

Genetic differentiation of adaptively host- shifted populations of *Rhagoletis* *cerasi*: incipient sympatric speciation?



Figure 1: Photo of *Rhagoletis cerasi* (<http://www.entomart.be/INS-0848.html>).

Dwin Grashof
S1082225
Bin3A

Hogeschool Leiden
Mentor: Jan Oliehoek
Internship: Naturalis Biodiversity Center
Supervisor: Rutger Vos

Contents

Samenvatting	3
Introduction	4
Speciation.....	4
Selection.....	5
<i>Rhagoletis cerasi</i>	6
This study	6
Objectives	7
Materials & Methods	8
Data	8
Pre-processing.....	8
Assembly	9
Alignment.....	9
BLAST.....	10
Flowchart	11
Results	12
Pre-processing.....	12
Assembly	15
Alignment.....	17
BLAST.....	17
Discussion	20
Data	20
Assembly	20
BLAST and Filtering.....	20
Conclusion	21
Further Studies	22
References	23

Samenvatting

In dit onderzoek werd gezocht naar genetische verschillen tussen twee individuen uit twee populaties van de Europese kersenboorvlieg, *Rhagoletis cerasi*. Deze verschillen kunnen aangeven dat er sympatrische soortvorming plaats vindt tussen deze twee populaties.

De bekende en niet-genetische verschillen zijn dat de ene populatie op een inheemse kamperfoelie zit en de andere populatie op een exotische kamperfoelie die in 1700 in Nederland is geïntroduceerd. Dit zou er voor kunnen zorgen dat de twee populaties uit elkaar groeien, bijvoorbeeld omdat hybriden een slechte overlevingskans hebben.

Naar deze verschillen zijn gezocht doormiddel van een zelf ontwikkelde pipeline. Deze pipeline bestaat uit twee grote stappen:

- 1) Pre-processing: In deze stap worden de illumina paired-end sequencing reads op kwaliteit gecontroleerd en worden de reads op basis van deze kwaliteit bijgesneden en/of verwijderd. Ook wordt er op adaptersequenties gezocht en deze worden eventueel weg gehaald.
- 2) Assembly: Alle files worden samen gebuikt om een pseudo-referentie genoom te maken van de *Rhagoletis cerasi*.

De reads zijn succesvol verwerkt en gecontroleerd. Er zijn geen adaptersequenties gevonden en de gemiddelde kwaliteit van de reads is nauwelijks vooruit gegaan in stap 1) omdat de reads al een hoge kwaliteitsscore hadden van gemiddeld 38.

De assembly zorgde echter voor slechtere resultaten. De gemiddelde contig lengte is 164 bp en de gemiddelde scaffold lengte is 262 bp (de reads waren 100 bp lang). Naast de standaard methode met een vooraf berekende k-mer lengte is ook geprobeerd andere k-mer lengtes te gebruiken. Ook is geprobeerd een enkel individu afzonderlijk te assembleren. Deze resultaten waren echter niet beter dan de eerste methode. Vervolgens is geprobeerd alle reads tegen een referentie genoom afkomstig van *Ceratitis capitata* te alignen. Dit leverde slechts een alignment-rate van 18% op.

Om de data nader te bekijken zijn contigs met een lengte van 500 bp geBLAST tegen de genoom database van NCBI. De resultaten duiden er op dat de data vervuild is met ander DNA. Ongeveer 50% is afkomstig van een bacterie en maar 35% afkomstig van de *Rhagoletis cerasi*.

Om deze vervuiling weg te halen is er een lokale BLAST uitgevoerd tegen een database bestaand uit het genoom van de evolutionair verwante vlieg *Ceratitis capitatae*. Na gefilterd te hebben op een e-value van $1e^{-10}$ waren nog maar 1% van de oorspronkelijke contigs over. De statistieken, zoals de gemiddelde lengte, van deze contigs zijn wel veel beter en dat maakt de contigs betrouwbaarder.

Uit deze nieuwe lijst met contigs zijn contigs gepakt met een lengte van 500 bp. Deze zijn vervolgens opnieuw tegen de database geBLAST om te kijken of de vervuiling is afgenomen. Volgens deze resultaten is de vervuiling afgenomen omdat nu ongeveer 75% van de data afkomstig is uit insecten. Deze nieuwe lijst met contigs kan vervolgens gebruikt worden voor vervolg onderzoek.

Dus ondanks de vervuiling van vreemd DNA kan nu verder gegaan worden met het zoeken naar genetische verschillen tussen de twee populaties om aan te tonen dat er mogelijk sympatrische soortvorming aan de orde is.

Introduction

Speciation

Speciation is the process of one species becoming two species. Different modes of speciation are recognized.

These modes, along with their cause, are shown in figure 2.

Allopatric speciation may occur after a change in the environment that separated populations and thus stopped gene flow between them.

Peripatric speciation is a special kind of allopatric speciation that happens when one of the isolated populations has very few individuals.

These few individuals may have a different genepool containing rare genes compared to the genepool of the main population. After a few generations all the individuals of the smaller population may have obtained these rare genes purely by drift. This combined with natural and sexual selection may change the population into a new species. In parapatric and sympatric speciation there is no specific extrinsic barrier to gene flow. Parapatric speciation occurs when there is a lack of random mating in a species. Individuals are more likely to mate with a specific individual rather than a random mate. This may cause divergence because of reduced gene flow between populations. Sympatric speciation may occur when individuals separate themselves

from the population. They can do that by “trying out” something new in contrast to the rest of the population. This may occasionally happen when, for example, herbivorous insects “try out” , that is, end upon a new host plant in the same geographic area. These insects then develop a preference for the new host plant (and for individuals with the same preference). This reduces gene flow between the two populations. These may be the first steps toward sympatric speciation.

However, gene flow between the two populations does not have to be reduced for sympatric speciation to happen. In some occasions speciation with-gene-flow occurs. This rarely happens, because the strength of selection has to overcome the homogenizing effect of gene flow.^[1]





Mode of speciation	Cause of speciation	
Allopatric	Geographically isolated populations	
<u>Peripatric</u>	A small population isolated at the edge of a larger population	
<u>Parapatric</u>	A continuously distributed population	
Sympatric	within the range of the ancestral population	

Figure 1: This figure shows the four different kinds of speciation: Allopatric, peripatric, , parapatric and sympatric speciation.^[1]

Selection

Natural selection can occur in a few different ways as shown in Figure 3. Figure 3 a) illustrates directional selection. This is where evolution focusses on a certain trait value benefitted from in nature and selects

towards it, thereby changing the mean trait value of that population. For example, when a bird population has intermediately-sized beaks with which they eat seeds, but then seeds of this size become rare. They may come under selective pressure for smaller beaks so they can eat smaller seeds instead. As a consequence, the average beak size in the population drops. Figure 3b) shows stabilizing selection. This happens when there is a broad range of traits within a species, but there is an optimal intermediate trait value. This optimal value is picked up by evolution and the extremes will disappear. Imagine a bird species with a lot of variety in beak sizes between individuals. When a small beak cannot obtain food because it is too small and a big beak fails as well, natural selection will drive the species to an intermediately-sized beak. The last type of selection is shown in figure 3c).

This is disruptive/divergent selection, which may play a major role in sympatric speciation. Disruptive selection takes place when a trait gets evolutionarily pushed towards extremes (or pulled apart). Take the same bird species from the example of figure 3a. With the intermediately-sized seeds still gone the birds adapt to the change, but this time also larger seeds are available and so large beaks get selected as well. This may result in two populations, or even species. One population with a small beak and one with a large beak, both selected to eat different kinds of food.

Once Darwin hypothesised that disruptive natural selection, where organisms adapt divergently to changing environments, is largely responsible for creating new species (Darwin, C. (1859) *On the Origin of Species*). Now we know that it is not necessarily the environment that has to change. The opportunities available to the individuals in a species can also start disruptive natural selection. When disruptive selection is happening, we know that gene flow between individuals in the two trait extremes may be reduced or disappear. This, however, does not have to be the case. As stated earlier, two populations may still speciate even though gene flow never stopped.^[2]

When two diverging populations interbreed they may produce maladapted hybrids, i.e. offspring with reduced survival or mating success. This then creates impetus to evolve pre-mating isolation. This is any mechanism that prevents the interbreeding of the two populations, also called as reinforcement. Reinforcement happens because individuals from the different, diverging populations might evolve to become more discriminating with respect to individuals of the other population to prevent hybridization.^[3]

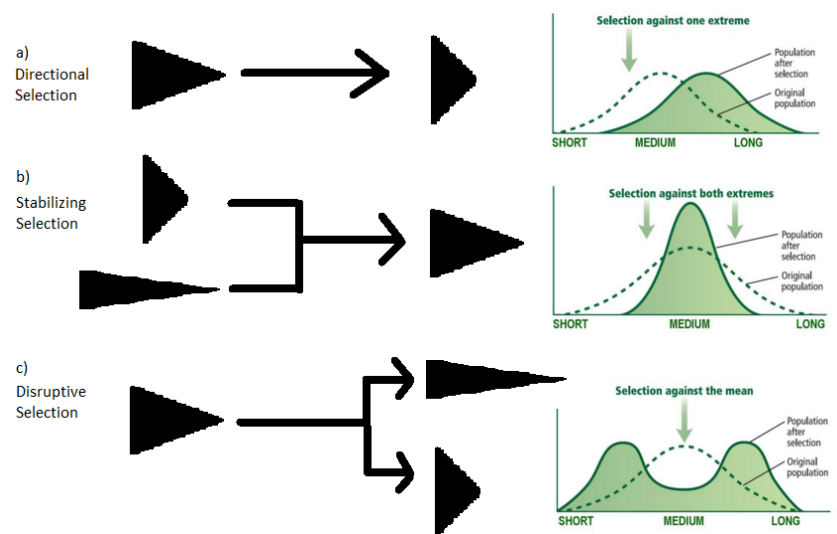


Figure 2: This is a schematic drawing of the different types of selection. A) shows directional selection, where the mean of the population changes. B) shows stabilizing selection, where the distribution becomes narrower. C) shows disruptive selection where the distribution goes from unimodal to binomial.

Rhagoletis cerasi

In this study we tried to find evidence for incipient sympatric speciation in the fruit fly *Rhagoletis cerasi*. Each individual *R. cerasi* has a preference for the host plant species it was born in. A big question is whether this preference is heritable or not. This means that not every *R. cerasi* lays her eggs in the same host plant species' fruit. In the case of our study

there is a clear difference in preference of host plant within *R. cerasi*. The difference we are interested in is preference of *R. cerasi* for two different species of honeysuckle, the Dutch native *Lonicera xylosteum* and the exotic *L. tatarica*, which was introduced to the Netherlands in mid-1700.^[4] Knowing this it is plausible to say that there was a host shift from *L. xylosteum* to *L. tatarica*.

When the larva comes out of the fruit it enters a pupa stage in the soil underneath its host plants. There the pupa overwinters to emerge as a fully-grown fly when the fruits are ripe. This fly is then part of a new generation.^[5]

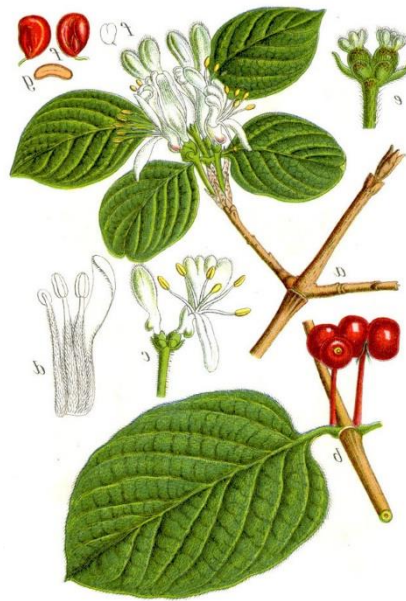


Figure 3^[6]: A) This is a schematic view of the *Lonicera xylosteum*.



B) This is a schematic view of the *Lonicera tatarica*.

This study

When there is, a constant difference in host plant preference we expect to see differences in the genome as well, even if gene flow never stopped between diverging populations. Experimental evidence for this has already been found by Feder et al.(2015) in species that are undergoing speciation with-gene-flow, for example in *R. pomonella*, a relative of the *R. cerasi*. This *R. pomonella* experienced a host shift as well, the native population lay their eggs in a hawthorn, but the host shifting population shifted to an apple tree as host plant. They found significant gene shifts and genome-wide variations within just one generation exposed to environmental change. They found these differences by comparing single-nucleotide polymorphisms (SNPs) between different populations.^[7]

According to that study the assumption can be made that there may be differences in the genomes of the *L. xylosteum R. cerasi* and the *L. tatarica R. cerasi* as well. The differences may lead to further divergence and may even lead to two different species.

The difference in the genome between both *R. cerasi* populations was planned to be analysed through a custom pipeline including the assembly of a pseudo-reference genome and subsequent SNP calling. The analysis of the SNPs was an optional part, for which there was not enough time to finish.

Objectives

In this project we looked at the genetic differences among three individuals from two populations of *Rhagoletis cerasi*, of which one individual belongs to a population that shifted to a new host plant. These individuals may be subject to disruptive ecological selection and may therefore represent the early stages of speciation-with-gene-flow.

The objective was to answer the question:

Is there genetic evidence to be found that corresponds to the separation of two populations *Rhagoletis cerasi*, which may be interpreted as the early stages of speciation-with-gene-flow?

Materials & Methods

Data

The collected data for this study consist of eighteen FASTQ files. In total, three individuals had been sequenced on the Illumina platform. Two of these three were caught on host plant XXX and the third one was caught on host plant YYY. The three sequenced individuals resulted in six FASTQ files for each individual, hence eighteen in total. These were all paired-end reads. Each organism was sequenced in two different lanes and two different runs. This was done to minimize artefacts.

Pre-processing

The first step in the custom made pipeline consisted of checking for adapter sequences and removing them when these sequences were found. Because Illumina sequences randomly one needs to add a short nucleotide sequence to the DNA fragments. These short added sequences are called adapter sequences or adapters for short. By adding these adapters to the DNA fragments they can, for example:

1. Bind to the flow cell of the sequencer to be read.
2. Be used for PCR enrichment to multiply the DNA fragments and ensure more data.
3. Be used for indexing (barcoding) of samples. This way multiple DNA libraries can be read simultaneously in the same lane. This process is called multiplexing.^[8]

These adapter sequences need to be removed from the end of the reads because they can cause confusion to different kinds of programs. This may result in a bad quality assembly. In our case, the removal of the adapter sequences was done by the program CutAdapt^[9]. This program is made for removing adapter sequences, primers and poly-A tails. Illumina uses standard adapter sequences for certain tasks. In this case Illumina used the normal “Illumina Paired End Adapter”, “Illumina Pared End PCR Primer” and “Illumina Pared End Sequencing Primer”. These three types of adapters have two different sequences. To filter the data both of them for all three adapter types were used. So in total six adapter sequences were used. (These adapter sequences were obtained from an article written in a blog by Graham Etherington^[10].) CutAdapt is designed to cut the adapter or part of the adapter off all the different reads. This program is used to clean the data received from the sequencer to optimize the outcome.

After trimming the adapter sequences the reads were trimmed to filter out low quality parts of a sequence or whole low quality reads. The trimming was done with a custom made script written in the programming language Python. This script trims in three steps (the quality scores are adjusted to the standard phred score. That means the original score minus platform specific standards):

1. The average quality score of a read is calculated first. If this average quality score is above a specified threshold (in this case a quality score of 25) the read may proceed to the next step. If a read doesn't meet the threshold the read is discarded.
2. The second step trims per base at a time. This is done from the 3' side of the read. If a base has a quality score lower than 25 the base is discarded. This is done till three bases are found with a quality score higher than the threshold of 25 in a row. Once this condition is met the trimming is stopped. Afterwards the read is reversed to trim the 5' side of the read. This side goes through the same process. Once this side is trimmed the read is reversed to its original state and passed on to step three.
3. This step discards reads that were significantly shortened through the trimming process. The reads with fewer than 30 base pairs left are discarded.

For quality control of the trimmed reads the program FastQC was used^[11]. FastQC is a Java based program written to visualise and analyse reads from high throughput sequencing pipelines. FastQC makes plots of the following statistics:

1. FastQC does basic statistics like the total number of reads, the range length of the reads and the GC content.

2. FastQC makes a boxplot where the range of the quality is set out against the position of the base pair in the read.
3. FastQC also calculates the average quality score per read. This can also be displayed in a plot.
4. FastQC also looks at the length distribution, duplication levels, overrepresented sequences and k-mer content in the dataset. This may help determine the quality of the dataset and work as a predictor for the quality of the assembly.

By using FastQC in combination with the custom made trim script the trimming process was tuned and optimized.

Assembly

To make a reference genome with the processed reads they were assembled with the program SOAPdenovo2^[12]. This is a short-read assembly tool made for *de novo* assemblies. SOAPdenovo is optimized for Illumina GA short reads and human-sized genomes. The assembly was done with all the eighteen trimmed files from the in total three individuals. By doing that the reference genome includes both populations and has a high coverage for the assembly. To make sure the optimal k-mer lengths was used in the assembly, KmerGenie was used^[13]. KmerGenie estimates the optimal k-mer length for the given reads. By calculating k-mer abundance histograms of all possible k-mers and searching for the highest distinct genomic k-mers in those histograms, the optimal k-mer length is chosen. KmerGenie runs through all the given files and calculates the best k-mer length for each file. When multiple files are used the average of those files is taken as the optimal k-mer length. This optimal k-mer length was used by SOAPdenovo for the assembly.

The other options for SOAPdenovo were left at their default values. They can be changed when needed. To check the output of the assembly, contigs consisting of five hundred base pairs were run through the Genome Database of BLAST^[14]. BLAST (Basic Local Alignment Search Tool) is a tool that finds similarities between biological sequences. Within BLAST different kinds of databases and algorithms can be selected. This tool can be used to compare nucleotide and protein sequences with themselves and each other. To check the purity of the data a standard nucleotide BLAST was used. BLAST will align all the 500 bp contigs against all the known genomes in the database. The best hits were chosen based on e-value lower than 1e-10 and percentage identity of at least 30%. The organisms belonging to the best hits should give an idea of the purity of the data.

Alignment

To improve results a different method was attempted as well. Instead of a *de novo* assembly the reads were aligned against an existing full reference genome of a related species. This way, a read related to the reference genome will align with it and may form a consensus sequence together with the other mapped reads. An advantage of this method is that only related reads are aligned. Reads can be aligned more easily and the process is much faster. A disadvantage is that some reads from the *Rhagoletis cerasi* may not be aligned because they differ too much from the reference genome. This will result in a gap in the consensus genome.

As a reference genome the genome of *Ceratitis capitata* was used. This insect is related enough to perhaps do a fairly good alignment. Two different tools were used to align. The first tool was Bowtie2^[15] and the second tool was BWA^[16]. Bowtie2 is a very fast and efficient tool for aligning against long reference genomes. It is optimised for reads around 50 to 100 base pairs and reference genomes around the length of mammals. Bowtie2 has a local alignment option that is useful for alignment with closely related genomes when there is no genome of the organism of interest itself. Bowtie2 will generate a SAM (Sequence Alignment Map) output file. This file can be processed to a reference genome using samtools^[17].

BWA (Burrows-Wheeler Aligner) is an aligner made to align low-divergent sequences against a large reference genome, for example mammalian. The BWA-MEM algorithm does a local alignment with the given reads. This algorithm is faster and more accurate compared to the other BWA algorithms. BWA-MEM is also optimised for 70-100 base pair Illumina reads. The generated output of BWA is similar to Bowtie2 and can be processed in the same way with samtools.

BLAST

To clear away the contamination as much as possible, the recovered contigs were locally BLASTed against the full genome of *Ceratitis capitata*. This was done using the blast+ toolkit made by NCBI^[18]. This toolkit can be used on a LINUX terminal. First a database of the *Ceratitis capitata* genome was made using the “makeblastdb” command of blast+. The second step was to blast the contigs against the database using the blastn program. To ensure the usability of the contigs, only those consisting of 200 base pairs or larger were included. To filter the hits automatically the maximum e-value was set to 1e-10. After the blastn step the found hits were sorted and filtered on a query coverage of at least 30%. The list of contig names of the remaining hits was then made unique (i.e. duplicates were filtered out). The contigs and contig names corresponding with the unique list were placed in a different file. By doing this the amount of non-insect contigs were reduced.

Flowchart

The whole process of this project is showed below in figure 5.

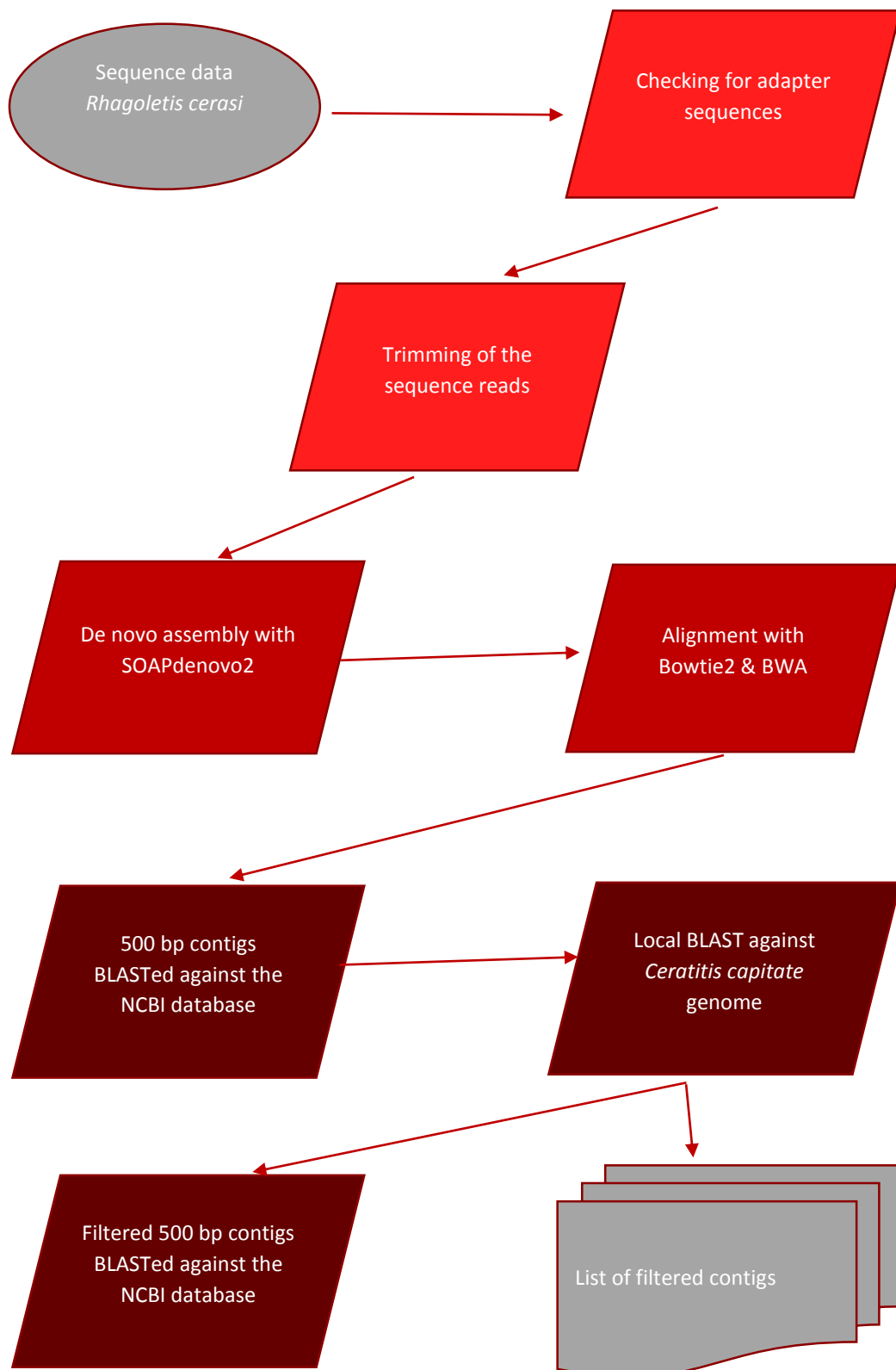


Figure 5: Flowchart of the process of this study.

Results

Pre-processing

The raw reads were loaded in FastQC as the first step. To determine the quality the following points were of importance: the number of reads per file, the average length per reads per file and the average quality score per read per file. These statistics are visualised in table 1. The first noticeable thing was that the files coming from the second run (the filenames beginning with the D) had double the amount of reads per file in comparison to their first-run file counterparts. Another thing to mention is that the files originating from the third individual had a higher read count, the amount of reads per file was almost doubled.

All the files had the same expected average read length of 100 base pairs, which is an Illumina standard. Furthermore, the average quality score of the reads had a constant value of 38-39. This is a fairly good quality score. This means that minimal trimming was required, so trimming went according to the method described in the Materials and Methods.

The first step in the pre-processing plan was to check for potential adapter sequences and remove them wherever present. The different files were all searched through by CutAdapt one by one. Each file had an adapter found percentage of 0.6%. This included partially found adapter sequences that made up most of the found adapter sequences. This might mean that there are no adapter sequences present in the data and that the hits are because of random occurrence. According to that conclusion the files were not processed further by CutAdapt.

The reads were processed by the trimming script without any problems. After this process the files were loaded in FastQC once again to see if there were any improvements. The results of this quality check are listed with the first results in table 1. The average length per read per file is now around 98-99 base pairs and the average quality score of the reads remained at 38-39. The results from before and after trimming can be compared to decide the course of action. The average length of the untrimmed reads is 2 base pairs longer than the trimmed reads, however, the average quality score was unchanged. Even though those were minor changes, the number of reads was reduced by 10%..

bf = before trimming af = after trimming QS = quality score	Description	Number of reads bf	Average length bf	Average QS bf	Number of reads af	Average length af	Average QS af
COA7AACXX_101851-02_TGACCA_L001_R1	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	15593638	100	39	14824943	99	39
COA7AACXX_101851-02_TGACCA_L001_R2	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	15593638	100	38	14824943	99	39
COA7AACXX_101851-02_TGACCA_L002_R1	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	15108254	100	39	14327871	98	39
COA7AACXX_101851-02_TGACCA_L002_R2	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	15108254	100	38	14327871	98	39
D0TRBACXX_101851-02_TGACCA_L004_R1	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	38792652	101	39	33670723	99	39
D0TRBACXX_101851-02_TGACCA_L004_R2	<u>Rhagoletis cerasi</u> , individual 1, host plant 1	38792652	101	39	33670723	99	39
COA7AACXX_101851-03_ACAGTG_L001_R1	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	14712832	100	38	14077434	99	38
COA7AACXX_101851-03_ACAGTG_L001_R2	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	14712832	100	38	14077434	99	38
COA7AACXX_101851-03_ACAGTG_L002_R1	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	14194007	100	38	13543796	98	38
COA7AACXX_101851-03_ACAGTG_L002_R2	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	14194007	100	38	13543796	98	38
D0TRBACXX_101851-03_ACAGTG_L004_R1	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	36790207	101	39	30783955	98	39
D0TRBACXX_101851-03_ACAGTG_L004_R2	<u>Rhagoletis cerasi</u> , individual 2, host plant 1	36790207	101	39	30783955	98	39
COA7AACXX_101851-06_GCCAAT_L001_R1	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	31352304	100	38	30342456	99	38
COA7AACXX_101851-06_GCCAAT_L001_R2	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	31352304	100	38	30342456	99	38
COA7AACXX_101851-06_GCCAAT_L002_R1	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	30631378	100	38	29581558	98	38
COA7AACXX_101851-06_GCCAAT_L002_R2	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	30631378	100	38	29581558	98	38
D0TRBACXX_101851-06_GCCAAT_L004_R1	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	70489665	101	38	61794911	98	38
D0TRBACXX_101851-06_GCCAAT_L004_R2	<u>Rhagoletis cerasi</u> , individual 3, host plant 2	70489665	101	38	61794911	98	38
Total		535329874	100	38	485895294	98	38

Table 1: This is an overview of the reads of all eighteen files before and after trimming.

FastQC produces a wide range of information, including plots of duplicate levels and overrepresented reads. This information can tell on forehand something about the probability of a successful assembly. This is because a duplication or overrepresented reads may interfere with the algorithm and cause problems in the created genome. The FastQC of plot the duplication levels is shown in figure 6. This figure shows a normal gradient for Illumina DNA data. The Sequence Duplication Level shown in the top of the figure is slightly higher than normal, but 20% isn't an alarming value. The other noticeable thing is the slight peak at the end of the line. This means that some reads are found significantly more than most reads. Most reads are found only one or two times, but some reads are found more than ten times in the whole file. It is common that this is seen in such a plot, but it is worth keeping in mind.

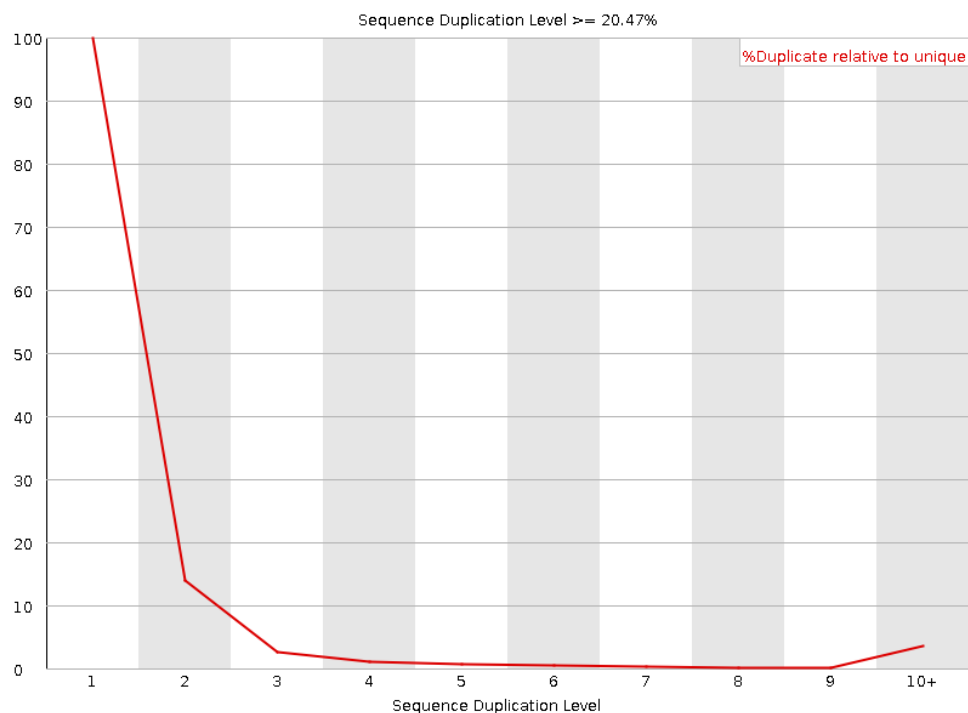


Figure 6: FastQC plot of the duplication levels of the trimmed C0A7AACXX_101851-02_TGACCA_L001_R1 file. The x-axis shows the number of times a read is found in the file. The y-axis shown the percentage of reads that are found.

To make sure that the “duplicated reads” will not give a problem with the assembly of the genome, FastQC checks for overrepresented reads. In the same file as the duplication level plot no overrepresented reads were found. This could indicate that the duplicated reads are not present in excess to the other reads according to FastQC. These FastQC checks were done to the whole dataset file by file. No file was significantly different from the file shown in figure 6.

Assembly

The next step after quality control and trimming was to do a de novo assembly with all the reads from all the populations and individuals. The optimal K-mer length for the de novo assembly was estimated by KmerGenie. All files were individually processed by KmerGenie. The results are shown in table 2. The prediction varies from a k-mer length of 17 to 35, with an average of 23. This average k-mer length of 23 was used in the de novo assembly with SAOPdenovo2. To visualize the k-mer length for each file in relation to the average a plot was made which shows the proportions (figure 7).

	Best predicted K size	Predicted assembly size
COA7AACXX_101851-02_TGACCA_L001_R1	25	3229756 bp
COA7AACXX_101851-02_TGACCA_L001_R2	27	3935211 bp
COA7AACXX_101851-02_TGACCA_L002_R1	21	2808305 bp
COA7AACXX_101851-02_TGACCA_L002_R2	21	3191481 bp
D0TRBACXX_101851-02_TGACCA_L004_R1	25	6669424 bp
D0TRBACXX_101851-02_TGACCA_L004_R2	19	5705615 bp
COA7AACXX_101851-03_ACAGTG_L001_R1	21	2526318 bp
COA7AACXX_101851-03_ACAGTG_L001_R2	21	2629027 bp
COA7AACXX_101851-03_ACAGTG_L002_R1	23	2001776 bp
COA7AACXX_101851-03_ACAGTG_L002_R2	19	2197144 bp
D0TRBACXX_101851-03_ACAGTG_L004_R1	27	5339366 bp
D0TRBACXX_101851-03_ACAGTG_L004_R2	21	5650695 bp
COA7AACXX_101851-06_GCCAAT_L001_R1	21	15228726 bp
COA7AACXX_101851-06_GCCAAT_L001_R2	27	15444998 bp
COA7AACXX_101851-06_GCCAAT_L002_R1	21	11819390 bp
COA7AACXX_101851-06_GCCAAT_L002_R2	29	16882964 bp
D0TRBACXX_101851-06_GCCAAT_L004_R1	17	23727503 bp
D0TRBACXX_101851-06_GCCAAT_L004_R2	35	17159698 bp

Table 2: This table shows the predictions from KmerGenie of each file individually.

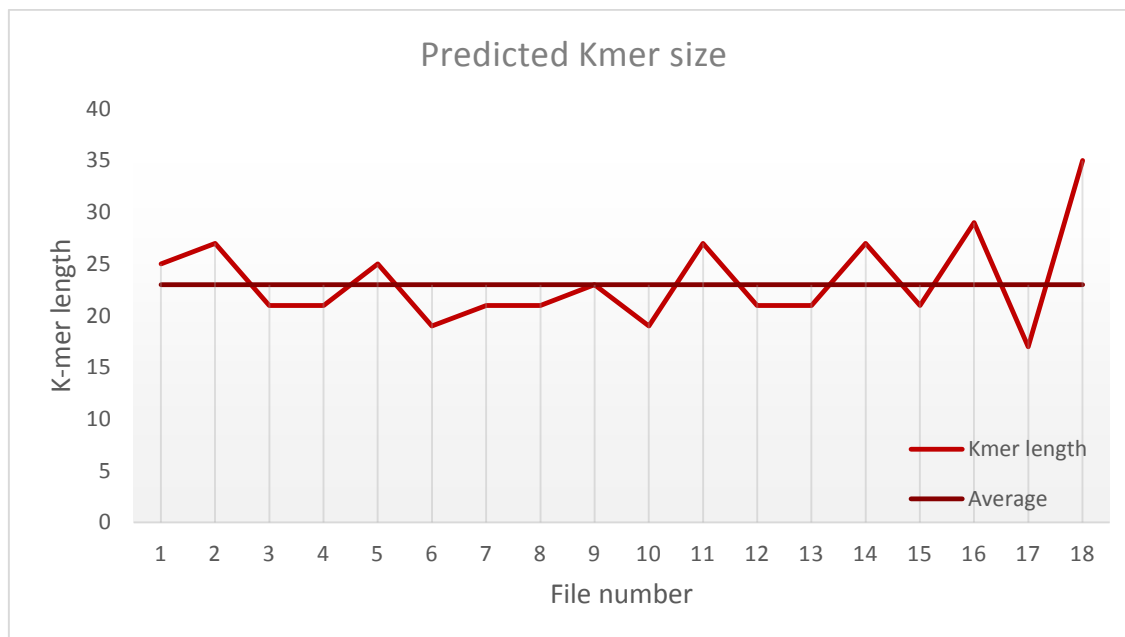


Figure 7: This plot shows the proportions of the differences in the predicted and average k-mer lengths.

On the journey to the optimal reference genome for subsequent SNP calling a few different approaches were used. The first and basic approach was to assemble all the different individuals into one pseudo-reference genome, assembled with a k-mer of 23. This resulted in the following contig and scaffold lengths and further statistics (Figure 8):

Average Contig length: 164.
Average Scaffold length: 262

Contig>100	6099793	98.60%
Contig>500	197151	3.19%
Contig>1K	34488	0.56%
Contig>10K	223	0.00%
Contig>100K	0	0.00%
Contig>1M	0	0.00%
scaffolds>100	4502329	98.70%
scaffolds>500	446170	9.78%
scaffolds>1K	114488	2.51%
scaffolds>10K	615	0.01%
scaffolds>100K	0	0.00%
scaffolds>1M	0	0.00%

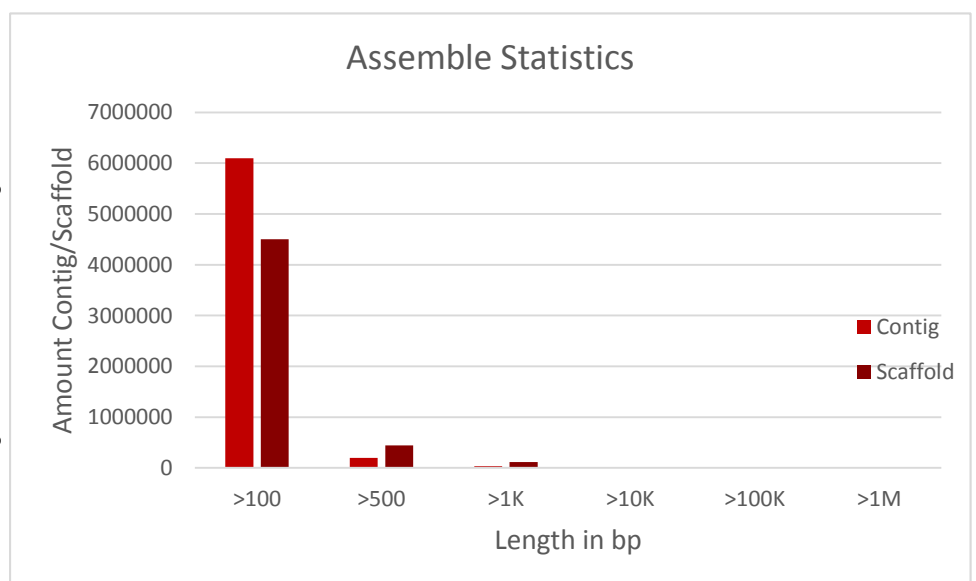


Figure 8: The assemble stats of an assembly with a k-mer length of 23. The x-axis shows the length of the contig or scaffold. The y-axis shows the amount of contigs or scaffolds.

The second approach was more of a k-mer check. To determine if the right k-mer was used on the data set, even though KmerGenie predicted the optimal k-mer lengths. This was done because of some outliers in the k-mer prediction. A second and third assembly was done, one with a k-mer size of 20 and one with 26. The results are showed below.

K-mer size of 20:
Average Contig length: 176
Average Scaffold length: 183

k-mer size of 26
Average Contig length: 167
Average Scaffold length: 176

The third approach was to assemble only one individual at a time. To do this the first individual from the first host plant was chosen. This individual was chosen because of a constant predicted k-mer size with little to no outliers. The average k-mer size of these files was 23 as well. The result was no better than the first assembly.

Alignment

The first alignment with Bowtie2 was a standard alignment with default parameters and all the data files from all the individuals. The alignment rate of this alignment was only 0.42%. So only 0.42% of all the reads were aligned against the reference genome. After this alignment a local alignment, again with Bowtie2, was done. These results were way better. The alignment rate of this alignment was 10.38%. The reference genome of these two alignments came from *Ceratitis capitata*. To get more insight in contamination of the data an alignment against a human genome was performed. This was a local alignment as well, because of the better results from the local alignment in the previous alignment. 5.39% was aligned against a human genome. This is half compared to the alignment against the *Ceratitis capitata* genome.

The second tool that was used for alignment was BWA. Equivalent assemblies were conducted with BWA. The first global assembly had an alignment-rate of 2%. This is, again, a very low alignment rate and it has no use to convert this result into a consensus sequence. To get a higher alignment rate a local alignment was performed using BWA. This alignment had an alignment rate of 18%. This is substantially higher than the executed alignments with Bowtie2. Despite this improvement no high-quality reference genome could be created. This was probably due to the contamination in the data.

To see if this higher alignment rate applies to an alignment with a human genome as well, a local BWA alignment against a human genome was performed. This resulted in an alignment rate of 11%. This is a significantly higher alignment rate as well, but this strongly hints that human DNA is present in the samples.

BLAST

To get more insight into the data and the presence of different organisms the BLAST genome database was used. With this database one is able to find the origin of a sequence and find out if the dataset is contaminated. A sample of “500 base pair” contigs, these contigs originated from the first and best assembly, were BLASTed against the database. The results will only give an idea of the contamination, but this method cannot prove contamination. The sample consisted of 243 contigs, all with a base pair length of 500. The results of this nucleotide blast (blastn) are displayed in a circle diagram in figure 9. The first thing that strikes out is the huge amount of hits correlating with bacterial genomes. More than 50% of the “500 base pair contigs” show strong similarity with bacteria. Three bacteria made up most of those 50%. These three bacteria are: *Klebsiella oxytoca*, *Enterobacter* and *Raoultella ornithinolytica*. The *Wolbachia* is found a fair amount of time as well, but not nearly as much as the other three. 40% of the hits refer to an insect genome. Most of those hits came from the *Ceratitis capitata*, which is also used as the reference genome for Bowtie2 and BWA. 15% of the hits referred to the human genome. All the hits had an e-value under $1e-10$. The hits were validated and checked on reliability. Of the 243 contigs 130 were not found in the BLAST database, so there is no telling where they are from.

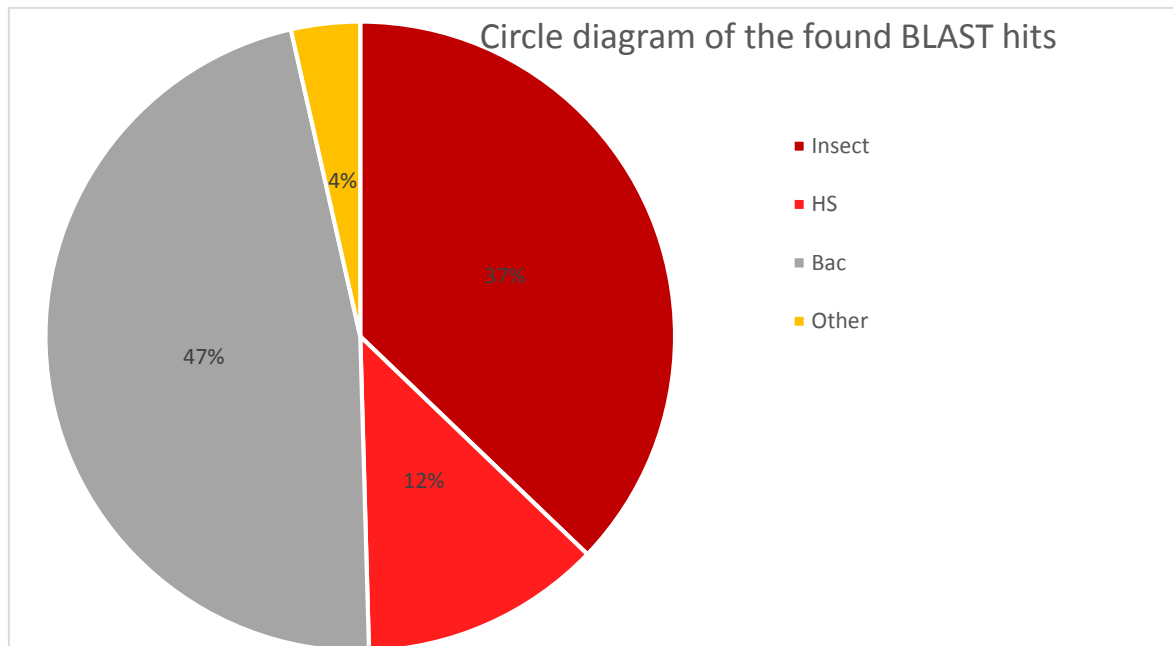


Figure 9: This circle diagram shows the BLAST hits with the “500 base pair contigs”. These contigs were blasted against the genome BLAST database with the blastn program.

After looking at these results, a local BLAST was conducted. This was a local blastn against a database made with a *Ceratitidis capitata* genome. This will reduce the amount of contaminated reads in the dataset. Only contigs larger than 199 base pairs were blasted. These were in total 1,591,921 contigs. This was a data reduction of 95%. After this BLAST the filtered and approved reads were transferred to a different file. The contigs from before and the contigs from after the blastn are compared in figure 10. The number of approved contigs was 259,174. This was a total data reduction of 99.2%. The average length of the contigs was 500 base pairs, this was a big improvement. Another good thing was that the largest contig is almost as big as the one from before the local blastn.

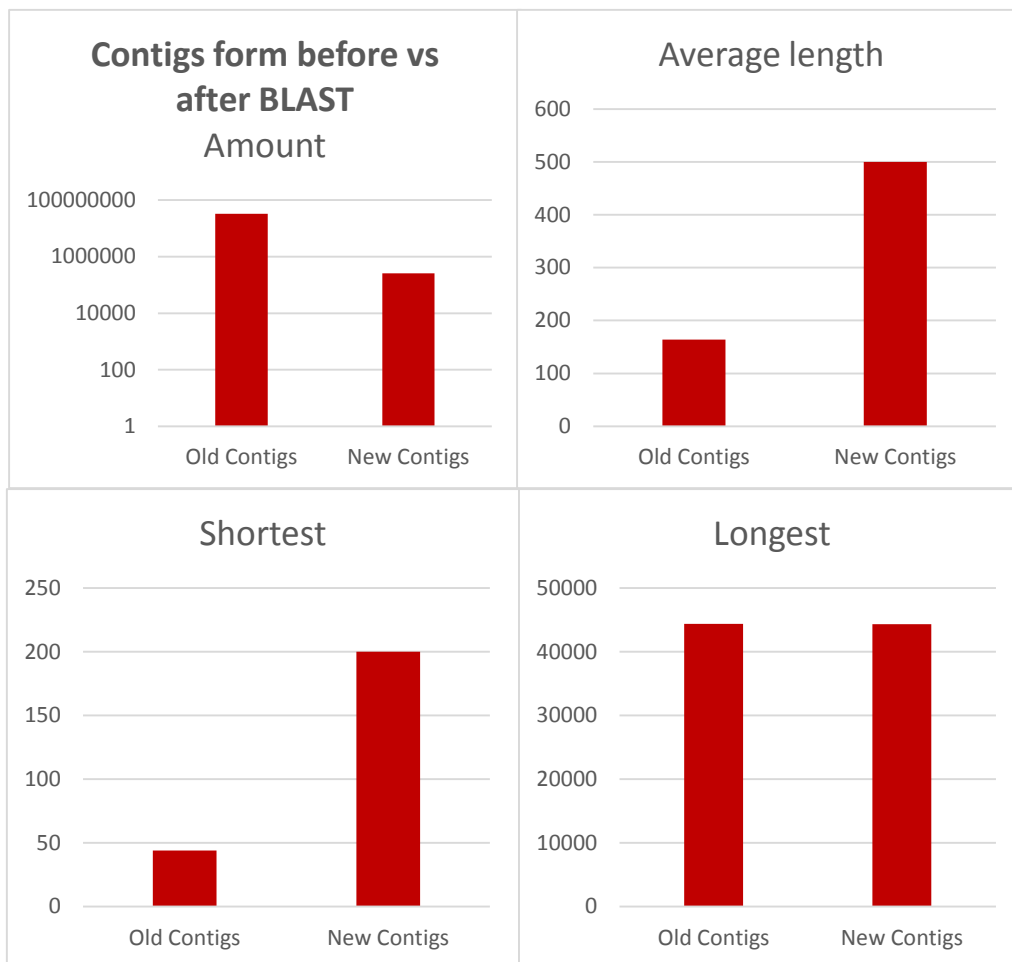


Figure 10: This figure shows the comparison and the differences between the contig dataset from before the local blastn and from after the local blastn. These two are compared in the amount of contigs in the file, the average length of the contigs, the shortest and the longest contig.

To check if there was a reduction in contaminated reads in the data set, contigs consisting of 500 base pairs were transferred to a different file to BLAST against the genome database again. The results are shown in figure 11. The amount of “500 base pair contigs” after the blast was 209. Of these 208, 109 contigs were not found in the BLAST database. The 100 contigs that were found show promising results. 73% of these hits show a strong similarity with insects. And mainly fruit flies to be more specific. This time no human contigs were found in the dataset and only 18 bacteria were found.

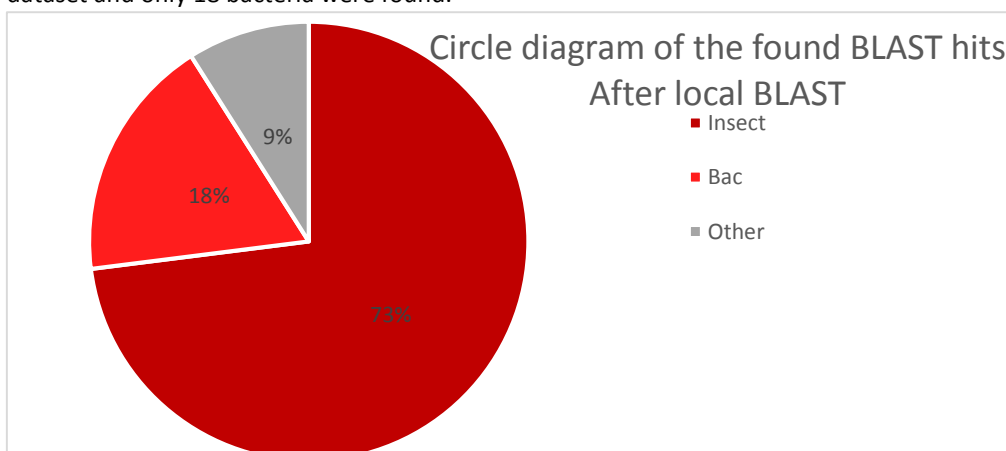


Figure 11: This circle diagram shows the BLAST hits with the “500 base pair contigs” after the local blastn. These contigs were blasted against the genome BLAST database with the blastn program.

Discussion

Data

The quality of obtained data was very high in terms of the phred scores. Because of that the trimming went without any problems and not much data was lost. The trimming did not improve the quality of the data much because the starting quality was already high. That means that the trimming was not really necessary for this data, but it did not do any harm to the next processes. Furthermore, the amount of data was enough to assemble and align with.

A problem might arise because of the lack of different individuals that were sequenced. If more individuals of both host plants were sequenced the amount of data would be more favourable and the certainty and quality of the assembly and alignment would be higher. This would be better for further analysis to find differences in the two populations of this organism for the SNP analysis.

FastQC also showed that the quality of the data was high. There is a minor concern because of the higher duplication levels. This might mean that some reads are sequenced more than others or the amplification went wrong here. It could also mean that there is a lot of repetitive DNA in the genome of this species. Perhaps these duplicated reads come from another organism. However, FastQC shows no warning concerning the overrepresented reads.

The check conducted with BLAST shows a strong possibility that the data were contaminated. It is usual that some bacterial sequences can be found in the data. This is because the way the samples are prepared and then sequenced. The flies were ground and DNA was extracted from them and everything that was in them. However, it is remarkable that 50%, according to the "500 base pair contigs" sample, was made up of bacteria. Another noticeable thing is the 15% human that is found. This could be because something originating from a human came in contact with the sample. This likely happened at the preparation stage.

Assembly

The assembly of the *Rhagoletis cerasi* is pretty ill considering the length of the contigs and scaffolds. This might be because of the contamination in the data. Another possibility is that there are repetitive parts in the genome of the *Rhagoletis cerasi*. SOAPdenovo2 might have a hard time assembling the repetitive parts. The combination of the two is also possible. The bad assembly will make it difficult to find concrete differences between the two populations and to compare their DNA.

BLAST and Filtering

The filtered contigs of 200 base pairs and more might make a new assembly or alignment possible. This assembly and alignment will not suffer from the contamination like the previous ones. By filtering the foreign DNA is removed, so eventually differences between the populations might be found instead of differences between *Rhagoletis Cerasi* and another organism.

The downside of this method was the loss of data. There is a high chance of removing some relevant *Rhagoletis Cerasi* contigs instead of the contigs corresponding to other organisms. This might result in an incomplete representation of the differences between the two populations.

The results show that contamination is greatly reduced by the local BLAST and the filtering afterwards. However, the data reduction is very severe. This total data reduction of 99,2% might be a stumbling block to a further project. The plus side is that the data is more reliable now and results with higher certainty might come out. This is because of the reduction in the contamination.

Because of this contamination the assembly and alignment took longer than first estimated. To filter out the useful data was a good step, but this took some time as well. This caused that no genomic comparison between the two populations could be conducted, so no evidence of speciation-with-gene-flow could be found.

Conclusion

As stated in the discussion, no evidence of speciation-with-gene-flow could be found because of the circumstances. However, a big part of the prework is already done. The data is already trimmed in a good manner. Contigs are made that could suffice as data for an analysis. These contigs are filtered and pre-processed to get the best results as possible.

To conclude the results:

The pre-processing went according to plan. It was not really necessary to process these reads, but it did no harm as well. Only 10% of the data is lost due to trimming and they are lost because they did not fulfil the criteria. The assembly resulted in quite bad results with a lot of very short contigs and scaffolds. However, after the BLAST step the average length of the contigs increased significantly. This might help with further assembly of the genome, or it can be used as data for a SNP analysis. The new dataset is almost fully decontaminated. That means the BLAST was a success and further research with this data is possible.

Further Studies

There are a lot of things that can be done after this project. For starters, this current project can be finished. This can be one in a few different ways.

- 1) The contigs can be used to do a further assembly or a alignment to make a good pseudo-reference genome of the *Rhagoletis cerasi*. When a good reference genome is made the original reads can be mapped against it. This might result in SNPs that thereafter can be used for variation analysis. The outcome might be the evidence of speciation-with-gene-flow that is the core question of this project.
- 2) The contigs can be used as an incomplete reference genome to produce a VCF file with significant differences between the populations. This will answer the core question of this study as well.

Another thing to do is to collect new and more data of these different populations. With better and new data one might produce a good reference genome without any or minimal contamination. When the data is abundant the found differences will be more significant. This reference genome can be verified with a search for core genes / household genes. If these genes are found and are found in a way usual for this kind of species the reference genome will be of a good quality. This is strongly recommended to do, and can be done through the tool Cegma.

References

- [1] **Understanding Evolution**. 2016. University of California Museum of Paleontology. 22 Februari 2016
<<http://evolution.berkeley.edu/>>
- [2] **Darwin, C.** (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London.
- [3] **Jonathan B. Losos, David A. Baum, Douglas J. Futuyma, Hopi E. Hoekstra, Richard E. Lenski, Allen J. Moore, Catherine L. Peichel, Dolph Schluter, Michael C. Whitlock**. 2014. *The Princeton Guide to Evolution*. 2014: 517
- [4] **Dwayne van der Klugt**. 2014. Adaptive Host-Shift of Fruit Flies (*Rhagoletis cerasi* & *R. alternata*, Diptera: Tephritidae) on Exotic Plants
- [5] **John L. Capinera**. 2008. Encyclopedia of Entomology 2nd Edition: 1367
- [6] **Johann Georg Sturm, Jacob Sturm**. 2010. <http://www.wikiwand.com/>
- [7] **Egan SP, Ragland GJ, Assour L, Powell TH, Hood GR, Emrich S, Nosil P, Feder JL**. 2015. Experimental evidence of genome-wide impact of ecological selection during early stages of speciation-with-gene-flow. Ecology Letters. **18**(8): 817-825
- [8] **Istvan Albert**. 2014. A Short Guide to Illumina Sequencing.
<<http://www.personal.psu.edu/iaa1/courses/illumina-sequencing.html> >
- [9] **Marcel Martin**. 2012. Cutadapt Removes Adapter Sequences From High-Throughput Sequencing Reads. EBMnet.journal **17**(1): 10-12
- [10] **Graham Etherington**. 2014. Why you should QC your reads AND your assembly.
<<http://grahametherington.blogspot.nl/2014/09/why-you-should-qc-your-reads-and-your.html>>
- [11] **Simon Andrews**. 2010. FastQC Babraham Bioinformatics.
- [12] **Ruibang Luo, Binghang Liu, Yinlong Xie, Zhenyu Li, Weihua Huang, Jianying Yuan, Guangzhu He, Yanxiang Chen, Qi Pan, Yunjie Liu, Jingbo Tang, Gengxiong Wu, Hao Zhang, Yujian Shi, Yong Liu, Chang Yu, Bo Wang, Yao Lu, Changlei Han, David W Cheung, Siu-Ming Yiu, Shaoliang Peng, Zhu Xiaoqian, Guangming Liu, Xiangke Liao, Yingrui Li, Huanming Yang, Jian Wang, Tak-Wah Lam2 and Jun Wang**. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience **1**:18
- [13] **Rayan Chikhi and Paul Medvedev**. 2013. Informed and automated k-mer size selection for genome assembly.
- [14] **NCBI, National Center for Biotechnology Information, U.S. National Library of Medicine**. BLAST, Basic Local Alignment Search Tool.
<http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHomeNew>
- [15] **Langmead B, Salzberg S**. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods. **9**: 357-359.
- [16] **Li H. and Durbin R**. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics. **25**: 1754-60. [PMID: 19451168]
- [17] **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R**. 2009. 1000 Genome Project Data Processing Subgroup (2009). "The Sequence Alignment/Map format and SAMtools". Bioinformatics **25**(16): 2078–2079.
- [18] **Camacho C1, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health**. 2009. BLAST+: architecture and applications. [PMID: 20003500]