Assembly of a foraminifera (meta)genome

Supervisors: Jeroen Pijpe
Arjen Speksnijder
Rutger Vos

Authors:

Bo Baas	s1107085
Julius van Schaik	s1100879
Jennifer Stolk	s1105938
Robert Zomerdijk	s1105934

Abstract

Due to the importance and complication of the foraminifera and it's symbiotic relations, they make for a very interesting research subject. Until now there hasn't been done much genomic research of the foraminifera.

The objective of this paper is to transform metagenomic data, consisting of *Foraminifera amphisorus* and its symbiotic relations, symbiodinium and bacteria, into a full draft genome of the *F. amphisorus*. This will be done by using a pipeline which transforms, filters and prepares the data for a de novo assembly.
After the de novo assembly the objective is to make a draft genome of the *F. amphisorus.*

If the project fails it could be an indication that the task is currently still too complicated to be solved with bio-informatics. But due to unforeseen complications the project hasn't been wholly completed and reasons for such will be explained.

Table of contents

An introduction

Foraminifera are single celled organisms that belong to the Protista kingdom. They are characterized by an external shell that is called a “test”. These organisms often live in marine environments. The abundance of these organisms shows they are an important part of underwater life and help scientists in their research. Because foraminifera’s show a mostly stable continuous evolutionary development, they are used to predict the age of marine rocks. Different species foraminifera live in different environments, so foraminifera is also used to discover past marine environments. [1]

Foraminifera often live in symbiosis with other organisms such as algae and bacteria. In the symbiotic relationships, the Foram is more often than not the host.[2]The organism living in symbiosis with *Foraminifera amhisorus* is *Symbiodinium microadriaticum. S. microadriaticum* are small photosynthetic dinoflagellates.

For this research we have been given a metagenome containing a mixture of a *F. amphisorus* genome, the genome of its symbiont *S. microadriaticum* and bacterial genomes which all live in symbiosis with the *F. amphisorus*. The samples are extracted from *F. amphisorus* in Jurien Bay, Parker Point and Horrocks all located in west Australia. The objective is to filter out the *S. microadriaticum* and bacterial genome, and then extract the Foram reads from the metagenome. Once the *F. amphisorus* reads have been extracted, a draft genome of the *F. amphisorus* will be created.

Failing to provide a draft genome will prove that it either isn’t possible or at least very challenging to separate the different genomes using bio-informatics and it should be separated by laboratory researchers.

Methods

Workflow

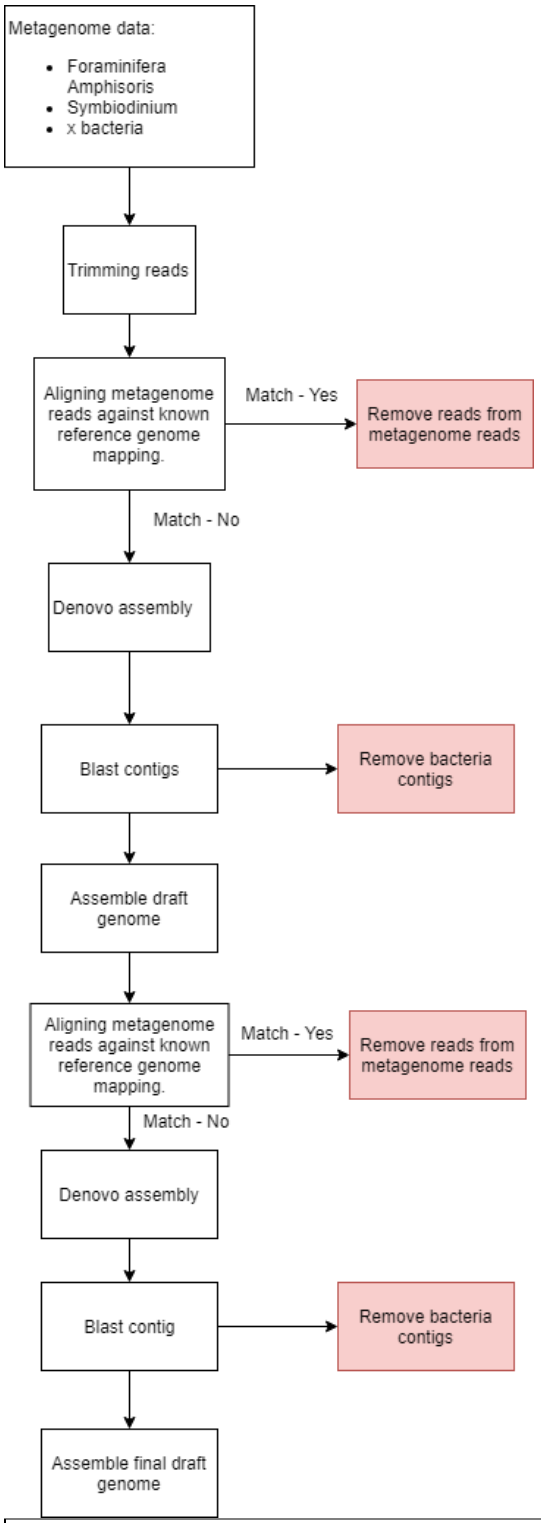

Figure 1: Flowchart of the steps in this project. 7-10 are repeat steps to ensure all non-foraminifera data is removed.

In order to make the project go as smoothly as possible we've created a workflow(figure 1). Each step of the workflow is described below.

Github

A github page has been created to ensure that all scripts and documents were available to all group members. All scripts can be found in the scripts folder. The link to the github page is: https://github.com/naturalis/foram-assembly

Data

The data used for this project was provided by Naturalis. The data consisted of short novaseq paired end reads in a fastq file. While sequencing researchers found a main difference between the reads found at different sample locations in Australia. They found that some samples had an 1 base pair difference in comparison with the others. This could mean that they are different species. With that in mind only reads from 1 type of sample were taken for this project. The reads that were used came from 3 different locations, in the table below is an overview of the reads.

File name	Explanation
H[1_1, 1_2, 2_1, 2_2, 3_1, 3_2, 4_1, 4_2, 5_1, 5_2, 6_1, 6_2, 7_1, 7_2, 8_1, 8_2, 9_1, 9_2, 10_1, 10_2].fastq	Fastq files of samples from Horrocks, Australia.

JB[1_1, 1_2 , 3_1, 3_2, 4_1, 4_2, 5_1, 5_2, 6_1, 6_2, 9_1, 9_2, 10_1, 10_2].fastq	Fastq files of samples from Jurien Bay, Australia.
PP[1_1, 1_2, 2_1, 2_2, 3_1, 3_2, 4_1, 4_2, 5_1, 5_2, 6_1, 6_2, 7_1, 7_2, 9_1, 9_2, 10_1, 10_2].fastq	Fastq files of samples from Parker Point, Australia.

In order to get the best results the reads had to be trimmed and filtered first.

Trimming reads

The reads were trimmed using a tool called fastP.[4] To prevent huge loss of data but still filter the reads the default settings were used. After filtering, the reads with a low quality score were removed from the dataset. The script that was used to perform this step is on the github page in the script folder. the script is called SnakeMake_foram. The step is performed by the lines of code 39 to 50.

Mapping

The next step to filter the reads even further was to map the reads to the reference genome of symbiodinium (NCBI ID: 13759)[5] in order to remove the symbiodinium reads from the metagenome pool.
For this step minimap2[6] was used with the default settings. After the mapping step symbiodinium was still in the dataset, but marked as mapped. In order to remove the symbodinium reads a tool called samtools was used.
The script that was used to perform this step is on the github page in the script folder. The script is called SnakeMake_foram. The step is performed by the lines of code 53 to 139.

Denovo assembly

For the denovo assembly the tool SOAPdenovo 2[7] was used. As input, soapdenovo needs a config file. This contains, among other things, the maximum reads length, the kmer and the paths to the file with the reversed reads and the file with the forward reads.

To ensure that SOAPdenovo runs smoothly, the forward and reversed file must be the same length. The files must also be sorted, the read pairs must be in the same place in both files. Fastq sort was used to sort the files.
After mapping, the files no longer had the same length. to solve this, a script has been written to remove the orphan reads (reads without a forward or reversed partner). The script that was used to perform this step is on the github page in the script folder. The script is called SnakeMake_matching. This step is performed by the lines of code 75 to 122.

The reads were then used as input for Kmergenie[8]. Kmergenie calculates the optimal kmer needed for the config file. SOAPdenovo needs an odd kmer number.

A small data set was used to ensure that SOAPdenovo works. This dataset is made using fastq sample with a sample percentage of 30%. A new config file has been created for the sample set.

The used config file is on the github page in the conf folder. The config file that is used is called config_file_sample.

Results

The three different kind of reads, Horrocks (H), Jurien Bay (JB), and Parker Point (PP) have been processed to create contigs to assemble a draft genome of the *F. amphisorus.*

Quality trimming

The first thing to do was to remove all low quality reads. The read was considered of low quality if the read quality score was lower than 20.

	# Horrocks (H) reads	**# Jurien Bay (JB) reads**	**# Parker Point (PP) reads**	**# Total reads**
Before trimming	246.097.014	138.160.606	260.837.464	645.095.084
After trimming	243.740.898	136.929.584	258.735.104	639.405.586
# deleted reads	2.356.116	1.231.022	2.102.360	5.689.498
% deleted reads	0.957 %	0.891 %	0.806 %	0.882 %

Table 1: Fastp quality trimming results, which shows the amount of reads before and after trimming and the amount of reads that were removed by Fastp.

As shown in table 1, just under one percent of the total reads have been removed due to low quality. Out of all reads, Horrocks has had the most amount of reads removed and Jurien Bay the least.

Reference genome mapping

Figure 1 shows the amount of reads filtered from the dataset. The orange part of the pie charts are the filtered symbiodinium reads, the reference genome. The green part of the pie charts are of the *Foraminifera* genome.

The graphs shows very similar results between the different kind of reads. Each of them had around 5% of its total reads filtered, which were the symbiodinium *reads.*

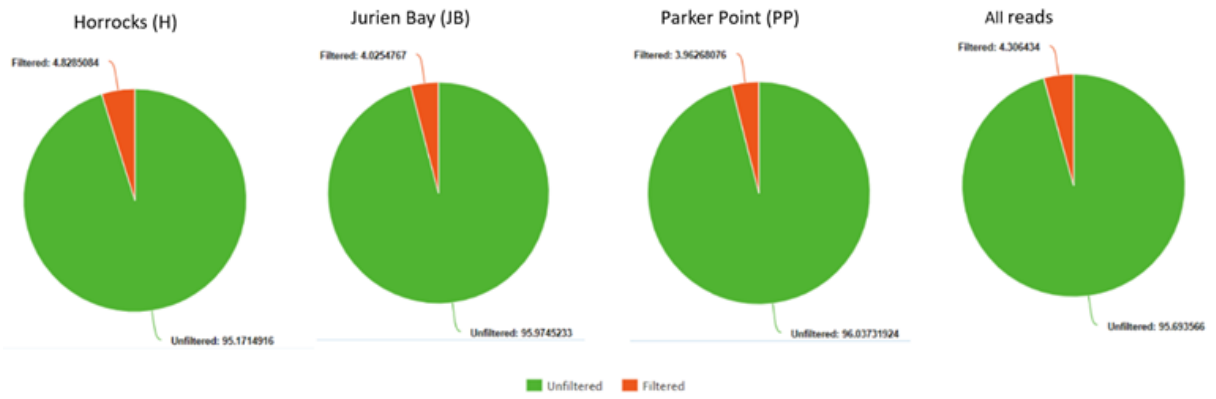

Figure 1: A display of the amount of mapped reads against the reference genome in percentages.

According to figure 1 it looks like only 5% of the metagenome is derived from the *S. microadriaticum*. If we take a closer look at the files created by the reference genome mapping. It showed the files were corrupted and had to be cleaned. The result of which is shown in figure 2.

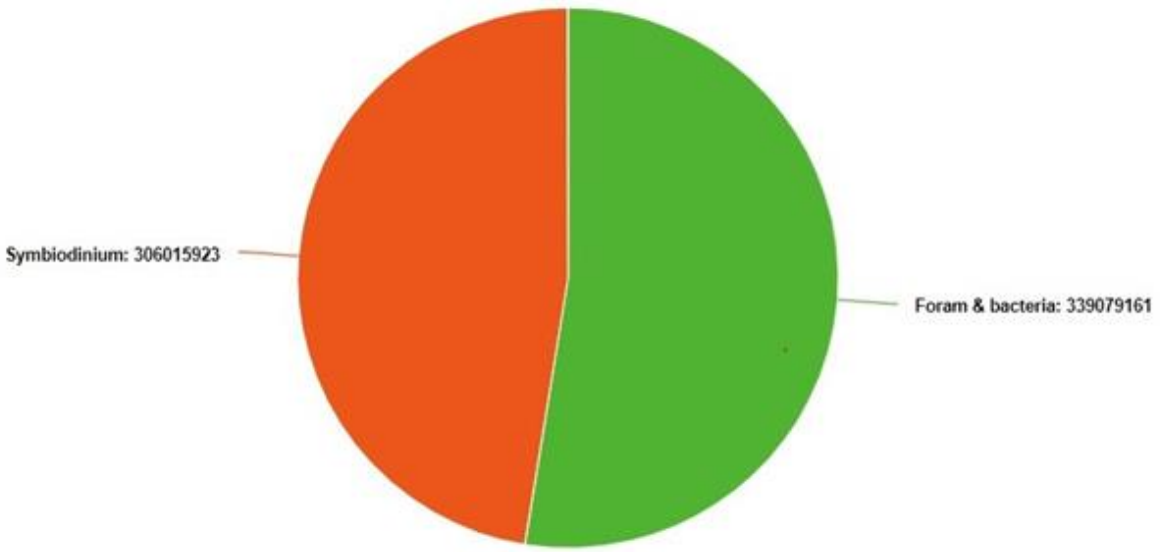

Figure 2: A display of the amount of mapped reads against the reference genome after they have been cleaned.

Figure 2 shows a much more significant change in the amount of filtered reads. Before only 5% of the total reads was filtered, now it is around 40%. The remaining reads contains the *F. amphisorus* and bacterial genomes, which are 339.079.161 reads.

Soapdenovo2 results

Table 2 shows the result of the soapdenovo2 run of 30% of the left-over dataset. The left-over dataset consists of the *F. amphisorus* genome and the bacterial genomes.

Soapdenovo2 statistics	
# Contigs	5.587.803
Mean size	379
Longest seq	31.528 nt
Shortest seq	100 nt
% Nucleotide A	27.47 %
% Nucleotide C	23.24 %
% Nucleotide G	22.67 %
% Nucleotide T	26.63 %
GC Content	45.90 %
N50	558

There are 5.587.803 contigs of which the longest is 31.528 nucleotides and the shortest is 100 nucleotides. The N50 is 558, which means 558 is the length of the shortest contig at 50% of the total genome length.

Discussion

There were several problems regarding the research. This includes three server breakdowns, which halted the process up to a week each time. This time was needed to reboot the server.

The mapping of the reads against the symbiodinium also took longer than expected, because minimap wasn't able to map the reads against the symbiodinium genome. We expected the problem to somewhere within the minimap settings and spent a lot of time changing the settings. In reality, the wrong reference reads where downloaded which made minimap unable to map the reads.

The symbiodinium reads that were filtered out was around 5% of the total reads. The expectation was to filter out around 5% to 10% with the mapping based on the reference genome.

The server instance made for the project had space issues. The total read input for SOAPdenovo before filtering out the orphans was around 200 GB. This caused a memory error when trying to run SOAPdenovo. The problem was quickly solved by allowing access to a bigger high memory server. However, the merged forward and reversed files where so big that they couldn't be sorted according to their IDs. The unmerged files where already deleted due to memory space issues. Thus all previous steps were repeated to sort the individual sample files before merging and to remove the orphans. After this was done the files where again merged.

When inputting this data in SOAPdenovo, it gave a format error back. After looking into the forward and reverse file, it was discovered that the splitting of the forward and reverse reads done with egrep caused this problem. The code resulted in unwanted -- in between lines. The extra lines disturbed the fastq format and made it unrecognizable for SOAPdenovo. A code was made to remove

these unwanted lines and after that de SOAPdenovo was sampled and ran without any further problems. Unfortunately only the 30% sample could be run in SOAPdenovo due to time limitations.

The contigs made both contain *F. amphisorus* and bacteria contigs. Since the mean GC content lies around the 45% it is assumed that the biggest part of the contigs are from the *F. amphisorus.* If the contigs were mostly bacterial contigs the GC content would be around 20-30% which is the usual GC content for bacterial genomes. If the longest contigs (around 30.000) are indeed mostly from *F. amphisorus*, then it is possible that the contigs resulting from the complete fastq files could be used to make a draft genome. Taking into account that contigs made with the whole dataset will probably be longer.

Because of the problems within the project, especially around SOAPdenovo, a lot of steps needed to be corrected. Our code deletes all previous steps due to the memory capacity. After discovering a new problem the whole pipeline needed to repeated. This caused a lot of time loss. Therefore, the part after the SOAPdenovo in the workflow could not be completed.

The last steps of the workflow will be reassigned to Richard. This includes the removing of the bacteria contigs and the assembly of the draft genome. Richard will also get extra longreads that will be used to make the draft genome.

Conclusion

Because of some unforeseen problems, we were unable to complete a draft genome. Judging for the contig lengths extracted from the 30% sample SOAPdenovo run it should be theoretical still be possible to make a draft genome. This is assuming that the contig lengths will be longer when using the complete pre-processed fastq data file and that the long contigs assembled are indeed from *F. amphisorus*. This project will be reassigned and the goal to assemble a draft genome of *F. amphisorus* could still be reached.

References

1. **FORAM FACTS. *Berkeley.* [Online] [Citaat van: 11 09 2019.] https://ucmp.berkeley.edu/fosrec/Wetmore.html .**

2. *Symbiosis and the Evolution of Larger Foraminifera.* J. J. Lee, M. E. McEnery, E. G. Kahn and F. L. Schuster. Vol. 25, No. 2, sl : Micropaleontology.

3. Shifu Chen, Yanqing Zhou, Yaru Chen, Jia Gu. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics.* [Online] https://doi.org/10.1093/bioinformatics/bty560.

4. Github opengene/fastp. 12/10/2019. https://github.com/OpenGene/fastp.

5. Symbiodinium sp. clade A Y106. 12/12/2019. https://www.ncbi.nlm.nih.gov/genome/13759?genome_assembly_id=388217.

6. Github minimap2. 12/10/2019. https://github.com/lh3/minimap2.

7. Aquaskyline/SOAPdenovo2. 12/10/2019. https://github.com/aquaskyline/SOAPdenovo2.

8. KmerGenie. 12/10/2019. http://kmergenie.bx.psu.edu/.