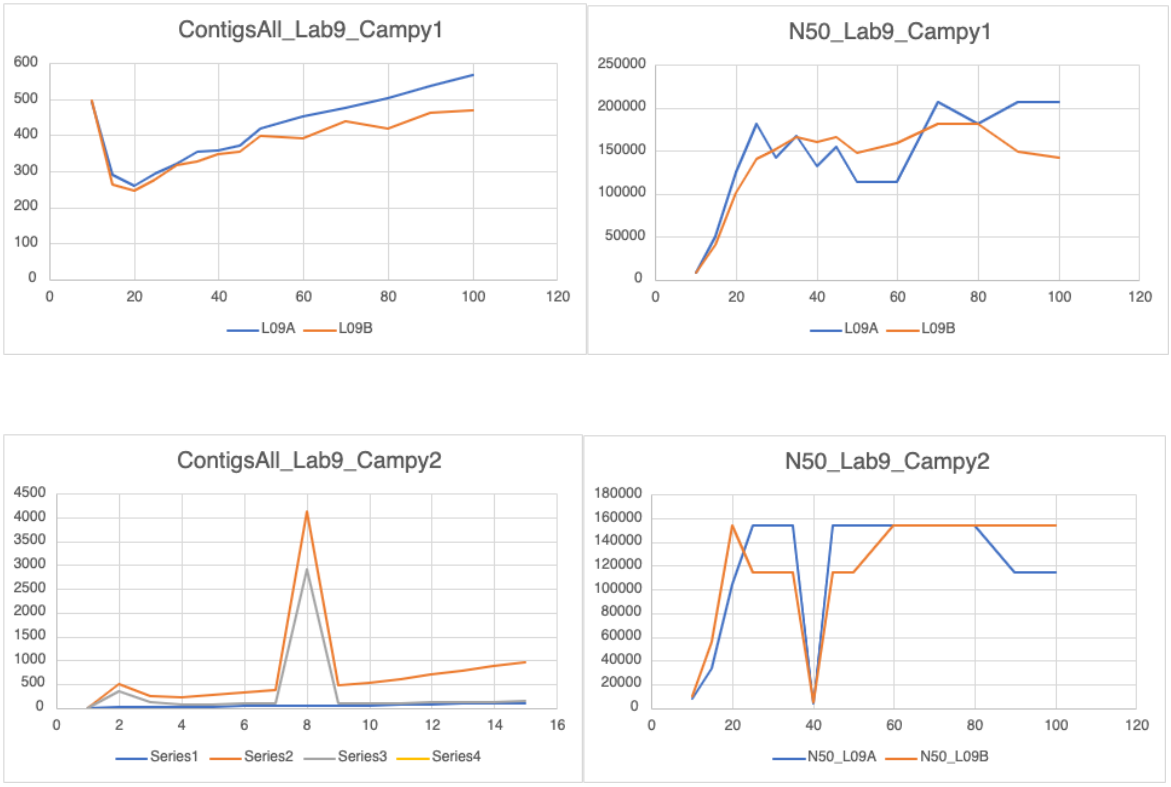


Figure 35*. The contigs and N50 values in different coverages for the L09, strains Campy1 and Campy2.*

For Lab10:

Lab10 sequenced two replicates with the same instruments NovaSeq 6000 and NEBNext Ultra II FS Library kit for both replicates. The read length was 2x151 bp. For the Campy1 the run gave 1.75 and 1.95 Gbases (1080 and 1204X coverage, respectively). For the Campy2 the run gave 1.73 and 1.77 Gbases (1006 and 1029X coverage, respectively). FastQC showed the error flags in “Per base sequence content” but it did not show deviating base composition of the very last base as the other labs, thus L10 passed “Per base sequence content” for both strains Campy1 and Campy2 (Figure 36).

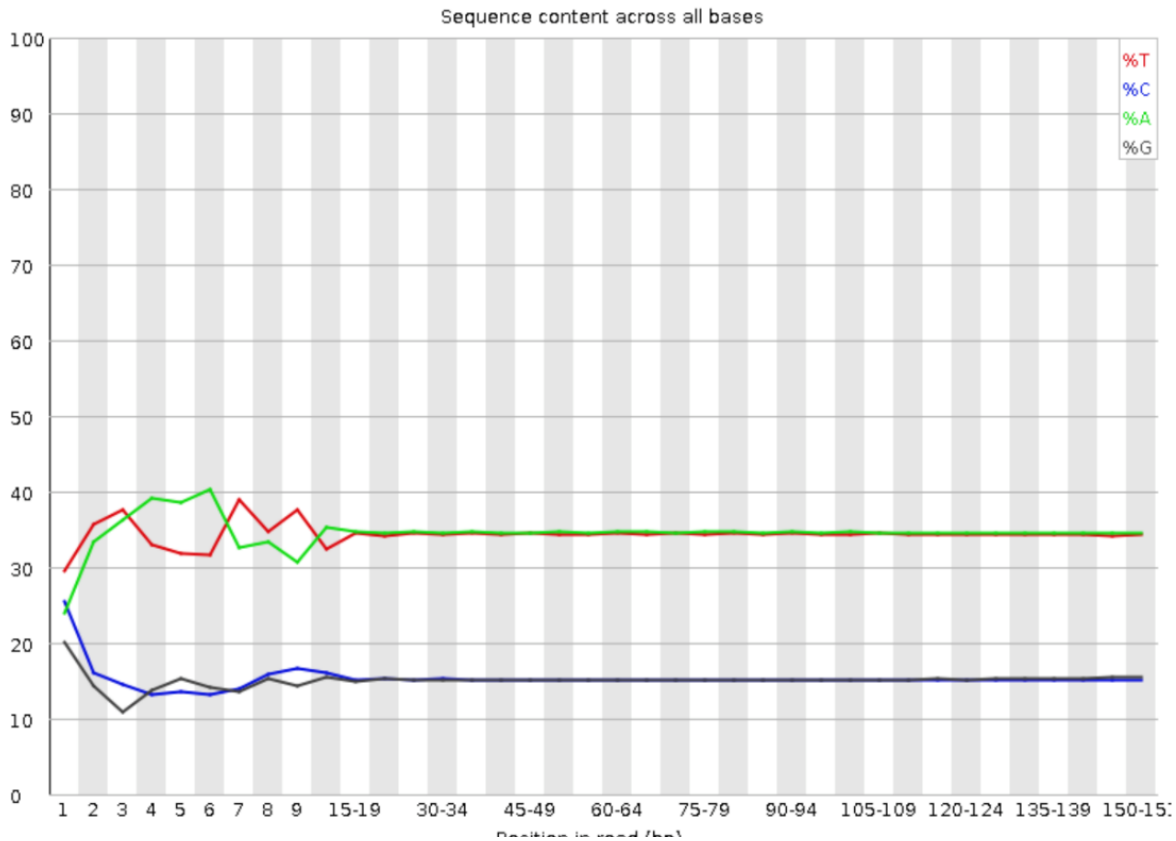
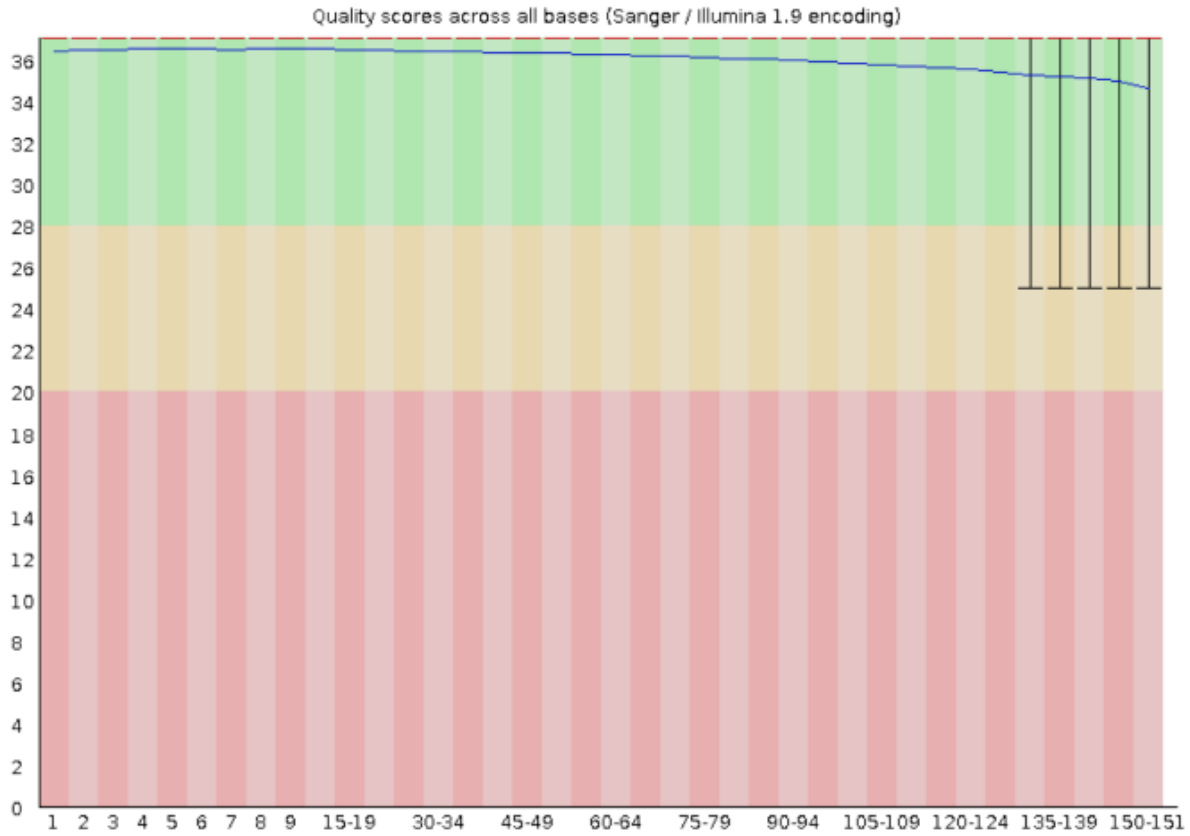


Figure 36. *The "Per base sequence content" showed an abnormal base composition in the first 9 bases but did not look similar to other labs. Also, the very last base showed no abnormal base composition.*

As the other labs the "Per base sequence quality "showed high quality along the whole reads (Figure 37).


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 15-19 30-34 45-49 60-64 75-79 90-94 105-109 120-124 135-139 150-151

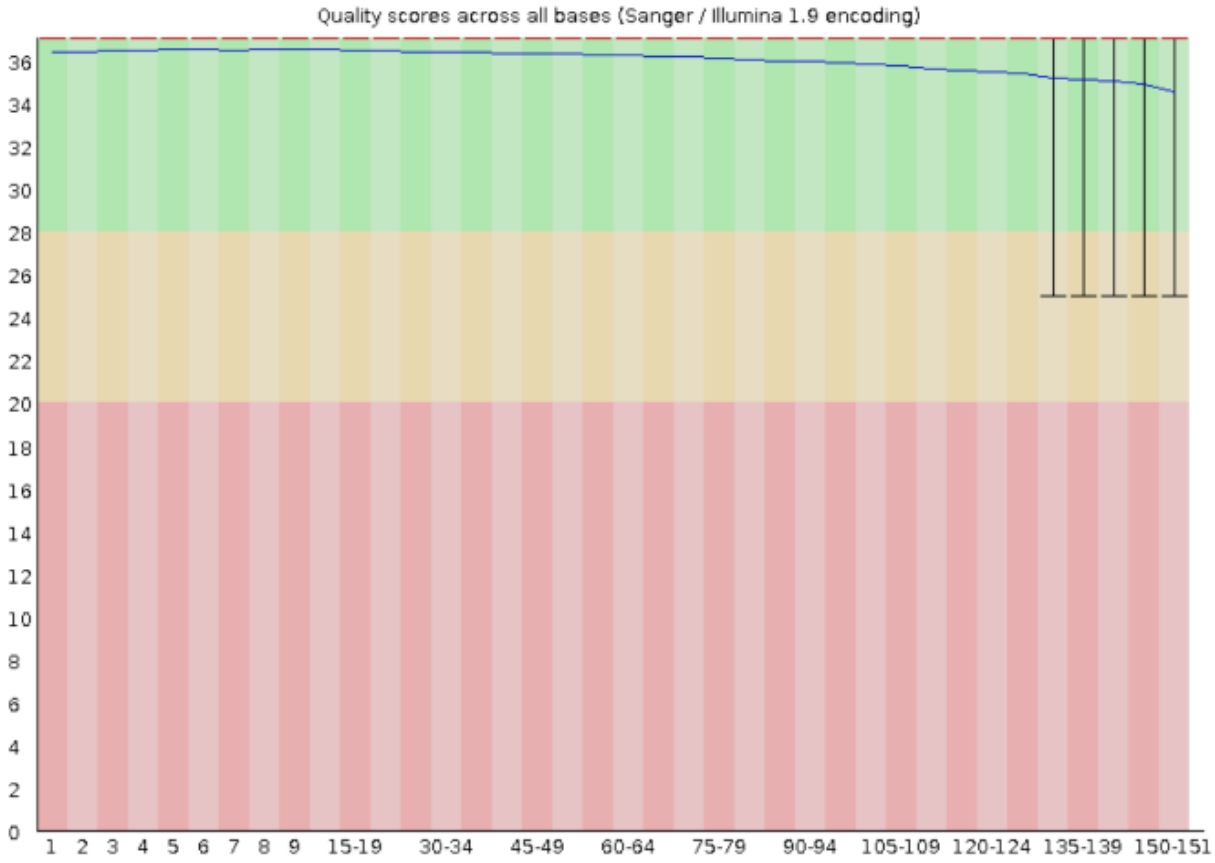


***Figure 37**. The "Per base sequence quality" of the L10A, L10B (NovaSeq) data for strain Campy1.*

The Kraken contamination check showed no sign of contamination from other species (Figure 38).

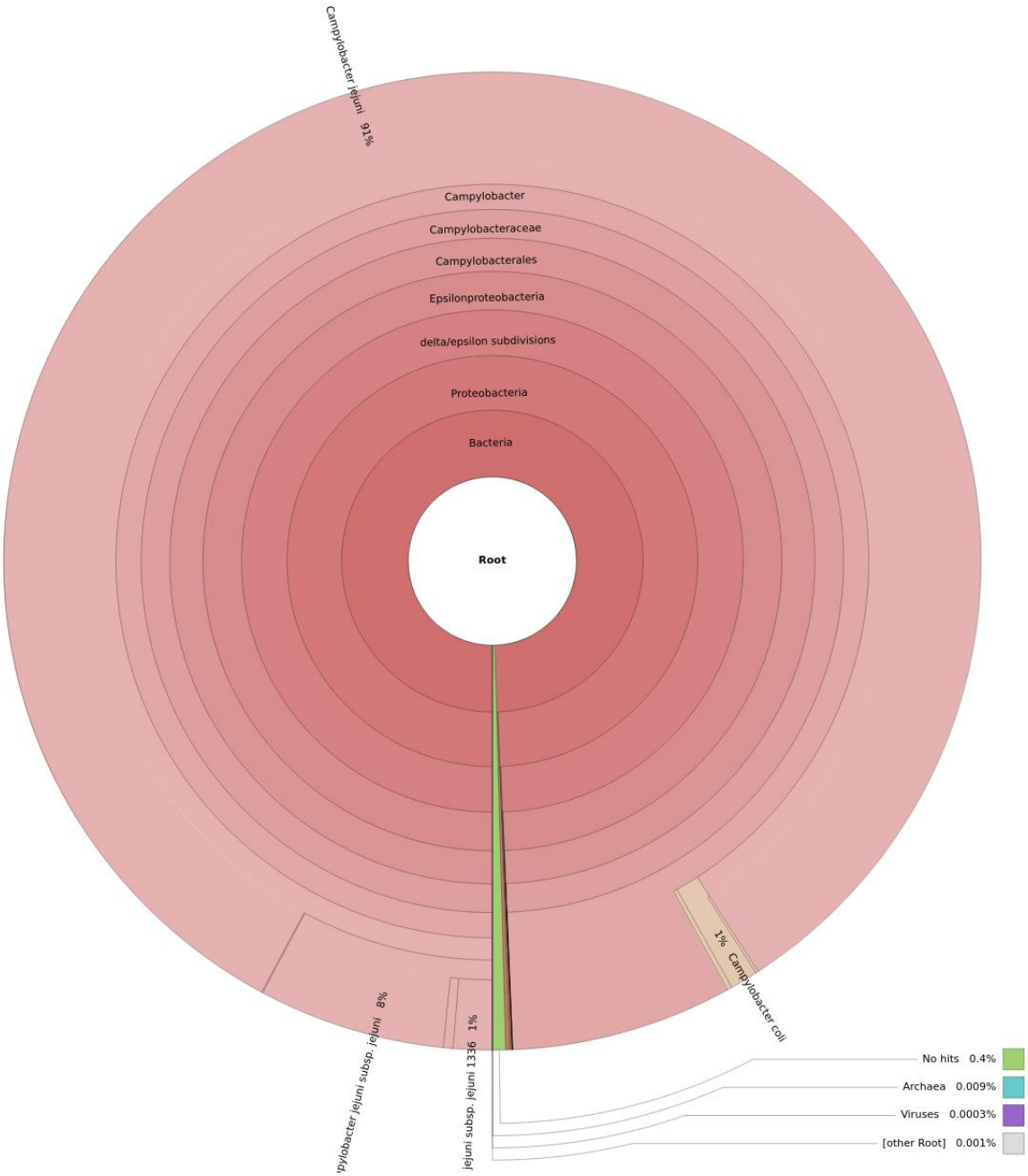
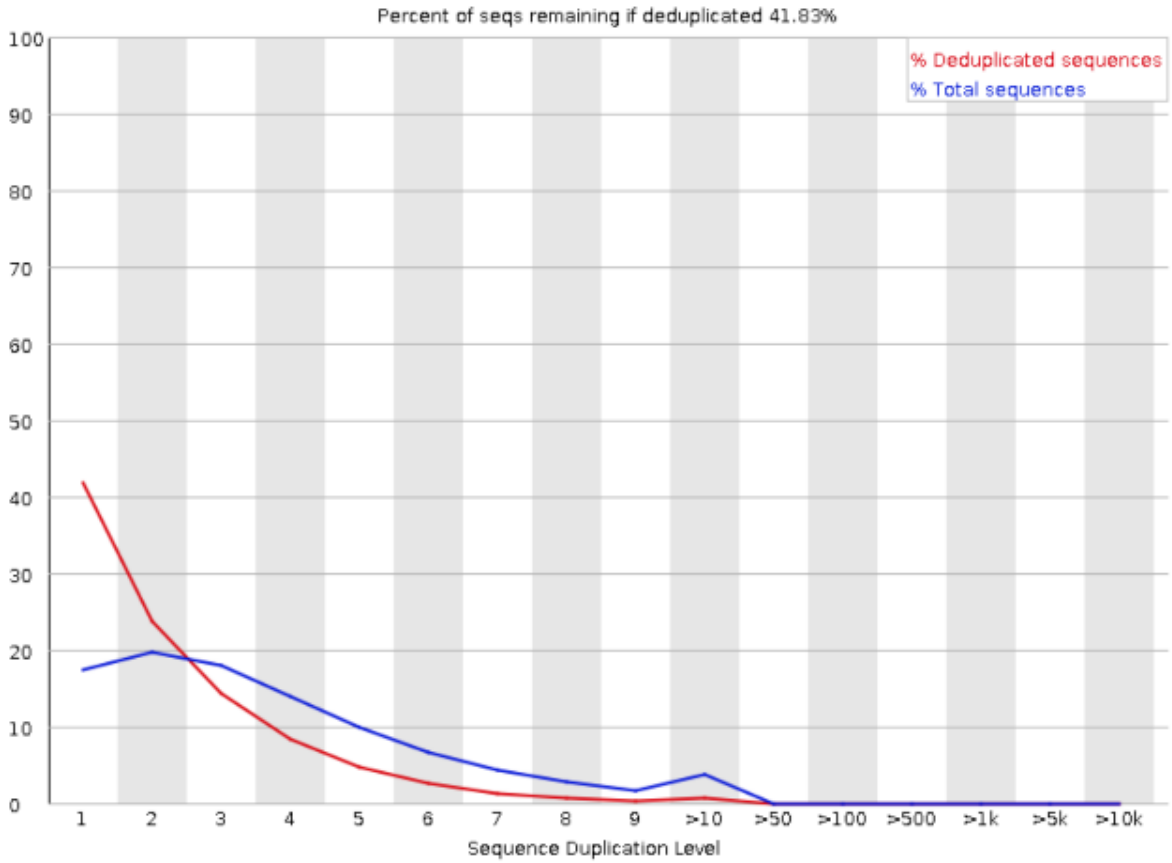


***Figure 38**. (Strain Campy1, sample L10A). As for the pervious labs minor amount of sequences was classified as other taxa than Campylobacter.*

FastQC-file for the L10 showed an error flag in sequence duplication levels as the L02C (Figure 39). By taking the word count of the fastq-file of L10A (strain Campy1) and dividing by 4 to get the number of sequences that corresponds to 1080X coverage, we then tested how many sequences are duplicated at 900X ,800x.... etc. We found that the error flag WARN started at 700X and disappeared at the coverage 300X (Figure 39).


Percent of seqs remaining if deduplicated 41.83%
% Deduplicated sequences
% Total sequences
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
>10
>50
>100
>500
>1k
>5k
>10k
Sequence Duplication Level

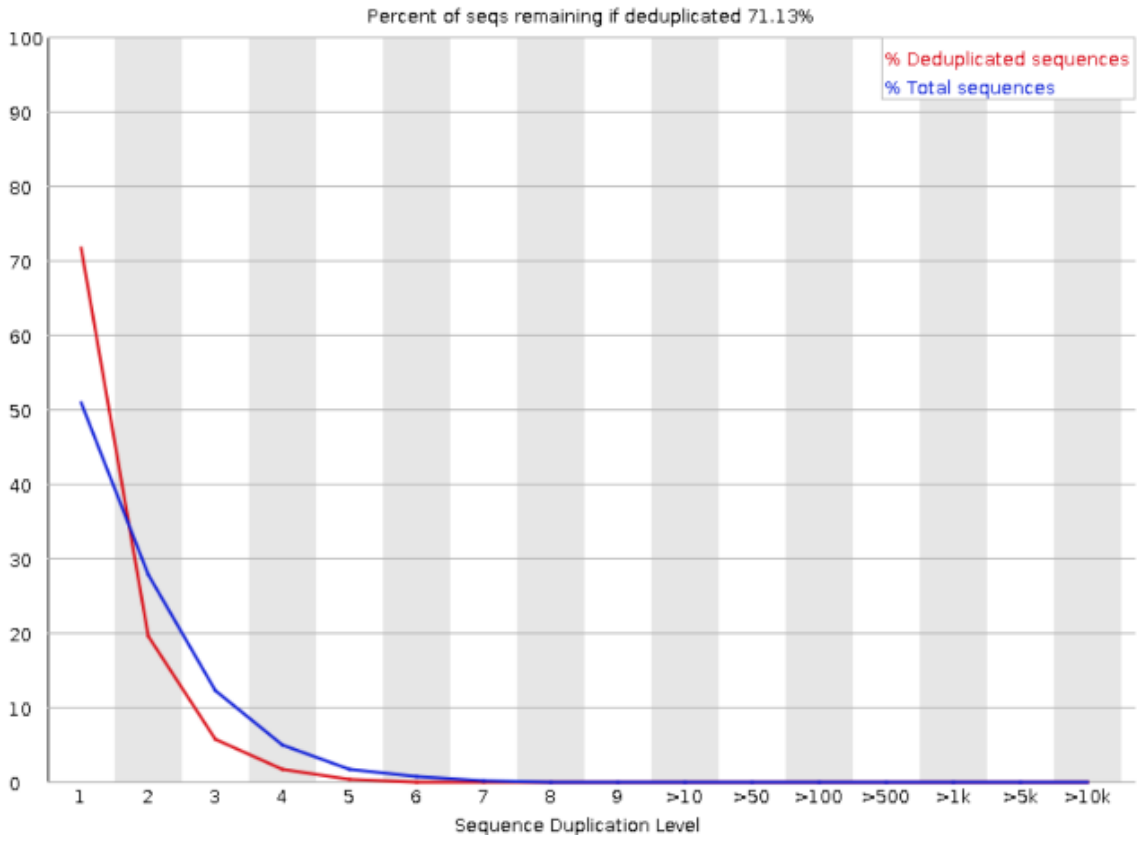
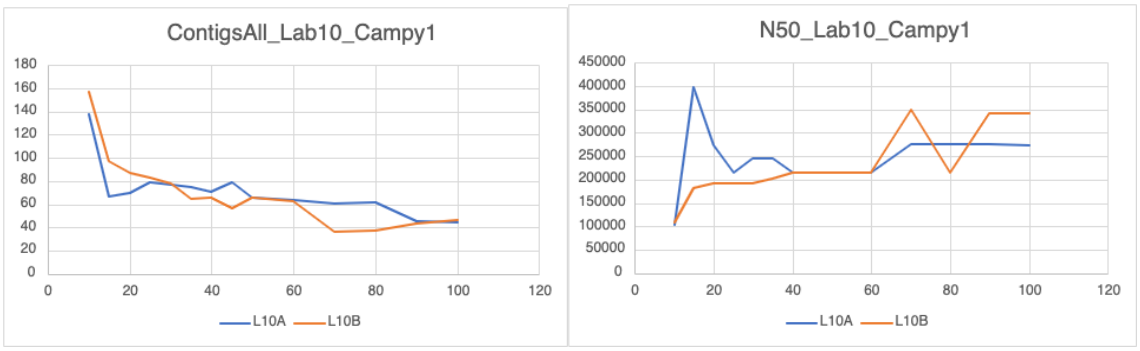


***Figure 39**. The error flag in "Sequence Duplication Levels" for the L10A (strain Campy1) at the coverage 1080X on the left versus no error flag at the coverage 300X on the right.*

Assembled data using SPAdes for both strains Campy1 and Campy2 showed that the contigs decreased radically until in the beginning then kept decreasing slightly until the end for both strains Campy1 and Campy2. N50 first showed sharp increase followed by a decrease at first for the L10A, strain Campy1 and then showed "up and down"-pattern for both L10A, L10B, strain Campy1. N50 for both replicates, strain Campy2 increased then showed constancy along the curve (Figure 40).



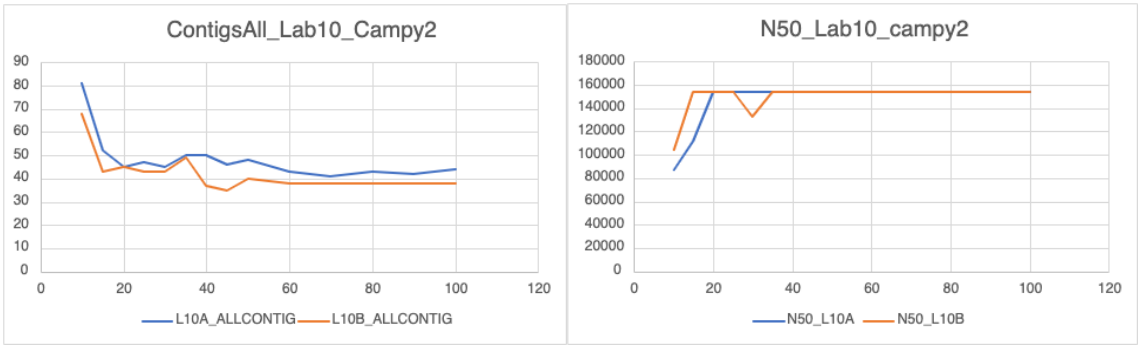


***Figure 40**. Change in total number of contigs and N50 values when amount of data used (coverage) was varied for the L10 strains (Campy1 and Campy2).*