UPPSALA UNIVERSITET

UPTEC X 22017

Examensarbete 30 hp

Juni 2022

Comparison of quality performance of whole genome sequencing analysis pipelines for foodborne pathogens

Chelsea Ramsin

Civilingenjörsprogrammet i molekylär bioteknik

UPPSALA UNIVERSITET

Comparison of quality performance of whole genome sequencing analysis pipelines for foodborne pathogens

Chelsea Ramsin

Abstract

Campylobacter is the leading cause of gastroenteritis worldwide and in Sweden there are official programs for the surveillance of the bacteria. One important objective with foodborne pathogen surveillance is molecular typing. As typing based on whole genome sequencing data is becoming more common, knowledge on how to set up analysis pipelines is essential to avoid variation in results. Here, typical whole genome sequencing pipelines are compared to a reference genome at different analysis stages to optimize assembly quality and typing results using cgMLST. The results show that read trimming is optimal to obtain high quality assemblies with SPAdes as well as for improving cgMLST results compared to when no read trimming was performed before assembling with SPAdes. The opposite was shown for SKESA where trimming beforehand had negative effects on the results, most likely due to SKESA having built in trimming properties. Additionally post assembly improvements had generally positive effects, however these effects were small.

Teknisk-naturvetenskapliga fakulteten

Uppsala universitet, Utgivningsort Uppsala

Handledare: Bo Segerman Ämnesgranskare: Lisa Klasson

Examinator: Siv Andersson

Bättre genomanalys minskar risken för matförgiftning!

Du vet förmodligen att man inte ska äta rå kyckling på grund av att man kan få Salmonella och bli matförgiftad. Men visste du att det inte bara är Salmonella som kan orsaka sjukdom vid konsumtion av rå kyckling? Kyckling bär även på en bakterie som heter Campylobacter som även den orsakar magsjuka. För att undvika stora utbrott av bakterien övervakar myndigheter i Sverige Campylobacter genom att ta prover från kyckling som undersöks för bakterien. En viktig del av övervakning och en eventuell smittspårning vid utbrott är att kunna identifiera olika typer av Campylobacter. Detta gör man genom att sekvensera hela bakteriens genom och ge gener varsitt ID så att man kan jämföra skillnader i gener mellan olika Campylobacter-prover. Det finns många olika verktyg man kan använda för genomanalys av en organism, och ett problem är att det är möjligt att få olika resultat beroende på vilka verktyg man använder. Det kan bland annat göra att generna får fel ID, vilket inte är bra. I det här projektet undersöktes olika datorprogram som används för genomanalys för att hitta vilka program som ger bäst resultat vid analys av Campylobacter, men också vilka program som gör att rätt ID ges till generna genom att jämföra med ett referensgenom med kända IDn för varje gen.

När man extraherat DNA från en bakterie så måste DNAt klippas upp i mindre bitar som sekvenseras på en maskin för att man ska få ut så kallade reads. Man kan använda trimningsprogram som tar bort delar av innehållet i readsen som är av dålig kvalité. Efter att man har fått sina reads använder man assemblyprogram för att få ut en assembly som representerar bakteriens genom. De program jag använt i denna rapport för assemblyn är SKESA och SPAdes. Intressant är att trimning, som används ofta och av väldigt många, gjorde att kvalitén på assemblies skapade med programmet SKESA blev sämre än när man inte trimmade readsen innan. Jämförelsevis så gjorde trimning av reads innan man assemblar med programmet SPAdes att det blev bättre assemblies. Det var också tydligt att SPAdes gav bättre resultat än SKESA.

Efter att man fått fram assemblies kan man finslipa dem, bland annat med programmet Pilon som försöker förbättra ens assembly. När Pilon användes på assemblies skapade med SKESA hade Pilon en positiv effekt, det blev förbättringar! Det blev förbättringar även när SPAdes användes, men då bara om man trimmat readsen innan man assemblade.

Slutsatsen som man kan dra av det här projektet är att vissa program fungerar bättre tillsammans än andra och att det går att rekommendera vissa program som ger bättre resultat när man sekvenserar Campylobacter och ska identifiera olika typer av bakterien. Denna information kan bidra till att mer optimala verktyg används vid övervakning av Campylobacter.

Table of contents

Abbreviations

MLST	Multi locus sequence typing
cgMLST	Core genome multi locus sequence typing
wgMLST	Whole genome multi locus sequence typing
PFGE	Pulse field gel electroforesis
DNA	Deoxyribonucleic acid

1 Introduction

Campylobacter is a genus with gram negative pathogenic bacteria carried asymptomatically predominantly by birds as well as other animals. Campylobacter causes campylobacteriosis which is responsible for the most gastroenteritis cases in humans worldwide, with symptoms such as fever, abdominal pain, nausea and diarrhea among others. Transmission of the bacteria to humans can happen in several ways, most commonly by handling or consuming undercooked contaminated poultry or other meat. Furthermore, consuming unpasteurized dairy products or drinking from contaminated water are also ways for the bacteria to infect to humans. *Campylobacter jejuni* is responsible for the vast majority of campylobacteriosis, followed by *Campylobacter coli* (SVA 2020a). *Campylobacter jejuni* has a relatively small genome with its genome length of approximately 1,600,000 base pairs. The actual length varies by strain. Approximately 94% of the genome is protein coding and the genome has few repeats. (Parkhill *et al.* 2000)

Occasionally, larger outbreaks of *Campylobacter* occur in Sweden and worldwide (SVA 2020a) and in order to prevent outbreaks from happening, surveillance of food borne pathogens is important (Lindsey *et al.* 2016). According to Thacker the definition of surveillance in epidemiology is the following: ” [..] the systematic collection, analysis, interpretation and timely dissemination for the planning of [...] public health programmes” (Thacker 1988). In Sweden there are official programs for the surveillance of *Campylobacter* in animals, specifically broiler chickens. Fecal samples are taken from 10 broiler chickens per slaughter batch and analyzed for *Campylobacter*. In humans, campylobacteriosis is a notifiable disease, i.e confirmed cases of *Campylobacter* infections are required to be reported to a regional disease control physician and the Swedish public health authority (SVA 2020a).

1.1 cgMLST and wgMLST, tools for surveilling *Campylobacter*

One of the major objectives of foodborne pathogen surveillance is to differentiate between populations of bacteria to be able to infer what strain or type is carried by for example a flock of broiler chickens or is responsible for an infection. Previously two methods called Multi locus sequence typing (MLST) and pulse field gel electrophoresis (PFGE) have been used to type differences between populations of food borne pathogens including *Campylobacter*. PFGE put simply is a gel electrophoresis performed on lysed isolates which have been fragmented with restriction enzymes. During the electrophoresis the fragments will be separated by size giving rise to patterns on the gel which act as a barcode. These barcodes can be used to differentiate between strains and allows for typing them (Schwartz & Cantor 1984; Sharma-Kuinkel *et al.* 2016). In MLST however, a rather small number of housekeeping genes are sequenced and alleles are identified from the sequences (Maiden *et al.* 1998). The number of housekeeping genes varies among studies but is generally between 7-11 genes, with seven being most commonly used (Maiden *et al.* 1998; Jolley & Maiden 2010; Dingle *et al.* 2001; Payne *et al.* 2020). Each allele in each sequence is given a number and the combination of numbers make up a so called sequence type (ST) which can be used as an identifier for a particular population of bacteria (Maiden *et al.* 1998). However, as progress in whole genome sequencing (WGS) has increased rapidly over the last couple of years, typing based on whole genomes or core genomes instead of a few housekeeping genes can be done to increase resolution. These methods are called cgMLST (core genome multi locus typing) and wgMLST (whole genome multi locus typing) respectively and each sequence at each locus is given an allele number similarly to regular MLST but on a larger scale. Thus cgMLST and wgMLST can distinguish more differences between different isolates than MLST can (Yan *et al.* 2021; Cody *et al.* 2013). Besides offering a high typing resolution, cgMLST and wgMLST are easily standardized due to utilizing bacterial gene schemes containing a fixed number of genes curated by different research groups. The schemes are used for typing and no reference genome is needed allowing for standardization and reproducibility (Deneke *et al.* 2021).

As an alternative to a gene-by-gene approach, i.e cgMLST and wgMLST, isolates can be typed with SNP analysis where SNPs (single nucleotide polymorhpisms) are identified. The SNP analysis approach offers an even higher resolution than gene-by-gene approaches, however it is more difficult to standardize among different laboratories due to recombination and the method requiring a reference genome. (Pearce *et al.* 2018)

1.2 *De novo* assembly

Since both cgMLST and wgMLST are based on core genomes or whole genomes, WGS is a mandatory precursor for the analyses. To assemble a *de novo* genome, DNA has to be sequenced on a sequencing machine to generate raw reads which are assembled into contigs and/or scaffolds. In the RefSeq database and among surveillance genomes, most genomes are assembled with reads generated by Illumina sequencers (Segerman 2020) and there are multiple Illumina sequencing machines and multiple library preparation kits commerically available (SVA 2020b; Segerman 2020). During the library preparation, adapters (short nucleotide sequences) are ligated to fragmented DNA which can adhere to the flowcell of the Illumina sequencing machine onto which the sequencing is carried out. Additionally, barcode sequences are added along with the adapters to allow for sample multiplexing. The barcode sequences therefore act as identifiers so individual samples can be distinguished.

After sequencing, raw DNA reads need to be assembled into *de novo* genomes. Available today are numerous different software used to assemble raw reads and they can differ in what algorithms are used and what parameters are available. In the RefSeq database, most bacterial genomes are assembled with the software SPAdes while most surveillance genomes are assembled with SKESA, developed by the NCBI (Segerman 2020). Additionally, before assembling the genome, raw reads can be trimmed to remove adapters and filter away reads that are too short. Trimming can additionally remove segments of the reads that are of bad quality. It is especially valuable to remove the end of the reads since the per base read quality generally decreases at the end (Bolger *et al.* 2014; Chen *et al.* 2018). While trimming can increase the quality of assemblies, there is a risk that trimming could have negative effects on the assembly. Trimming could for example lead to more fragmented assemblies (Del Fabbro *et al.* 2013). The assembly program SKESA has built in adapter and quality trimming (Souvorov *et al.* 2018) while SPAdes does not (Prjibelski *et al.* 2020).

Furthermore assemblies can be processed further to try to improve the assemblies. There is a possibility that DNA from a previous sequencing run is left in the sequencing machine. This could lead to DNA from an entirely different organism being sequenced and assembled together with your samples and thus contaminating them. The contigs produced by contaminants are usually small and therefore it could be beneficial to remove small contigs from the assembly to hopefully eliminate contaminating DNA (SVA 2020b). Additionally there are software aimed to improve assemblies using read alignment analysis. The software Pilon is one of those software and maps reads back to the assembly to find inconsistencies. Pilon then aims to fix the inconsistencies and is able to reassemble genome regions. (Walker *et al.* 2014).

The chosen software for an assembly, along with trimming of raw reads and any post assembly improvements may affect the quality of the assembly. Library preparation, read depth or coverage may also have an effect (SVA 2020b). Since cgMLST analysis is based on WGS, the assembly quality could consequently affect the cgMLST results. One study performed a cgMLST analysis on assemblies assembled with raw reads of different coverages, spanning the range 10x to 500x, and used among others SPAdes and SKESA assemblers. It was observed that SPAdes required coverages of at least 30x while SKESA needed coverages of 40-60x to not get a high cgMLST error rate (Liu *et al.* 2021). A different study found that a coverage of 40x is beneficial for cgMLST analyses based on SPAdes assemblies (Palma *et al.* 2022) and also found that wet lab work had no effect on the results. However none of these studies investigated Campylobacter.

1.3 Aims and purpose

The purpose of this project is to better understand and optimize typical whole genome sequencing pipelines by measuring and comparing different quality aspects of WGS data at different analysis stages. This could make it possible to recommend how analysis pipelines should be set up for Campylobacter and could reduce variation in analysis results.

2 Material and methods

In this section the materials used is presented along with all methods from the pipeline creation, how to run the pipeline and how results are compiled.

Scripts and the scientific workflow can be found at the following GitHub page: https://github.com/chels0/Quality_performance_of_WGS_analysis_pipelines

2.1 Sequencing data

The raw data is *Campylobacter jejuni* ST-464 raw reads (n=25) of varying quality sequenced by different reference laboratories for Campylobacter in the EU. Half of the raw reads were sequenced from genomic DNA given to the different laboratories by SVA (the Swedish National Veterinary Institute) which is the official EU reference laboratory for Campylobacter. The other half of the raw reads were sequenced from lyophilised cultures, also given to the laboratories by SVA. All raw reads are Illumina reads generated either with the Illumina DNA Prep kit or Nextera XT DNA Library Preparation kit. The majority of laboratories used Illumina MiSeq as sequencer (SVA 2020b) with readlengths of 300. However some laboratories used readlengths of 150 and 250. The coverage of the reads were downsampled to 20x, 50x and 100x. However a few samples did not reach a coverage of 100x.

A complete genome from the same ST-464 isolate is used to benchmark the assemblies against and this is considered to be a correct assembly. The complete genome was sequenced with both long read and short read sequencing using Oxford nanopore technology and Illumina respectively. (SVA 2020b)

2.2 Creating a scientific workflow for *de novo* assembly

A scientific workflow with software encompassing the analysis from quality control, trimming, assembly, contig size filtering, assembly improvement and assembly validation was developed in Nextflow. The workflow takes reads in the form of forward and reverse fastq files from one or more samples as input along with a reference genome, originating from the same isolate as the samples. For an overview of what software were used in the workflow along with their inputs and outputs as well as how the different software is linked, see Figure 1.

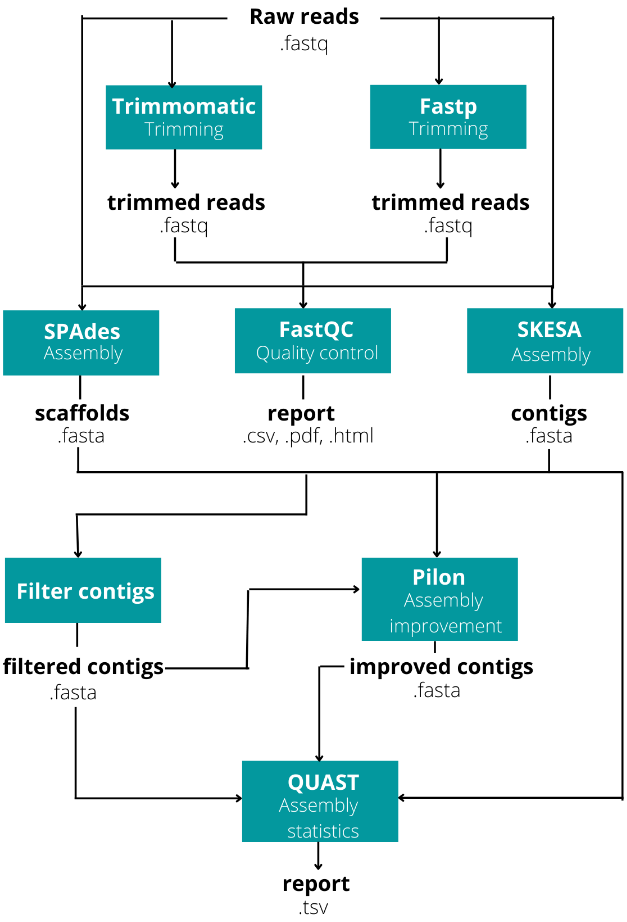


Figure 1: Structure of the Nextflow workflow, including software with their respective inputs and outputs

2.2.1 Software details

Trimmomatic (Bolger *et al.* 2014) and Fastp (Chen *et al.* 2018) were used to trim low quality reads from the read data. Fastp was run with default parameters while Trimmomatic was run with parameters 'LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 ILLUMINACLIP:NexteraPE-PE.fa:2:30:10:2:true' for all samples generated with

Nextera library preparation kits and parameters ’LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36 ILLUMINACLIP:’TruSeq3-PE-2.fa:2:30:10:2:true’ for the samples PT28-1-27 and PT28-3-27 which were generated with a TruSeq library preparation kit.

FastQC (Andrews *et al.* 2012) as well as Fastp were used for quality control of trimmed reads after trimming had been performed. FastQC was run with default parameters while Fastp was run with flags -A -L -Q -G to utilize the quality control component of the software without trimming.

The assembly programs SPAdes (Prjibelski *et al.* 2020) and SKESA (Souvorov *et al.* 2018) were used for assembling the reads into scaffolds and contigs. SKESA was run with default parameters and SPAdes was run with options –isolate or –-careful to investigate differences in results between the options.

The filtering step removed short contigs from assemblies meeting a certain basepair amount threshold, here contigs of size 200 and 500 basepairs.

The software Pilon (Walker *et al.* 2014) was used to improve the assemblies using the assemblies and BAM files as input. The BAM file was generated by mapping raw reads back to the generated assembly using Bowtie2 using default parameters. Pilon was run with default parameters as well.

QUAST (Gurevich *et al.* 2013) was used to evaluate assemblies based on various metrics using default parameters.

2.3 Running different pipelines with the scientific workflow

First, one WGS pipeline setup was generated automatically for each possible software combination by choosing one of the trimming options, one assembler and then any post assembly improvement option, either filtering, Pilon, both or nothing at all. Each pipeline was then run with the scientific workflow. Thus each iteration had a unique setup of software combinations. Then, for each pipeline, the software MultiQC (Ewels *et al.* 2016) was used to compile all assembly output statistics for each sample generated with QUAST. The following four QUAST metrics were plotted as bar plots for each pipeline: N50 values, number of contigs, genome fraction in percent (the percentage of aligned bases to the reference genome) and misassemblies. The N50 value can be defined as such: If contigs are ordered from biggest to smallest until they make up 50% of the genome, the N50 value is the length of the smallest contig. This metric is often used to describe the completeness of the genome.

2.4 cgMLST analysis

The software chewBBACA (Silva *et al.* 2018) was used to perform a cgMLST analysis by taking all assemblies for each combination along with the reference genome as input. The cgMLST scheme for *Campylobacter jejuni* used for the allele calling in the cgMLST analysis was from the Innuendo project (Llarena *et al.* 2018) and contained 678 loci from the core genome of *Campylobacter jejuni.* The output from the cgMLST analysis was the allele numbers at every loci for each sample and the reference. However one of the samples was discarded as it had been assembled with the wrong settings.

Next, the allele number at each loci for the samples in the chewBBACA output were subtracted with the references' allele number for every pipeline. Consequently, an allele number of zero represents loci where the allele numbers of the reference and the sample is the same. Thus the allele calling for that particular loci and sample is deemed correct. Inversely, a non-zero allele number represents a wrong called allele, missing alleles or alleles not present in the schema. Additionally the number of non-zero allele numbers for every sample was summed up for each pipeline. Since there are a total of 678 loci for each sample, the combined amount of loci for the summed up samples (n=25) is 16,950 for each pipeline. As such, if every single sample differed from the reference for a pipeline there would be 16,950 differences.

Furthermore a pairwise comparison between pipelines was performed. To avoid confounding variables and multivariate analysis, comparisons were only made between pipelines that were identical except for when one of the pipelines strictly had one added software. This was done to find loci which differed from each other after the reference's allele number had been subtracted from each samples' allele number. The total number of loci which differed between the compared pipelines was counted along with how many of the differences between the pipelines were corrections, errors and changes from one error to another. A correction is when the pipeline with additional software had an allele number of zero while the other pipeline had an allele number of non-zero. An error is when the pipeline with more software had a non-zero allele number while the other pipeline had an allele number of zero. A change from one error to another is when both pipelines had non-zero values not equalling each other. The differences, proportion of corrections, proportion of errors and proportion of changes were plotted as stacked bar plots.

3 Results

38 different pipelines containing software common for the analysis of *Campylobacter jejuni* have been compared on the basis of how pre-processing of reads and post assembly improvements affect assemblies generated with SKESA and SPAdes. The main results are assembly quality statistics from QUAST as well as how many differences and what types of different allele calling differences are observed between the assemblies and the reference genome in a cgMLST analysis.

3.1 Coverage and its effect on assembly quality and cgMLST

Reads were downsampled to coverages of 20x, 50x and 100x to investigate what effect the different coverages had on the assembly quality and the cgMLST analysis. Regarding assembly quality, increasing coverage decreased the number of contigs (Fig 2A) and misassemblies (Fig 2D) for all pipelines which is desirable. Meanwhile increasing coverage yields higher N50 values (Fig 2B) and genome fraction percentages (Fig 2C) for all pipelines which is ideal for these metrics. It is also noticeable that pipelines including SPAdes performed better than pipelines including SKESA since the median N50 values and genome fractions increased while the median amount of misassemblies and contigs decreased.

Additionally, increasing coverage produced fewer allele calling differences from the reference for every pipeline (Fig 3). For exact values see Appendix A table A2.

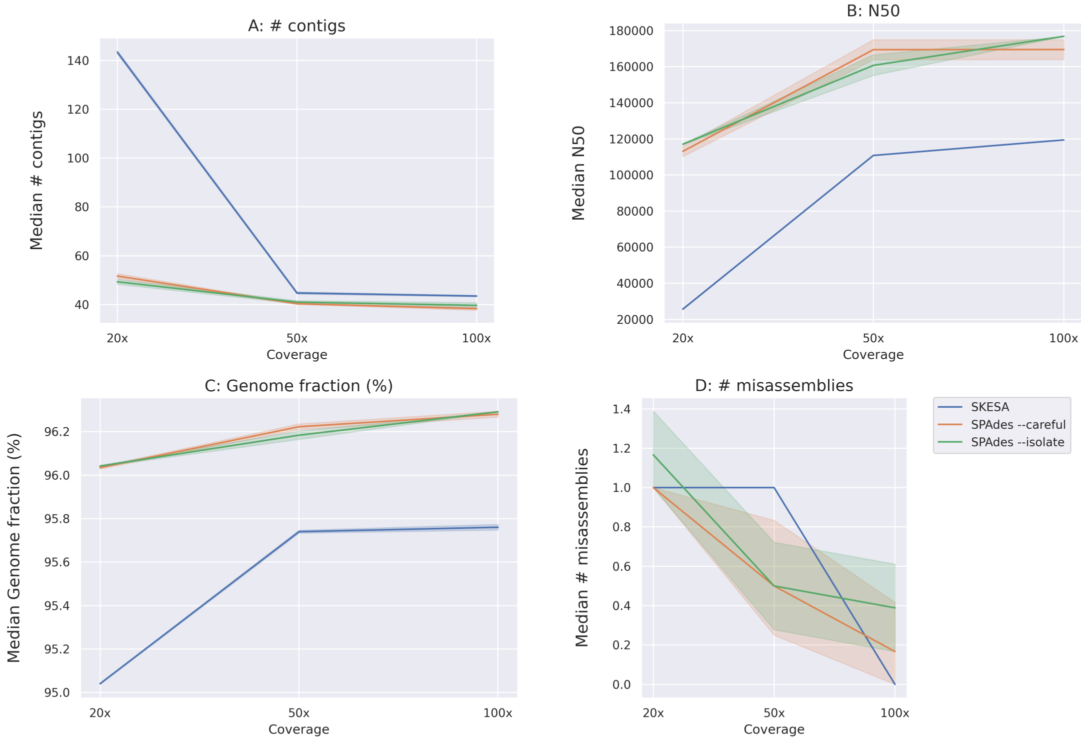


Figure 2: Line plots depicting the distribution of the median QUAST metric values generated by all pipelines aggregated on assembler option. Bold lines represent the mean median QUAST metric for the aggregated pipelines in regards to assembler option and the transparent areas show the 95% confidence interval. ***(A)*** *The median number of contigs* ***(B)*** *The median N50 values.* ***(C)*** *The median genome fraction* ***(D)*** *The median number of misassemblies*

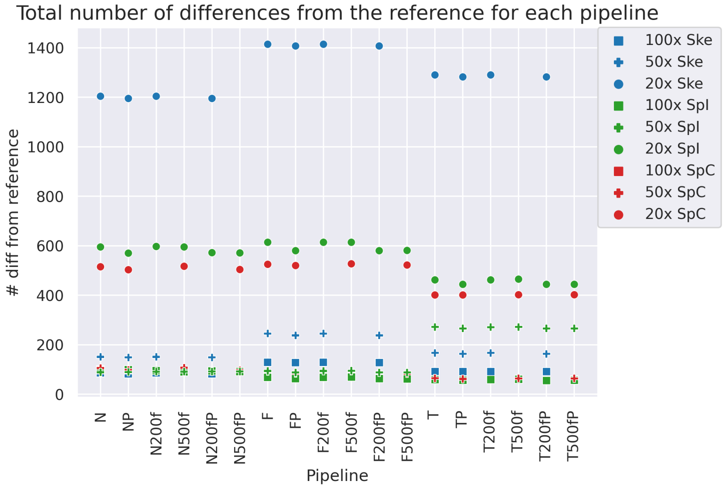


Figure 3: The total number of alleles which differed from the reference for all pipelines in regards to coverage (20x, 50x and 100x). Ske= SKESA, Spl= SPAdes –isolate, SpC= SPAdes –careful, N= No trimming, T= Trimmomatic, F= Fastp, P= Pilon, 200f= filtering contigs of size 200, 500f= filtering contigs of size 500.

Considering the observation that increasing coverage generally improves all metrics, only results for coverages of 100x are shown here on forward. For coverages of 20x and 50x, see appendices B-E.

3.2 The effect of pre assembly read-trimming on assembly quality and cgMLST results

Trimming is often recommended to filter away bases with lower quality and in this section of the results the actual effect of trimming on the assembly quality is investigated by comparing SKESA and SPAdes assemblies generated from untrimmed reds with assemblies generated with trimmed reads without post assembly improvements.

For SPAdes I observed that the median number of contigs decreased when Trimmomatic was used compared to when no trimming had been done before assembling. The spread in values was also decreased as values cluster more closely to the median, as seen by the small interquartile range compared to no trimming. However, even though most assemblies improved after trimming with Trimmomatic, a few assemblies seem to have worsened (Fig 4A). The N50 values for Trimmomatic and no trimming are similar to each other, with both having the same median and similar spread as well as minimum/maximum values (Fig 4B). The median genome fraction is slightly higher for assemblies which had been assembled with reads previously trimmed with Trimmomatic compared to when no read trimming had been done beforehand. Additionally using Trimmomatic raised the lower quartile and the minimum value which is satisfactory (Fig 4C). Trimming with Trimmomatic did not affect the amount of misassemblies as the results are identical to no trimming (Fig 4D).

For SKESA the following was illustrated. Regarding total number of contigs, the median was the same for untrimmed assemblies and assemblies trimmed with Trimmomatic (Fig 4A). However for Trimmomatic the upper quartile decreased slightly while the outliers got worse. The median N50 values was the same when no trimming had been done beforehand and when trimming had been used. However not trimming before assembling yielded more assemblies with high N50 values compared to when Trimmomatic had been used beforehand (Fig 4B). The median genome fraction is higher when Trimmomatic had been used, however the smallest genome fraction value decreased compared to when no trimming had been done (Fig 4C). Trimmomatic had no large effect on the amount of misassemblies however it did remove an outlier present when no trimming had been done (Fig 4D).

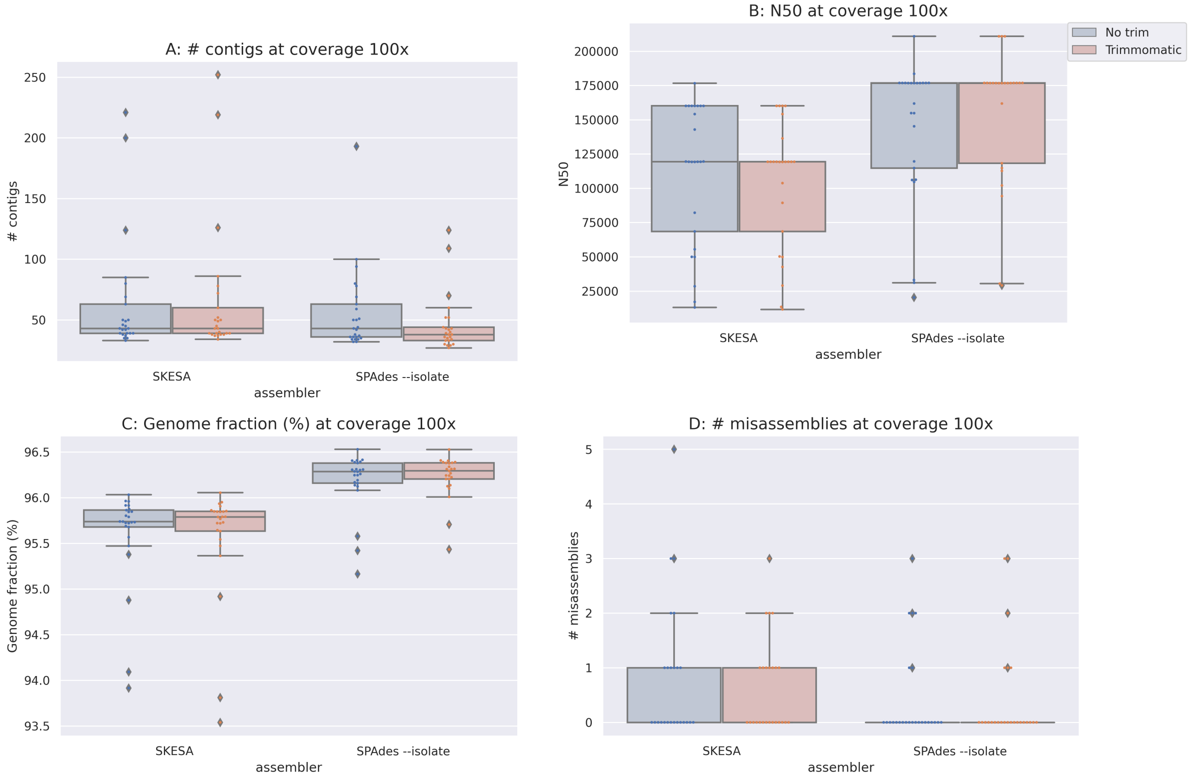


Figure 4: QUAST metrics for SPAdes and SKESA assemblies generated by reads trimmed with Trimmomatic (red) and untrimmed reads (blue) at coverage 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

Trimmomatic is not the only trimming software available, another common software is Fastp and the two were compared to each other to evaluate which trimming software was best paired with SKESA and SPAdes. For SKESA, Trimmomatic and Fastp perform equally for number of contigs (Fig 5A) and amount of misassemblies (Fig 5D), however Fastp had higher outliers than Trimmomatic for both metrics, which was not ideal. The median N50 values was the same for both Trimmomatic and Fastp. However using Fastp before assembling yielded more assemblies with high N50 values compared to when Trimmomatic had been used beforehand (Fig 5B). Trimmomatic scored better than Fastp regarding genome fraction as the median is higher and the lower quartile is higher (Fig 5C).

For SPAdes –isolate, Trimmomatic and Fastp had similar medians to each other for the following metrics: number of contigs (Fig 5A), N50 values (Fig. 5B) and genome fraction (Fig 5C). Additionally, Trimmomatic had a smaller upper quartile than Fastp regarding number of contigs and a larger lower quartile regarding genome fraction, both of which are more ideal. Adding to that, outliers strayed less from the rest of the values for Trimmomatic. For figure 5B the median N50 value is the same for both trimming options and the interquartile ranges are similar, however Trimmomatic had a longer bottom whisker, meaning 25% of the data points are relatively low. Worth noting is that Fastp had outliers with lower N50 values than Trimmomatic. Trimmomatic scored better than Fastp when investigating misassemblies as can be seen in figure 5D where the median is zero for Trimmomatic and 1 one for Fastp.

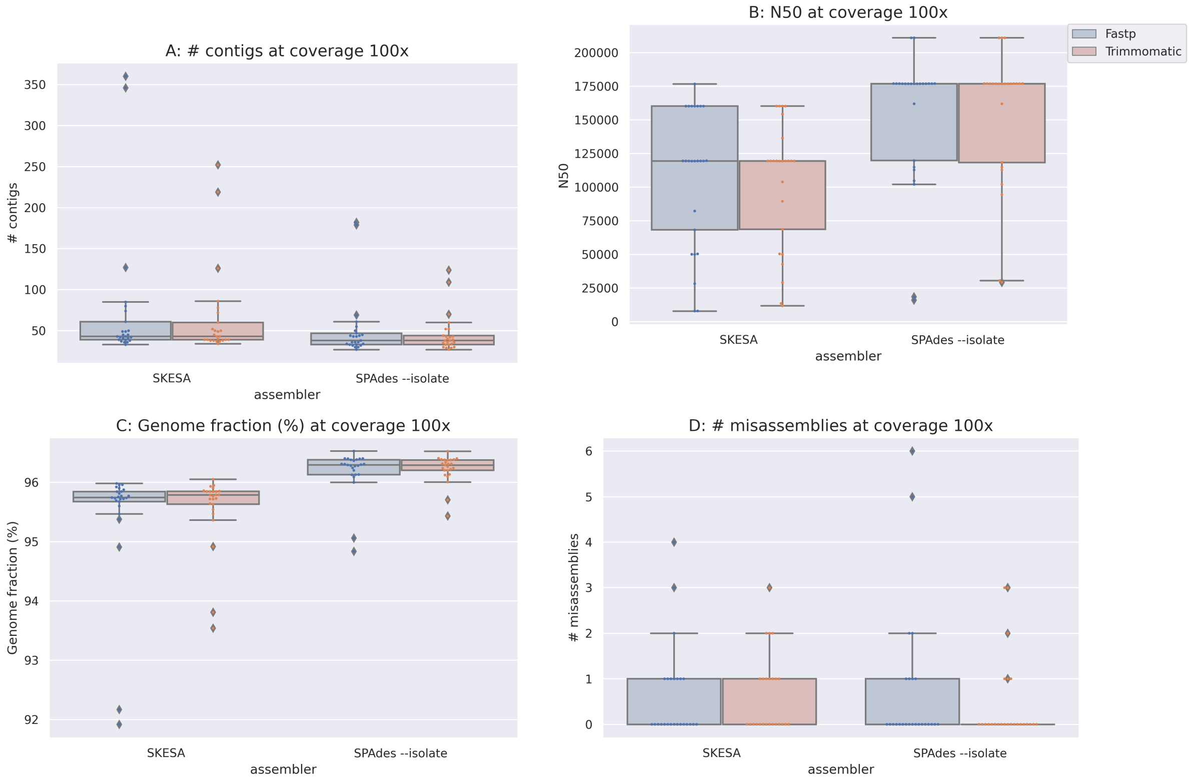


Figure 5: QUAST metrics for SPAdes and SKESA assemblies generated by reads trimmed with Trimmomatic (red) and reads trimmed with Fastp (blue) at coverage 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

After the cgMLST analysis had been performed and the total amount of loci differences from the reference for every sample had been summed up, SKESA assemblies generated from untrimmed reads had a combined total of 86 differences for all samples out of 16,950 possible while there were 98 differences for SPAdes assemblies trimmed beforehand, also out of 16,950 possible (Fig 3). Trimming before using SKESA introduces more new errors than corrections regardless of trimming software, however Fastp introduces more errors than Trimmomatic. For SPAdes, Trimmomatic had a net positive effect on the allele calling as more corrections than errors were introduced while Fastp had a net negative effect (Fig 6).

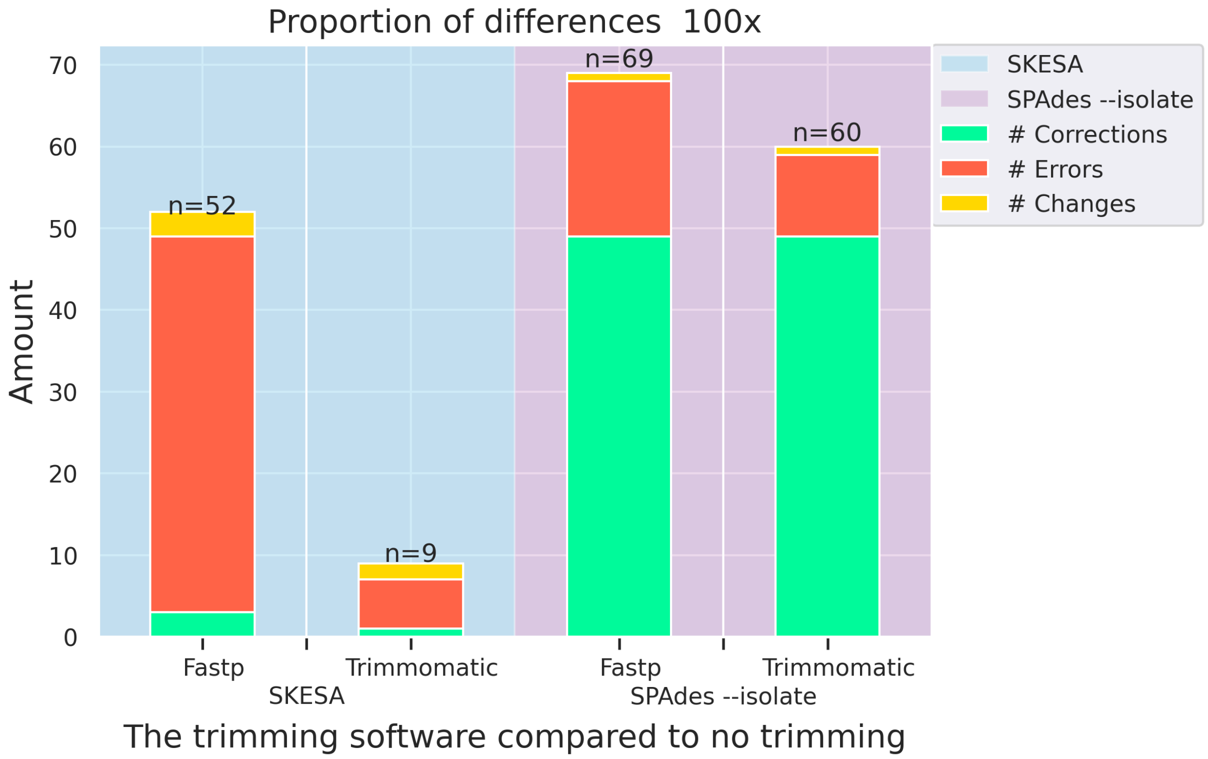


Figure 6: Differences in allele calling between trimmed assemblies and untrimmed assemblies (without post assembly improvements). n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive).Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). On the x-axis is software, y-axis is the amount of each change. The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple).

3.2.1 Results for other coverages than 100x

The observations seen at coverage 100x could generally be seen at coverages of 20x and 50x as well, both regarding assembly statistics (Appendix B, Fig B1 and Fig B2) and the cgMLST results (Appendix C, Fig C1). However at coverages of 50x, neither Trimmomatic or Fastp perform well when used before SPAdes with Trimmomatic producing several times more errors than Fastp, see appendix E figure E1.

3.2.2 SPAdes –isolate compared to SPAdes –careful

The SPAdes manual has two recommended modes which are incompatible with each other and it is unclear which would suit best for this data set. The two modes, –isolate and –careful, were therefore compared to each other to investigate which is the most optimal to use in regards to assembly quality and cgMLST analysis. The box plots for –careful and –isolate are more similar to each other than when SPAdes and SKESA were compared in regards to QUAST metrics, see appendix D figure D1 and D2. As such assembly quality is not enough to infer which option is best. However the cgMLST results are more clear. After the cgMLST analysis had been performed and the total amount of loci differences from the reference for every sample had been summed up, SPAdes –careful assemblies generated from untrimmed reads had a combined total of 99 differences for all samples out of 16,950 possible (Fig 3. Using Trimmomatic before SPAdes –careful yielded a higher correction rate than when Fastp was used beforehand. Comparing –careful and –isolate it is evident that the proportion of corrections, errors and changes are similar to each other with –isolate having a slightly higher proportion of corrections. However SPAdes –careful had fewer total amount of differences. At coverages of 50x however –careful performs better than –isolate since it is possible to gain a net positive amount of corrections when using SPAdes –careful (Fig 7).

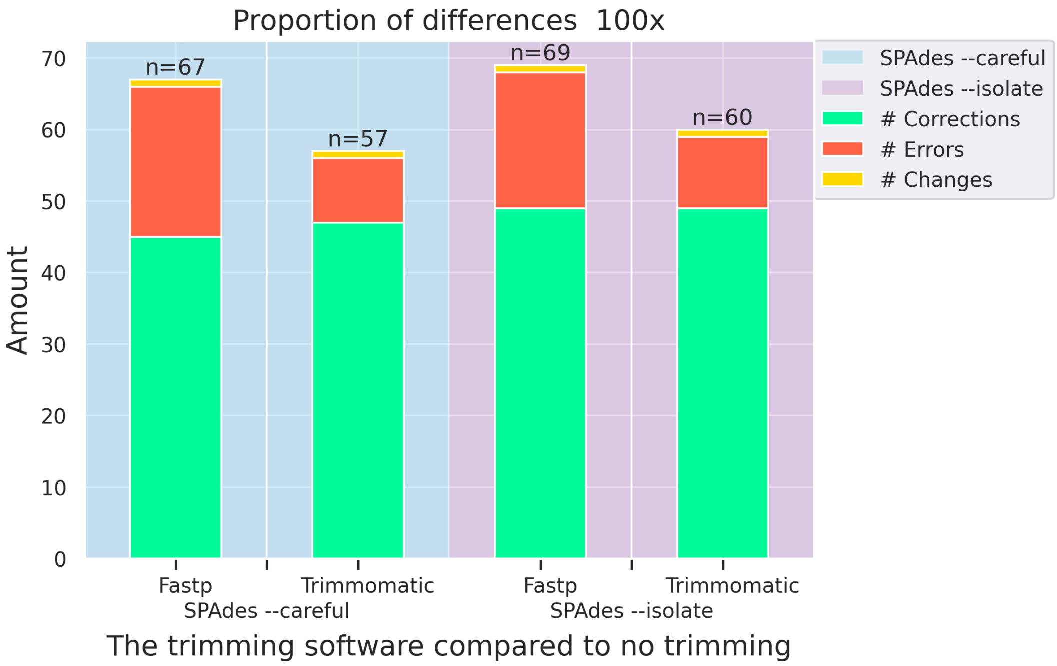


Figure 7: Differences in allele calling between trimmed assemblies and untrimmed assemblies (without post assembly improvements). n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive).Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). On the x-axis is software, y-axis is the amount of each change. The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple).

3.3 The effects of post assembly improvements

In this section of the results, SKESA and SPAdes assemblies without post assembly improvements have been compared to SKESA and SPAdes assemblies with post assembly improvements.

In figure 8 the effect Pilon had on QUAST metrics is shown. Pilon had a slight negative effect on the median values and spread for SKESA regarding the amount of contigs at coverage 100x (Fig 8A) while having a positive effect on the median values for SKESA regarding genome fraction (Fig 8B). Other metrics and coverages were not affected, see appendix C figure C1. For SPAdes –isolate however, Pilon had an effect on the amount of misassemblies where the medians were negatively affected as they were raised from zero to 1 misassembly when Trimmomatic or no trimming was used beforehand. This can be seen at other coverages as well, see appendix B. Additionally the same effect can be seen for spades –careful, see appendix D

Pilon did however have positive effects, albeit small, on the cgMLST results for both SPAdes and SKESA. The number of corrections made by Pilon was high for most pipelines, with the exceptions where it was used on assemblies generated with SPAdes –isolate by untrimmed reads as well as when used on assemblies generated with SKESA by reads trimmed with Trimmomatic (Fig 9).

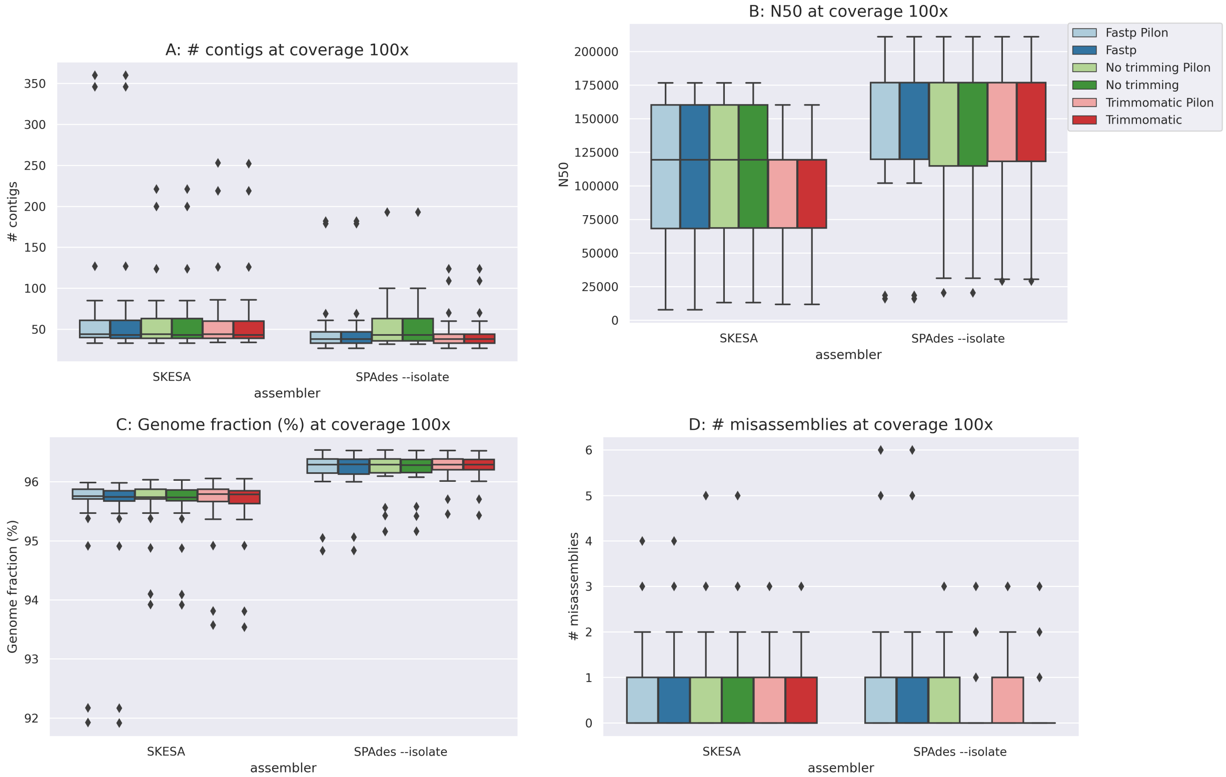


Figure 8: QUAST metrics for SPAdes and SKESA assemblies improved by Pilon at coverage 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

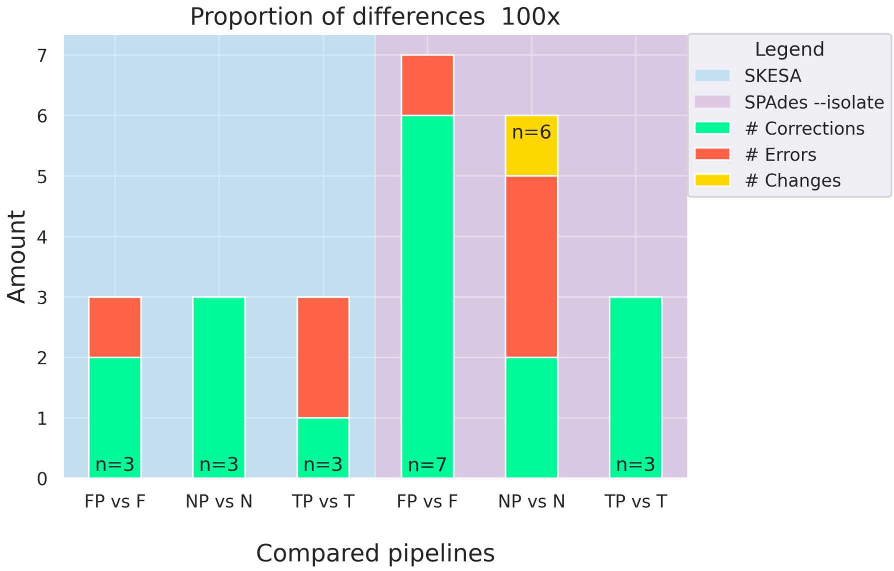


Figure 9: Differences in allele calling between assemblies which had not been improved by Pilon and assemblies improved by Pilon. n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive). Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple) and the y-axis is the amount of each change. On the x-axis is software. FP vs F = the pipeline including Fastp+Pilon compared to the pipeline including Fastp only. NP vs N = the pipeline including No trimming+Pilon compared to the pipeline including No trimming only. TP vs T = the pipeline including Trimmomatic+Pilon compared to the pipeline including Trimmomatic only.

A filtering step was used to remove small contigs in case there was left over DNA from a different organism in the sequencing machine which got assembled. Filtering had no discernible effect on the QUAST metrics for neither SKESA or SPAdes (including –careful) while having an effect on the cgMLST results. Filtering had no effect on SKESA assemblies while for SPAdes, filtering introduced a few corrections when no trimming had been done beforehand (Fig 10). When filtering contigs of length 500, most combinations were negatively or neutrally affected for both SPAdes –isolate and SPAdes –careful for all coverages with the exception for pipelines including trimming and SPAdes –careful at coverage 100x, see appendix E figure E1, where effect filtering had was mostly positive.

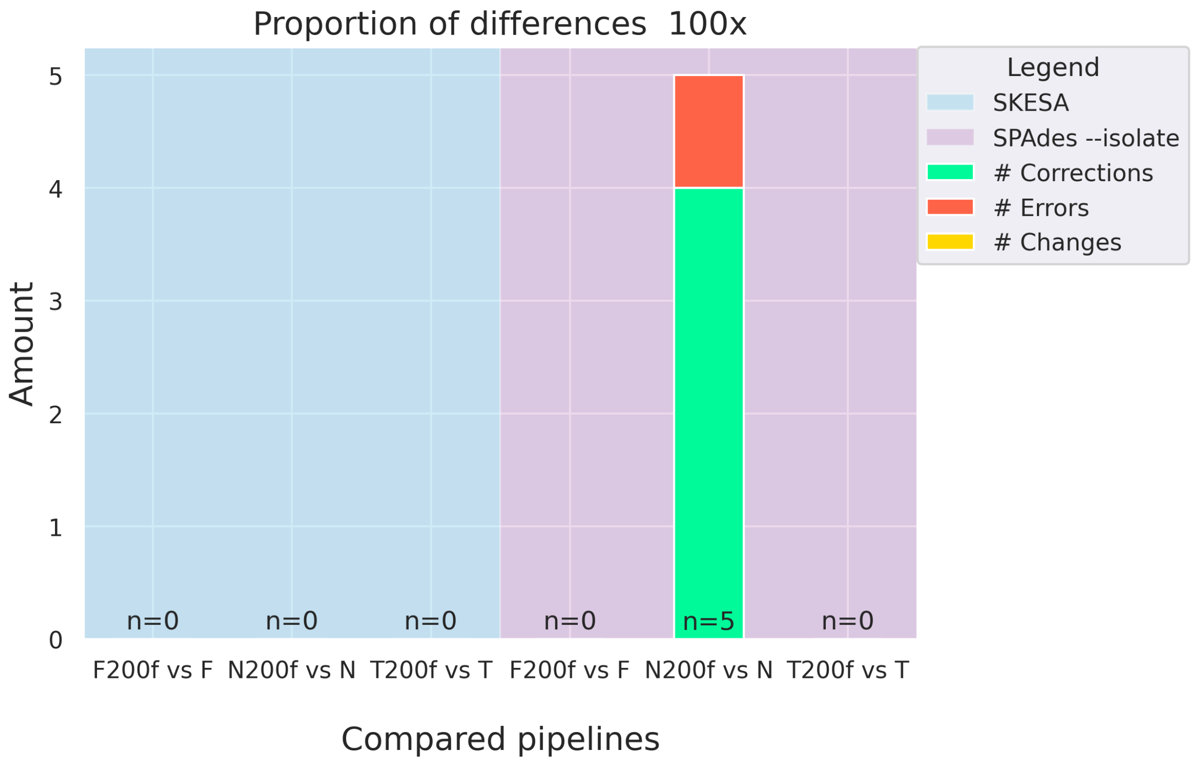


Figure 10: Differences in allele calling between assemblies without a contig size filter of 200 and assemblies with a contig size filter of 200. n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive). Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple) and the y-axis is the amount of each change. On the x-axis is software. FP vs F = the pipeline including Fastp+filtering of size 200 compared to the pipeline including Fastp only. N200f vs N = the pipeline including No trimming + filtering of size 200 compared to the pipeline including No trimming only. T200f vs T = the pipeline including Trimmomatic + filtering of size 200 compared to the pipeline including Trimmomatic only.

When comparing pipelines with both filtering and Pilon to pipelines with only one of the post assembly improvements, Pilon together with filtering introduced more corrections than when only filtering was used. Similarly when Pilon and filtering were used together more corrections were introduced compared to when only Pilon was used. However there are a few exceptions. See appendix E figure E1 and appendix C figure C1.

4 Discussion

Since *Campylobacter* is the leading cause of gastroenteritis in humans it is of great importance to be able to properly classify different strains/sequence types for the surveillance of the organism to avoid larger outbreaks. In this project 38 typically used WGS pipelines, encompassing software used for assembling and typing *Campylobacter*, have been compared using read data of varying quality from *Campylobacter jejuni*. A higher coverage produced higher quality assemblies in general and less deviance in the allele calling between the assemblies and a reference (deemed correct) as well as less deviance in allele calling between different pipelines. Coverages of 20x gave significantly worse results compared to 50x and 100x for all pipelines.

Trimming before assembling with SKESA had no obvious benefit regarding QUAST output statistics as most metrics were negatively affected when trimming, i.e worse median or spread in values. Exceptions where trimming might have positive effects are amount of contigs (coverage 50x and 100x) and genome fraction (coverage 100x). What is more, trimming even had detrimental effects on the cgMLST analysis as the error rates were significantly higher than the correction rate. This coupled with the fact that SKESA assemblies generated with untrimmed reads differ less from the reference than SKESA assemblies generated by trimmed reads give strong indications to there being no benefit to trimming reads before assembling with SKESA. This could be because SKESA has an inbuilt trimming system and thus there is no need for any pre-trimming by external trimming software. This observation is of importance since it might be a common occurrence to trim reads as a force of habit if trimming is a standardized part of ones pipeline. Regarding post assembly improvements, filtering contigs of size 200 only had an effect on the cgMLST analysis for one pipeline containing no trimming and SPAdes –isolate. Larger positive effects can be seen for the cgMLST analysis when contigs of size 500 were removed, especially at coverage 100x. However in the grand scale of things, there were few loci differences overall when taking into account that there are a total of 16,950 possible loci that could differ considering I summed all assemblies' differences from the reference. Furthermore, Pilon generally improved the cgMLST analysis compared to pipelines where it was not used. However in the grand scale of things, the amount of differences is small when taking into account that there are a total of 16,950 possible loci that could differ considering I summed together all assemblies' differences from the reference. SKESA achieves significantly better results at coverages of 100x which concurs with a previous study, even though it focused on a different organism, where the authors found that SKESA needed a higher coverage than 50x to give satisfactory cgMLST results (Liu *et al.* 2021).

Furthermore, trimming reads before using SPAdes give better results than trimming reads before using SKESA, regardless of coverage, both in regards to QUAST output statistics and the cgMLST analysis. Not trimming reads before using SPAdes gave rise to assemblies which also performed better than SKESA assemblies based on untrimmed reads in regards to the QUAST output statistics at all coverages as well as the cgMLST analysis at coverages of 20x and 50x (Fig 3.

The two SPAdes settings –isolate and –careful were compared to each other due to the SPAdes manual recommending both while they are incompatible with each other. SPAdes –isolate is recommended for high coverage isolates while SPAdes –careful is used to minimize the number of mismatches and insertions/deletions. Generally, SPAdes –isolate seemed to benefit the most from Trimmomatic as the correction rates in the cgMLST analysis were higher than for Fastp. An exception to this could be seen at coverages of 50x where Fastp produced better results, both in the cgMLST analysis and the box plots illustrating the QUAST metrics. The reason for why SPAdes –isolate performed worse than SPAdes –careful could be because of stochastic problems leading to the software not working ideally at certain coverages. This concurs with the manual which recommends the setting to be used at high coverages. SPAdes –careful could be used instead of SPAdes –isolate if the read coverage is around 50x and SPAdes –careful is paired with Trimmomatic. It is less clear for SPAdes –careful which trimming option is best based solely on the QUAST metrics since both trimming options had strengths and weaknesses. However when considering the fact that Trimmomatic led to more corrections in the cgMLST analysis than Fastp, Trimmomatic might be the best trimming tool for SPAdes –careful as well. Regarding post assembly improvements, Pilon had a generally positive effect on SPAdes –isolate assemblies while for –careful it was generally positive at coverages of 50x and 100x. Pilon should be avoided for SPAdes –isolate assemblies generated from untrimmed reads at all coverages as well as –careful assemblies at low coverages. However since the amount of differences is relatively small, especially at coverages of 50x and 100x, Pilon does not have a major impact on the cgMLST analysis as a whole, although what little effect it has is generally positive. Generally positive effects can be seen for the cgMLST analysis when contigs of size 500 were removed, especially at coverage 100x, but again the amount of differences is relatively small.

As of today, most studies investigating parameters which affect cgMLST analysis have not investigated any pre-processing of reads or post-processing of assemblies, but have instead focused more deeply on coverage, read length and assembler (Liu *et al.* 2021; Palma *et al.* 2022). As such this project might provide a broader overview of what parameters are of importance when setting up pipelines allowing for further in depth investigation based on the results found here in the future.

4.1 Pipeline recommendations

The possible recommendations that could be made for *Campylobacter jejuni* based on this project are the following: Avoid coverages of around 20x, trim reads before assembling with SPAdes and use the setting –careful if coverage is around 50x, do not combine SKESA with trimming software. Pilon can be used on SPAdes assemblies generated from trimmed reads while Pilon can be used on SKESA assemblies generated from untrimmed reads. Whether or not the filtering should be used is difficult to say without further studies investigating what effects the removal of small contigs have on the cgMLST analysis. An alternative to filtering could be to change the index on the sequencing machine so that the new sequencing run has different IDs than the previous run. That way one could possibly avoid contaminations without filtering.

It seems like Trimmomatic gave better cgMLST results compared to Fastp. However it is more difficult to give recommendations on specific trimming software since both Fastp and Trimmomatic in theory could achieve very similar results by tweaking the settings for each software. There is no guarantee that the settings used in this project are optimal for either software, perhaps the settings could be tweaked to fit this data set better or even fit individual read data better. However due to time limits and the scope of the project this could not be investigated.

4.2 Study limitations and future studies

One study limitation is that assemblies were not investigated individually in the cgMLST analysis as this would have generated too much data. Instead the amount of loci which deviated from the reference for every assembly were summed up within each pipeline. A consequence of this is that there are a few assemblies with more loci differences than others which skews the total sum of all assemblies' differences for every pipeline towards a larger deviance, even though the majority of assemblies differ very little from the reference. If instead each individual assembly was considered, it might have been possible to give recommendations on how to set up pipelines for read data of worse quality to improve both assemblies and cgMLST results. It would also have been possible to investigate how the wet-lab part of the analysis possibly could affect the results since there are various different DNA extraction kits, library preparation kits and sequencing machines etc.

A study which investigated pipelines for SNP-analysis found that each pipeline had variation in performance in regards to different species (Pearce *et al.* 2018). Thus it would be interesting to expand this project to a different genus, for example Salmonella which is another food borne pathogen which causes gastroenteritis in humans. Because of the small genome of *Campylobacter* and small number of repeats, the complexity level for assembling the genome is reduced compared to larger genomes with higher levels of repeats. Due to the fact that organisms of separate genera differ in genome sizes and the amount of repeats, it would be beneficial to try the different pipelines with other organisms to see how robust the pipelines are and what pipelines are best suited for specific organisms. Salmonella is a good candidate since the INNUENDO project has cgMLST schemes for Salmonella.

4.3 Conclusions

How WGS pipelines are set up have effects on assembly quality as well as cgMLST results. Thus it is possible to give recommendations on how WGS pipelines for *Campylobacter jejuni* should be arranged. To optimize both assembly quality as well as cgMLST results, a coverage of at least 50x should be obtained if assembling with SPAdes, while a higher coverage is needed for SKESA. Pipelines incorporating SKESA as assembler should not include a trimming software as the quality of the cgMLST analysis worsened when SKESA assemblies were trimmed with either Fastp or Trimmomatic. Pipelines including SPAdes should include a trimming software since the pipelines produced higher quality assemblies as well as less deviance in allele calling from a reference when read trimming had been performed before assembling. The SPAdes settings –isolate and –careful yield similar results except at coverages of 50x where –careful exclusively should be used. Furthermore, adding the software Pilon to the pipelines is recommended. It is more difficult to give recommendations on specific trimming software as there are an immense plethora of settings which could affect assemblies and cgMLST analysis both positively and negatively, regardless of software chosen. It is worth keeping in mind that these recommendations do not necessarily extend to other organisms, however this is worth investigating in future studies.

5 Acknowledgements

I would like to thank my supervisor Bo Segerman for giving me the opportunity to work on this project as well as for providing me with all the resources and support I needed as well for guiding me throughout the project. I would also like to thank my subject reviewer Lisa Klasson for giving me constructive and supportive feedback on my work.

References

Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. 2012. FastQC. Babraham Institute.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34: i884–i890.

Cody AJ, McCarthy ND, Jansen van Rensburg M, Isinkaye T, Bentley SD, Parkhill J, Dingle KE, Bowler ICJW, Jolley KA, Maiden MCJ. 2013. Real-Time Genomic Epidemiological Evaluation of Human Campylobacter Isolates by Use of Whole-Genome Multilocus Sequence Typing. Journal of Clinical Microbiology 51: 2526–2534.

Del Fabbro C, Scalabrin S, Morgante M, Giorgi FM. 2013. An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis. PLoS ONE 8: e85024.

Deneke C, Uelze L, Brendebach H, Tausch SH, Malorny B. 2021. Decentralized Investigation of Bacterial Outbreaks Based on Hashed cgMLST. Frontiers in Microbiology 12.

Dingle KE, Colles FM, Wareing DRA, Ure R, Fox AJ, Bolton FE, Bootsma HJ, Willems RJL, Urwin R, Maiden MCJ. 2001. Multilocus Sequence Typing System for Campylobacter jejuni. Journal of Clinical Microbiology 39: 14–23.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32: 3047–3048.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–1075.

Jolley KA, Maiden MC. 2010. BIGSdb: Scalable analysis of bacterial genome variation at the population level. BMC Bioinformatics 11: 595.

Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. 2016. Implementation of Whole Genome Sequencing (WGS) for Identification and Characterization of Shiga Toxin-Producing Escherichia coli (STEC) in the United States. Frontiers in Microbiology 7.

Liu YY, Chen BH, Chen CC, Chiou CS. 2021. Assessment of metrics in next-generation sequencing experiments for use in core-genome multilocus sequence type. PeerJ 9: e11842.

Llarena AK, Ribeiro-Gonçalves BF, Nuno Silva D, Halkilahti J, Machado MP, Da Silva MS, Jaakkonen A, Isidro J, Hämäläinen C, Joenperä J, Borges V, Viera L, Gomes JP, Correia C, Lunden J, Laukkanen-Ninios R, Fredriksson-Ahomaa M, Bikandi J, Millan RS, Martinez-Ballesteros I, Laorden L, Mäesaar M, Grantina-Ievina L, Hilbert F, Garaizar J, Oleastro M, Nevas M, Salmenlinna S, Hakkinen M, Carriço JA, Rossi M. 2018. INNUENDO: A cross-sectoral platform for the integration of genomics in the surveillance of food-borne pathogens. EFSA Supporting Publications 15: 1498E. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2903/sp.efsa.2018.EN-1498.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proceedings of the National Academy of Sciences 95: 3140–3145. Publisher: Proceedings of the National Academy of Sciences.

Palma F, Mangone I, Janowicz A, Moura A, Chiaverini A, Torresi M, Garofolo G, Criscuolo A, Brisse S, Di Pasquale A, Cammà C, Radomski N. 2022. In vitro and in silico parameters for precise cgMLST typing of Listeria monocytogenes. BMC Genomics 23: 235.

Parkhill J, Wren BW, Mungall K, Ketley JM, Churcher C, Basham D, Chillingworth T, Davies RM, Feltwell T, Holroyd S, Jagels K, Karlyshev AV, Moule S, Pallen MJ, Penn CW, Quail MA, Rajandream MA, Rutherford KM, van Vliet AHM, Whitehead S, Barrell BG. 2000. The genome sequence of the food-borne pathogen Campylobacter jejuni reveals hypervariable sequences. Nature 403: 665–668. Number: 6770 Publisher: Nature Publishing Group.

Payne M, Kaur S, Wang Q, Hennessy D, Luo L, Octavia S, Tanaka MM, Sintchenko V, Lan R. 2020. Multilevel genome typing: genomics-guided scalable resolution typing of microbial pathogens. Eurosurveillance 25: 1900519.

Pearce ME, Alikhan NF, Dallman TJ, Zhou Z, Grant K, Maiden MC. 2018. Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. International Journal of Food Microbiology 274: 1–11.

Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020. Using SPAdes De Novo Assembler. Current Protocols in Bioinformatics 70: e102. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpbi.102.

Schwartz DC, Cantor CR. 1984. Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. Cell 37: 67–75.

Segerman B. 2020. The Most Frequently Used Sequencing Technologies and Assembly Methods in Different Time Segments of the Bacterial Surveillance and RefSeq Genome Databases. Frontiers in Cellular and Infection Microbiology 10: 527102.

Sharma-Kuinkel BK, Rude TH, Fowler VG. 2016. Pulse Field Gel Electrophoresis. Methods in molecular biology (Clifton, N.J.) 1373: 117–130.

Silva M, Machado MP, Silva DN, Rossi M, Moran-Gilad J, Santos S, Ramirez M, Carriço JA. 2018. chewBBACA: A complete suite for gene-by-gene schema creation and strain identification. Microbial Genomics 4.

Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biology 19: 153.

SVA. 2020a. Surveillance of infectious diseases in animals and humans in Sweden 2020 146.

SVA. 2020b. Proficiency test number 28 whole genome sequencing of campylobacter .

Thacker SB. 1988. A METHOD FOR EVALUATING SYSTEMS· OF EPIDEMIOLOGICAL SURVEILLANCE• 8.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE 9: e112963. Publisher: Public Library of Science.

Yan S, Zhang W, Li C, Liu X, Zhu L, Chen L, Yang B. 2021. Serotyping, MLST, and Core Genome MLST Analysis of Salmonella enterica From Different Sources in China During 2004–2019. Frontiers in Microbiology 12: 688614.

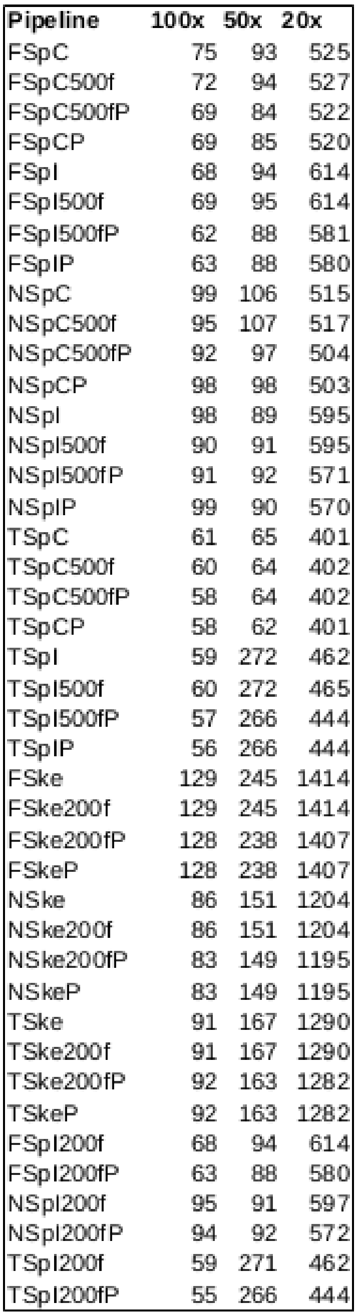
Appendix A

Appendix showing the denotations for each software which makes up each pipeline's name as well as the total number of allele differences from the reference for each pipeline.

Table A1: Analysis steps in pipelines and denotations

Software	Denotation
SPAdes –isolate	SpI
SPAdes –careful	SpC
SKESA	Ske
Trimmomatic	T
Fastp	F
No trimming	N
Pilon	P
Filtering contigs of size 200	200f
Filtering contigs of size 500	500f

Table A2: The amount of alleles which differed from the reference for all pipelines at all coverages



Pipeline	100x	50x	20x
FSpC	75	93	525
FSpC500f	72	94	527
FSpC500fP	69	84	522
FSpCP	69	85	520
FSpI	68	94	614
FSpI500f	69	95	614
FSpI500fP	62	88	581
FSpIP	63	88	580
NSpC	99	106	515
NSpC500f	95	107	517
NSpC500fP	92	97	504
NSpCP	98	98	503
NSpI	98	89	595
NSpI500f	90	91	595
NSpI500fP	91	92	571
NSpIP	99	90	570
TSpC	61	65	401
TSpC500f	60	64	402
TSpC500fP	58	64	402
TSpCP	58	62	401
TSpI	59	272	462
TSpI500f	60	272	465
TSpI500fP	57	266	444
TSpIP	56	266	444
FSke	129	245	1414
FSke200f	129	245	1414
FSke200fP	128	238	1407
FSkeP	128	238	1407
NSke	86	151	1204
NSke200f	86	151	1204
NSke200fP	83	149	1195
NSkeP	83	149	1195
TSke	91	167	1290
TSke200f	91	167	1290
TSke200fP	92	163	1282
TSkeP	92	163	1282
FSpI200f	68	94	614
FSpI200fP	63	88	580
NSpI200f	95	91	597
NSpI200fP	94	92	572
TSpI200f	59	271	462
TSpI200fP	55	266	444

Appendix B

This appendix shows QUAST metrics for all coverages for SPAdes and SKESA comparisons.

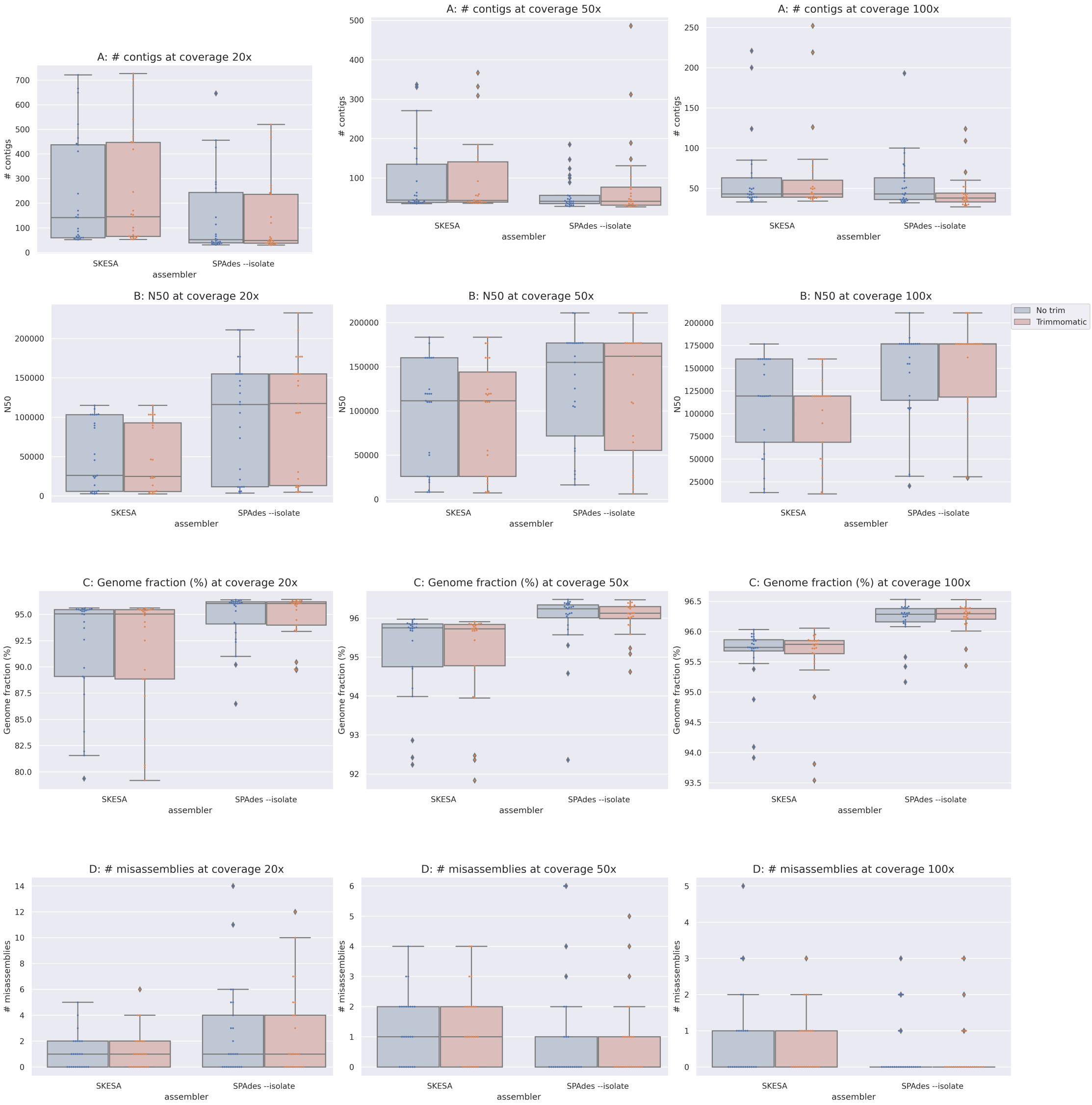


Figure B1: QUAST metrics for SPAdes and SKESA assemblies generated by trimmed reads (red) and untrimmed reads (blue) at coverages of 20x, 50x and 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

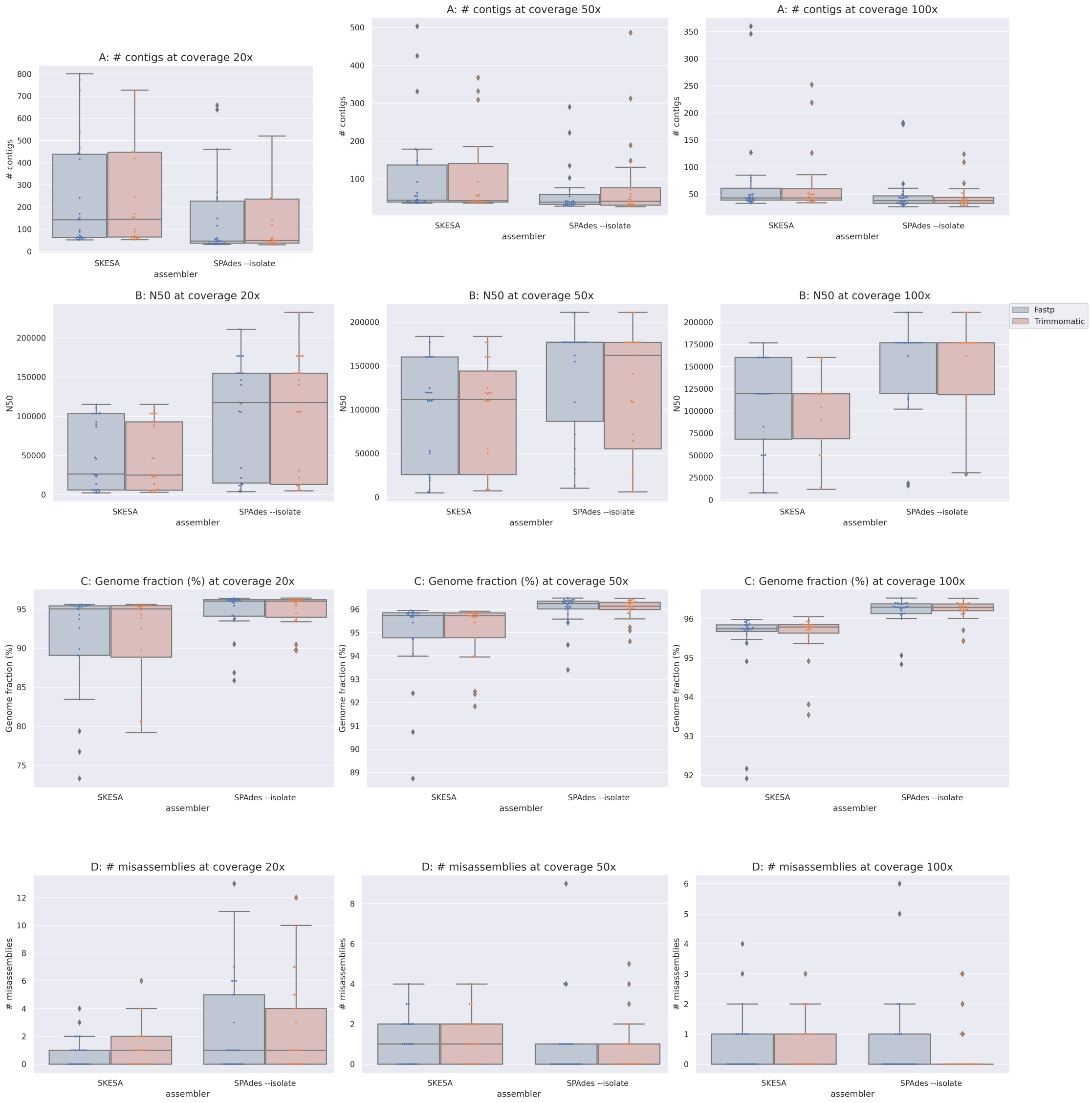


Figure B2: QUAST metrics for SPAdes and SKESA assemblies generated by trimmed reads with Trimmomatic (red) and reads trimmed with Fastp (blue) at coverages of 20x, 50x and 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

Appendix C

This appendix shows the chewBBACA comparisons for all coverages when SPAdes is compared to SKESA.

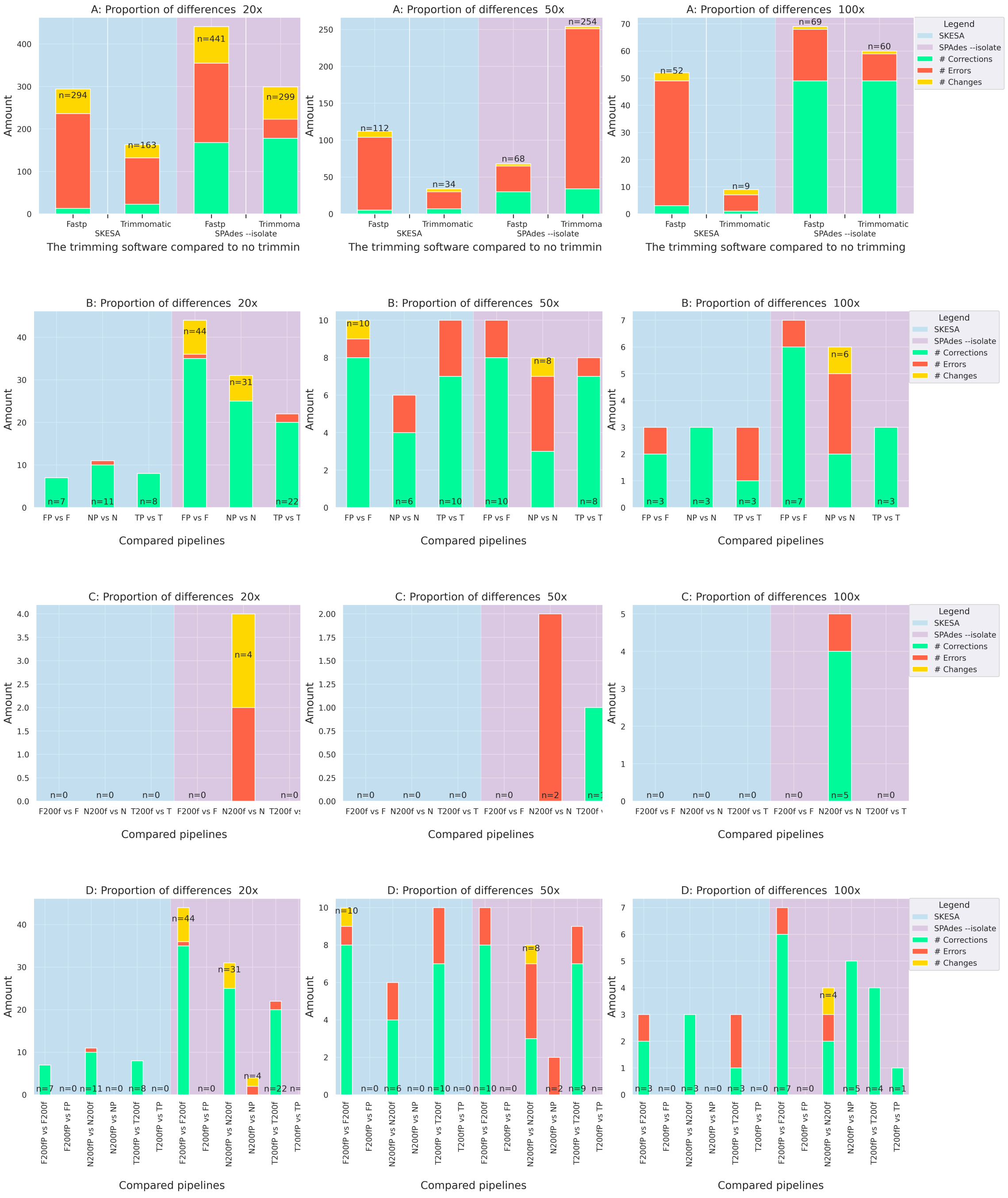


Figure C1: Differences in allele calling between different pipelines. n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive).Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). On the x-axis is software, y-axis is the amount of each change. The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple). ***(A)*** *The difference between Trimmomatic and no trimming as well as Fastp and no trimming.* ***(B)*** *The difference between filtering contigs of size 200 an no filtering for all trimming options.* ***(C)*** *The difference between using pilon an not using pilon for all trimming options.* ***(D)*** *The difference using pilon together with filtering for all trimming options.*

Appendix D

This appendix shows the QUAST metrics for all coverages when SPAdes –careful and SPAdes –isolate are compared.

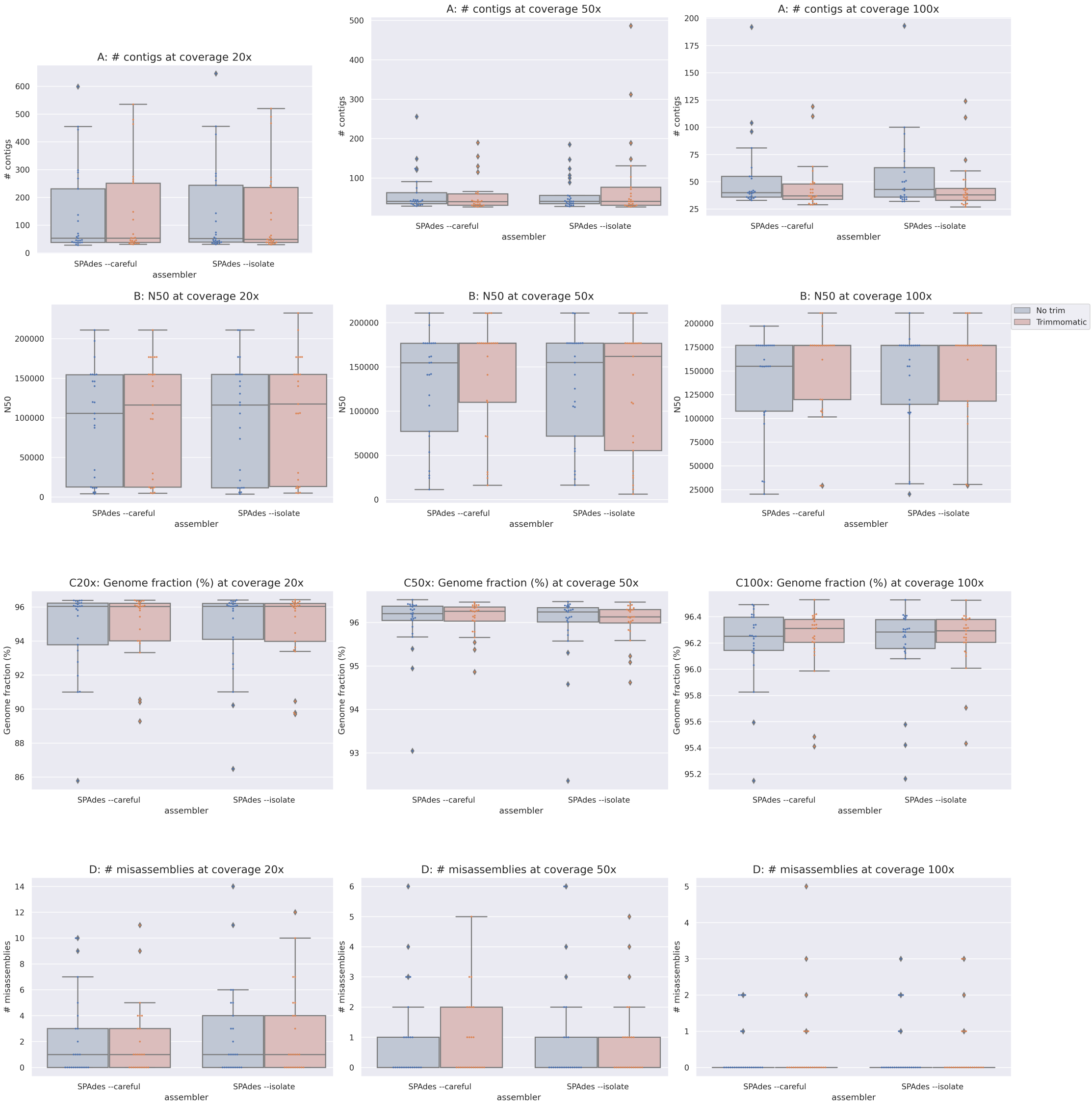


Figure D1: QUAST metrics for SPAdes –isolate and SPAdes –careful assemblies generated by reads trimmed with Trimmomatic (red) and untrimmed reads (blue) at coverages of 20x, 50x and 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

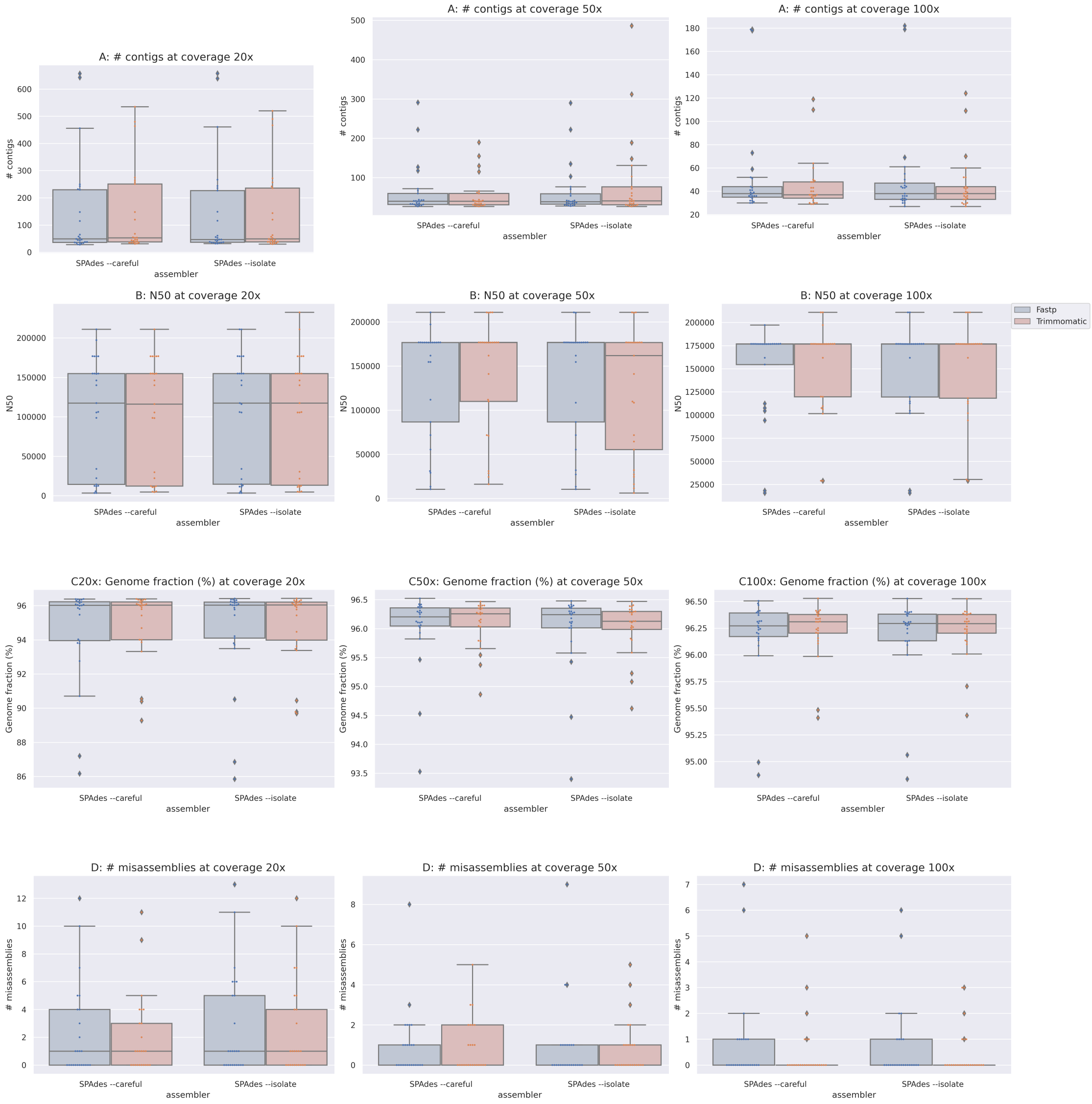


Figure D2: QUAST metrics for SPAdes –isolate and SPAdes –careful assemblies generated by reads trimmed with Trimmomatic (red) and reads trimmed with Fastp (blue) at coverages of 20x, 50x and 100x. The box represents the span av values for 50% of assemblies. The middle line inside the box represents the median value. The whiskers extending from the box plots each represent the span of values for 25% of the data points, with the bottom line being the minimum value and the top line being the maximum value. ***(A)*** *Number of contigs* ***(B)*** *N50 values* ***(C)*** *Genome fraction* ***(D)*** *Number of misassemblies (i.e number of relocations, translocations or inversions)*

Appendix E

chewBBACA comparison all coverages for SPAdes –isolate vs SPAdes –careful

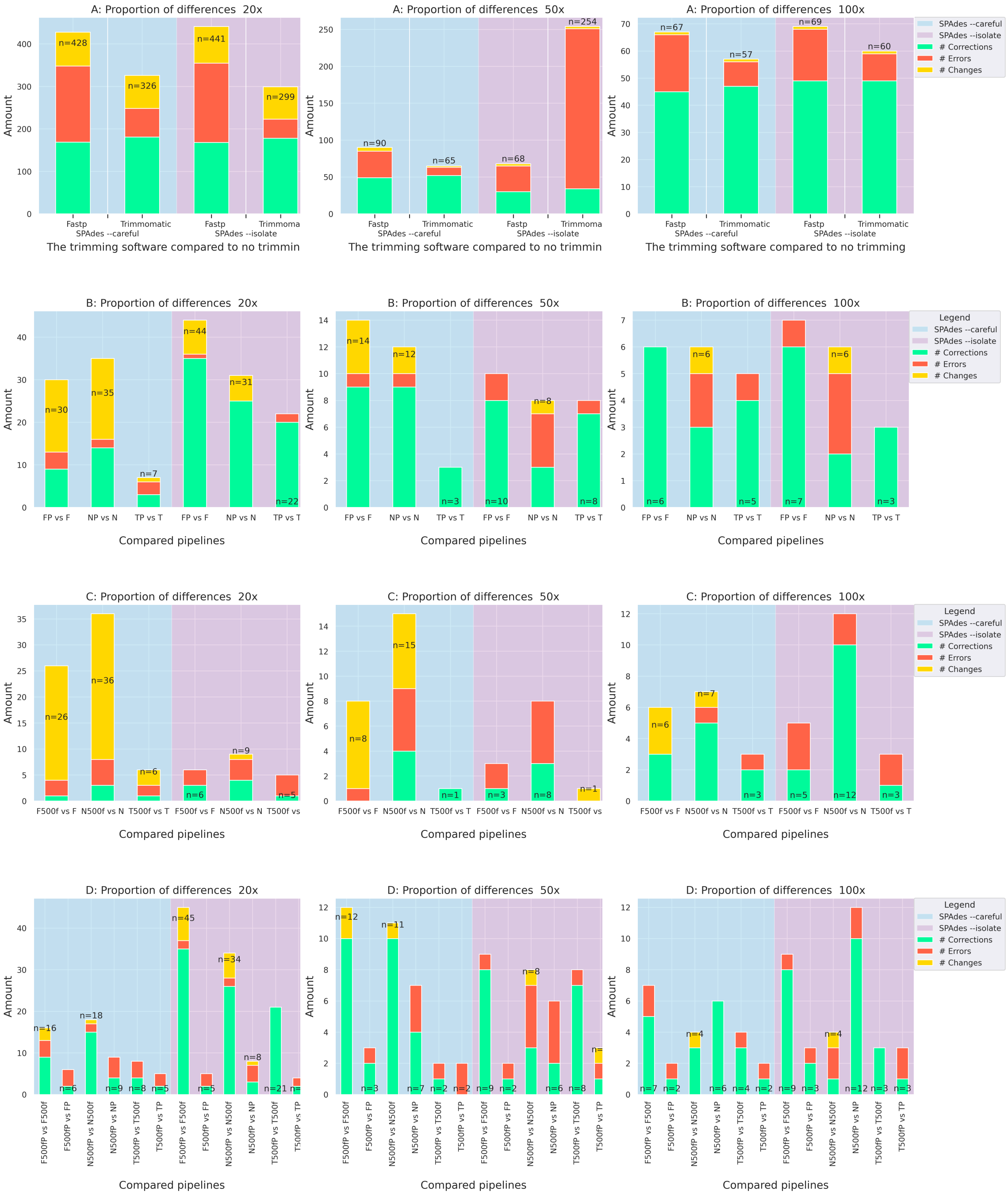


Figure E1: Differences in allele calling between different pipelines. n is the total amount of different alleles observed. Corrections are the amount of loci which were corrected after trimming without post assembly corrections (positive).Errors are the amount of loci where trimming introduced an error not found in the allele calling of the untrimmed assembly (negative). The changes are changes from one error to another error (neutral). On the x-axis is software, y-axis is the amount of each change. The plots are divided into SKESA (light blue) and SPAdes –isolate (light purple). ***(A)*** *The difference between Trimmomatic and no trimming as well as Fastp and no trimming.* ***(B)*** *The difference between filtering contigs of size 200 an no filtering for all trimming options.* ***(C)*** *The difference between using pilon an not using pilon for all trimming options.* ***(D)*** *The difference using pilon together with filtering for all trimming options.*