Utrecht University
Department of Biology
Department of Science

Graduate School of Life Sciences
Molecular and Cellular Lifesciences
Major Research Project

Analysis of Metagenomic Chromosome Conformation Capture methods and MetaTOR pipeline

Author:
Jolien Rietkerk
6485677

Daily Supervisor:
Dr. I. Stringlis
Examiners:
Dr. R. de Jonge
Prof. Dr. G. van den Ackervecken

January 9, 2020

Laymans Summary

In recent years, researchers have found bacteria that have a beneficial effect on the immune system of plants. By interacting with the plants at the roots, these bacteria are able to improve the immune response of plants, even before any attack on the plants health has been made.

Research has been focussed on identification of such beneficial bacteria and the level to which they are present in the soil. This, in combination with uncovering the way that these bacteria interact with the plant could result in promising tools to improve crop yields, by using the pant - bacteria interactions to improve plant health and protection.

The identification of these bacteria can be done with several methods, which range from laborious lab work to increasingly complicated data analysis. In 2014, researchers in France [] described a method called Metagenomic Chromosome Conformation Capture (Meta3C), which shows promise to improve the detection and decrease the time that laboratory methods take. In short, this Meta3C method uses a chemical that fixates the original three-dimensional state of all the DNA (including that of organisms like bacteria) present in the soil. Next, the DNA is identified and the three – dimensional information is represented by a data set of DNA – DNA interactions. This data can then be analysed and assessed using the computational analysis tool called MetaTOR, which was developed by the same authors as the Meta3C method, in 2019.

In this report, we were interested in applying this new method to detect bacteria in our soil samples, directly. We show that the results described by the French researchers can be reproduced to some extent, on pure bacterial samples. However, optimization on soil samples was complicated and we suggest different sample types for future optimization and research of the method. Additionally, we conclude that the MetaTOR pipeline can be used to analyse the data made by the Meta3C method, but is not as straightforward as described by the authors, since some errors made the analysis very time consuming and results less reliable. Additionally, out of the different sources available for the MetaTOR data analysis tool, none were fully complete in their description of MetaTOR, complicating data analysis and data interpretation even further. Nevertheless, the success of the laboratory methods in our hands confirms the promise of this method and we believe that further research and optimization of the laboratory method and computational method will help identification of bacteria in soil and shed light on the way that bacteria interact with the plant at the root.

Abstract

The plant microbiome is a complex community of microbes. It is able to help the plant by retrieving nutrients and some bacteria close to the roots (in the rhizosphere) are able to induce resistance in the plant. To identify these key players and the mechanism that lies behind their interaction with the plant, research has been investigating which microbes makeup the microbiome. There are several approaches to this research, one of them being a genetic approach, called metagenomics. Within metagenomics, amplicon sequencing and whole metagenomic shotgun sequencing are used to create a representation of microbes present in a sample using the genomic information available. However, these methods come with a plethora of downsides including sequencing bias and database bias created during computational analysis. To circumvent these downsides and biases Metagenomic Chromosome Conformation Capture (Meta3C) and the accompanying computational analysis pipeline called MetaTOR were developed by Marbouty et al.[1] and Baudry et al. respectively [2]. The Meta3C method captures 3D conformation, by means of proximity ligation, resulting in interaction frequencies of these interactions. Addition of this data to a metagenomic assembly of contigs creates a network. In the MetaTOR pipeline, this network is clustered with the Louvain algorithm [3], creating bins that represent all available genomic information in the sample.

In the current work we aimed to replicate and optimize the Meta3C protocol in the lab, as well as inspect the MetaTOR pipeline for computational analysis of the interaction data created by the protocol. Second, we compare three resources for execution of the MetaTOR pipeline and analysis its output.

Contents

Chapter 1

Introduction

The plant root microbiome

The **rhizosphere** is a thin layer of soil lining the outside of the plant root. Here, the plant exudes a large part of its produced carbon. This makes the rhizosphere a nutrient rich environment of which various microorganisms take advantage. As a result, the rhizosphere microbial community is very densely populated, compared to bulk soil. This difference in population density is described as the **rhizosphere effect**. In contrast to the increased population density, the population diversity of the community is decreased in the rhizosphere. This suggests that the **plant root microbiome** consists of selected members with specific functions in the community. [4] [5][6]

This suggestion was confirmed when interactions of the plant root microbiome with the plant were found. For example, the plant can attract microbes from the the microbiome via production of root-exuded molecules [7] [8]. Interestingly, the plant root microbiome has shown to include beneficial microbes that promote plant growth and induce plant immunity in a mechanism called induced systemic resistance, or **ISR**. These beneficial microbes are called plant growth promoting rhizobacteria (**PGPR**) and fungi (**PGPF**). By local interactions at the plant root PGPR and PGPF are able to induce ISR. [6] [5] [9] For example, the bacterium *Pseudomonas simiae* WCS417r has been shown to enhance resistance against pathogen *Fusarium oxysporum* through its interaction at the plant root. [10] [11]

Because of these discoveries the plant root microbiome holds promise for biotechnological applications and agricultural endeavours to increase crop yields.[4] [5] Thus, current research is aimed at the discovery of more beneficial microbes and the identification of the mechanisms that govern these plant-microbe interaction.[12] [13] [14]

Practical application of microbiome research

Originally, this research was conducted by isolation and cultivation of microbes from the rhizosphere. Here, a representation of the microbiome is created by growing the samples in different media and isolating colonies based on physiological differences. Unfortunately, rhizosphere and soil samples are complex and have a highly diverse compared to non-environmental samples, which can not be estimated previous to isolation. This makes these studies hard to replicate. The high complexity of these samples would also require many different plate types and therefore much sample volume in order to obtain a maximum amount of microbes from the sample. Additionally, these techniques are biased towards microbes that are easily kept in culture conditions, and to the media available for selection. The combination of these complications makes this type of research cumbersome and hard to execute.[4] [15]

Metagenomic studies

With the development of next generation sequencing techniques (**NGS**) an improvement on the cultivation techniques has been found in **metagenomics**. Metagenomic studies aim to recover all genomic information in a complex sample by using NGS. In microbial studies this translates to recovering a true representation of the microbial community present in a complex sample. [4] [13].

Amplicon Sequencing

Metagenomic studies can be conducted in several ways. In plant research, **amplicon sequencing** is most applied. Amplicon sequencing uses a polymerase chain reaction (**PCR**) on a DNA sequence that is shared between as many organisms in the sample as possible. The sequence that is most often used in identification of the bacteria in the microbiome is the gene encoding the 16S ribosomal part, because it is shared between most prokaryotes and eukaryotes. The resulting DNA sequences are then determined with NGS and mapped to a database. This mapping assigns an operational taxonomic unit, or **OTU's**, to the sequences. These OTU's are often combined with the relative abundances **RA** of sequenced reads to represent the amount in which the OTU (representing a microbial community) is present in a sample. This creates an indication of the microbiome composition. This is why amplicon sequencing generates a better and more complete view of the microbiome than techniques based on cultivation. It is also cost effective as large amount of samples can be analysed in short time and with relatively little resources. Unfortunately, this technique also has several down sides. [8] [13]

For example, complications are created in amplification of sequences because the genomic region targeted by primers can differ in length, even when the sequences are highly similar, like the 16S region. The effect of these longer sequences can be seen in the RA levels of reads, which complicates determination of the microbe abundances present in a sample. Additionally, the taxonomic levels to which the composition of the microbiome can be determined is limited, because primers are designed for highly comparative sequences.

In amplicon sequencing the taxonomic levels reached in analysis are often phyla, and less often genera. The level of species cannot be reached using amplication sequencing, since genomic differences that separate species from each other are not detected in the amplified shared sequences. Lastly, read analysis is complicated because assignment of OTUs is dependent on how the reads map to the database. This makes it more likely that organisms that have not been found before get assigned a false identifier. This results in an untrue representation of the levels in which organisms are represented leaving them undetected. Additionally, this affects RAs of other communities in the final microbiome representation as well [16][15].

Whole Metagenome Shotgun Sequencing

Next to Amplicon Sequencing, Whole Metagenome Shotgun sequencing (**WMS**) has been increasingly used in metagenomic studies, because it has important improvements over amplicon sequencing. In WMS, lysis of the sample cells and shearing of the DNA into equal sized pieces creates a mixture of all genomic information available in the sample. The sheared DNA pieces are all targeted for amplification, whereas in Amplicon Sequencing the amplification was limited to specific target regions. This is the main advantage of WMS over amplicon sequencing.

The increased genomic information available in WMS allows for a better detection of microbes in the sample. In WMS, taxonomic levels of phyla and even species have been detected. This also leads to easier discovery and identification of previously unidentified microbes. However, the complexity of the data set of small reads complicates the identification of individual communities. To understand why, we look into the steps taken in data analysis of WMS data sets. [16][13]

The first step in analysis of the short reads of a WMS data set is the assembly of the short reads into contigs using sequence overlap.[17] In metagenomic samples this is complex, because sequences that are closely related might cause the assembler to halt prematurely, which leaves the assembly highly fragmented. In contrast, if the parameters that are used are too lenient, the amount of reads that are assembled together increases and contigs containing DNA that originates from various species or phyla are created. This way, the diversity of the sample can be decreased.[16]

A way to create assemblies that are more complete (or less fragmented assemblies with longer contigs) is the use of long read methods such as the minION[18]. These have been effectively used in metagenomics studies, but it remains costly and the single nucleotide error rates are higher than those obtained by short read sequencing methods. A combination of long reads with shotgun sequencing might be possible, but this would increase the costs of the experiments and the sample volume required for the experiment would increase. This is why it is recommended to use the highly fragmented assemblies of short reads and **bin** the reads to decrease the assembly fragmentation.[19] [13]

In binning, contigs are grouped together based on various characteristics, including comparison on taxonomic level and sequence overlap. The former is used in WMS, after binning is complete, final bins (called

Core Communities, or **CC's**) would, theoretically, represent all genomic information of a single organism in the sample. In practice this is often not the case as binning sequences in a complex and large sample is difficult for various reasons. For example, it is hard to determine which sequence of two highly similar sequences belongs to which CC.

To help binning, some sequence characteristics can be used. For example, RA (relative abundance resulting from NGS techniques) of sequences can be used in WMS analysis as a way to identify individual species in a sample, because sequences from the same species are likely to have a similar level of abundance in the sample. Unfortunately, these similarities are not always unique, as some species could be equally abundant in the sample.

Another property used to help binning sequences is the GC content of DNA. Naturally, GC contents are highly similar in related organisms and are therefor limited in their help for binning.[13] [16] Binning techniques such as CONCOCT [20] and MetaBAT[21] both use these qualities to bin sequences.

When assembly and binning are completed to a level that is sufficient according to the researcher, the CC's are annotated.By means of gene prediction and read-database mapping, annotation algorithm and pipelines assign a taxonomic identity and gene content to a CC. This gene prediction creates an opportunity to do a metagenomic function prediction to possibly identify enriched pathways in the sample. Because identification is done based on bins of reads in WMS, contrary to a single amplicon, the false positive rates associated with read identification are lower in WMS than in amplicon sequencing. Additionally, previous unidentified sequences are more likely to be labelled as unidentified when identification of reads is based on more sequence information. [13] [16]

Besides the use of GC content and relative abundances after creating the assembly, binning can also be improved prior to assembly, by so called taxonomy-first methods(like applied in the Kaiju application[22] used by Stringlis et al.[8]). By taking individual reads and assigning them a taxonomic identity a group of reads with the same identity should represent the whole genome of that organism. However, the data base bias in this technique is, because underrepresented organisms in the database will be identified in a lesser amount. This favours the amount of organisms in the sample to those most represented in the database. Additionally, due to the likelihood is low that a sequence is mapped to a closely related organism, rather than labeled as absent in the database. [13] [23]

In summary, the assemblies of WMS are an improvement over amplicon sequencing as it gives a more complete representation of the microbes present in a sample. However, the resulting assemblies are often still highly fragmented and binning of the assembled contigs into communities is complicated. Naturally, improvements are still sought in these techniques. [13]

Chromosome Conformation Capture

A suitable, cost effective solution has been presented by Marbouty et al. [1] [24] [2]. They describe the technique called Metagenomic Chromosome Conformation Capture (**meta3C**) (figure 1.2), based on the original Chromosome Conformation Capture method (**3C**) by Dekker et al. [25].

The 3C and Meta3C methods capture DNA interactions by fixating those interactions which are in close proximity of each other. These interactions can be used to improve binning of contigs, as well as improve the assembly of individual genomes. The 3C method uses the chemical fixative Formaldehyde to keep the DNA fixed in its conformation. Subsequently, the DNA is digested with one or more restriction enzymes. This way small sequences that were in close proximity to each other are kept together while others disperse in the sample away from this interaction. Next, the sample is diluted to increase the dispersion of other sequences and the 'close proximity DNA sequences' are re-ligated. This **proximity-ligation** creates a sequence that is representative of the genomic interactions these sequences had prior to fixation. These interactions can then be used to create a network and the network is clustered, based on the amount of interactions that occur between parts of the DNA. The short range interactions are indications that sequences belong to the same genome or even same chromosome, whereas long range interactions are indications of separate genomes interacting. Naturally, the short range interactions occur more often than long range interactions. This translates to their **interaction frequencies** which are the output of Meta3C experiments. The short range interaction frequencies are high, whereas long range interaction frequencies are lower. This creates a characteristic on which we can bin sequences in the created network of contigs and interaction frequencies associated with these contigs. As shown in figure 1.1 MetaTOR uses an algorithm that does this. High interacting sequences are clustered together and there are as little as possible interactions between clusters.

As in 3C methods, the short range interaction data can also be applied to improve assemblies of individual genomes as well as scaffolding chromosomes and find their 3D conformation. [25] [26] [2]

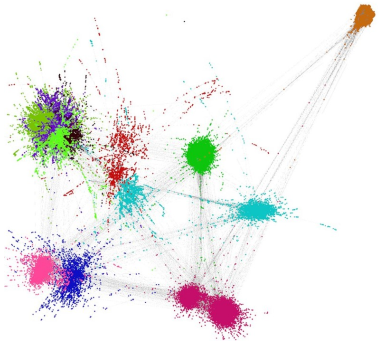


Figure 1.1: Adapted from Marbouty et al. 2014 [1]. This is a visual representation of the network created by the Meta3C analysis pipeline tool called MetaTOR. The nodes in this network represent contigs, whereas the edges are weighted by interaction frequencies. The long range interactions captured in Meta3C overall have a lower frequency than the interactions that are short range. This can be seen in this figure by the low amount of lines connecting nodes between clustered areas in the network, compared to the areas that are highly closely clustered. Naturally, genomic content of a cell is in contact with itself more than it is in contact with genomic content of another cell (or other species in this case). The individual colours indicate the different taxonomic identities associated with the reads. One cluster is multi colour, indicating that clustering of the network was not fully efficient. In MetaTOR this is solved by re-iterating over this sub-network.

Metagenomic Chromosome Conformation Capture

The first steps of Meta3C (step 1 through 6 in figure 1.2) are the same as in other 3C methods, except for the type of sample used, as Meta3C was created on a mixture of bacterial cultures and 3C methods are applied on single bacterial cultures. For this reason the subsequent analysis steps (step 7 through 9 in figure 1.2) of Meta3C are different. In analysis, the reads are assembled *de novo* based on sequence overlap. Next, the interaction data is applied to create a network of contigs. To obtain individual CCs from this mixture, the Louvain algorithm [3] is iterated over the network of connected contigs to create a minimal amount of interactions between CC's , but a maximal amount of interactions within CC's. This is shown in in figure 1.1 where a network after clustering is shown. The CC's are then annotated to show the final representation of the complex sample (see figure 1.2 and 2.14). [1][26][2]

The development of this method was done by Marbouty et al. [1] with an initial experiment using three bacterial strains. Here, the authors showed that these interactions decrease the fragmentation of a metagenomic assembly. In their first experiment the reference genomes and interaction data were combined to identify three core communities and construct a three dimensional (**3D**) representation of these communities. The promise of the interaction data to help in binning the communities lead to another application of the analysis method where the reads resulting from the Meta3C NGS were assembled *de novo* and the interaction data was used on this assembly to decrease the level of fragmented contigs in the assembly. [1]

In their complementary experiments the authors used eleven yeast species and an environmental soil sample to show the promise of this *de novo* assembly method. Using the same analytical approach *de novo* assemblies of the genomes were made. However, their ability to reconstruct the genomes from these binned communities was harder due to a lower read depth in this experiment. Additionally, the soil sample was enriched in Luria Broth (**LB**) medium prior to Meta3C processing which induced a bias alike the one described above for cultivation based techniques. [1]

Follow-up papers by the same authors emphasize the promise of this new metagenomic technique even further. By use of a mouse gut microbiome sample, that genomic interactions between larger sequences are detectable, as they show by detecting phage-host interactions in this complex sample.[26]

Finally, their publication describing the computational pipeline ([2]) used for the full analysis of metagenomic Meta3C projects called MetaTOR (see figure 2.14) shows how the use of interaction data in the as-

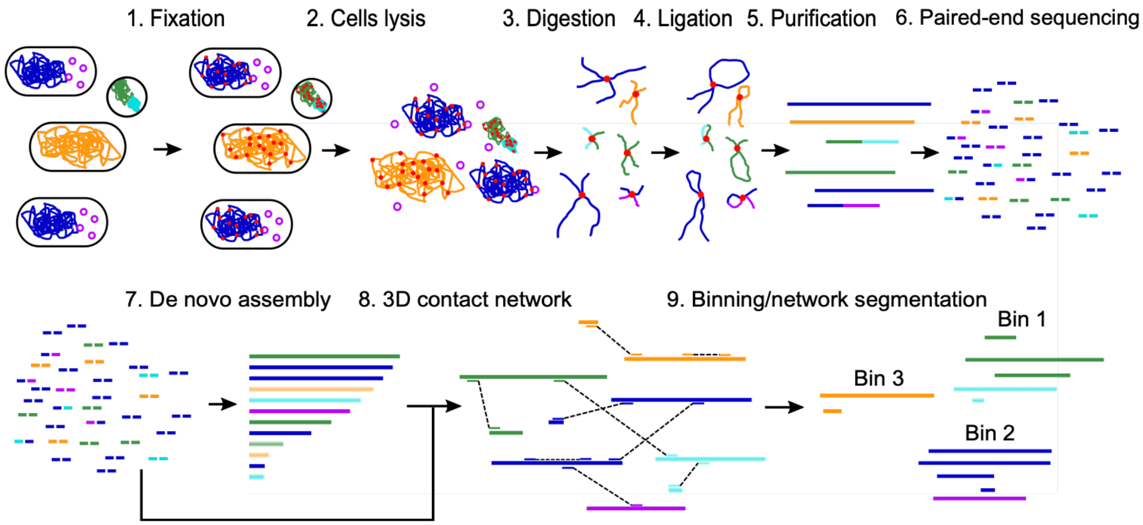


Figure 1.2: Schematic step-wise description of the Metagenomic Chromosome Conformation Capture method. Adapted from figure 1 by Foutel-Rodier et al. 2018[27]. **1.Fixation** Formaldehyde fixation makes covalent bonds between DNA pieces and DNA-protein complexes, fixating the genomic 3D information **2.Cell lysis** Mechanical (bead beating) and chemical (SDS) lysis are combined to have as little bias as possible for lysis of microbes in the sample. **3.Digestion** Using HpaII, a 4 cutting digestion protein, all DNA in the sample is digested into small pieces. At this point formaldehyde fixation keeps the DNA pieces that are close in 3D bonded. **4.Ligation** By diluting the sample highly, only those DNA pieces still bonded by the formaldehyde fixation can ligate together, creating a 2D representation of the genomic interactions. **5.Purification** Reversing the formaldehyde fixation and extracting the ligated DNA interactions from the sample are necessary to prepare the sample for step **6.Paired-end sequencing** where we use Illumina Next Generation Sequencing to get the DNA sequences of genomic interactions. Next during computational analysis step **7.*De novo* assembly** where the reads are used in non-paired end mode, will create a range of contigs representative of the genomic information in the samples. The reads can then be mapped against this assembly (indicated by the undergoing arrow) which is using the interaction data to create step **8. 3D contact network**. Here all the different contigs are linked together using the aforementioned interaction data making the last step a lot more efficient then when interaction data is not present. **9. Binning/network segmentation** in the MetaTOR pipeline an algorithm called Louvain algorithm [3] is used in 100 iterations to search the most optimal binning of the different interaction network clusters. These clusters are nearly all representative of a single genome, coming from a single microbe colony in the sample. The difference between a single genome or the mix of two (for example bin2 has 2 genomes) can be seen by comparing relative abundances, where the clusters containing more than one genome have an exceptionally high abundance in the sample. An iteration over this bin would probably help separate these two.[27]

sembly of meta genomes improves metagenomic assemblies, compared to MetaBAT [21] and CONCOCT[20] pipelines described above.

Overall the Meta3C method shows to be very promising for metagenomic studies. For this reason we aimed to use this method in our research to identify key players in the plant root microbiome.

The plant-microbe interaction research that was a starting point for this project, which requires optimization of the protocol for environmental/soil samples as the original protocol was created on mixed cultured bacterial communities. Marbouty et al, [1] have applied Meta3C to an environmental soil sample before, however they induced a cultivation bias by enriching the sample in LB medium overnight, prior to WMS. We aim to prevent such a bias by applying the method directly on soil.

The approach

Initial attempts on the application of Meta3C to soil environmental sample have shown that soil highly contaminates the sample, preventing creation of a sequencing library (figure 2.1, I. Stringlis not published).

We hypothesize that the main contaminants of soil are humic substances (**HS**). These compounds are negatively charged and can encompass DNA-protein complexes in the sample with a negative charge.[28] Because DNA has the same negative charge, DNA extraction is incomplete, because the humic substances will be extracted as well.

In order to eliminate HS we added steps of the PowerSoil DNA Isolation Kit (further referred to as 'the kit') that hold the 'Inhibitor Removal Technology'®. These buffers use a process called **flocculation** to remove HS from a sample prior to DNA analysis and so would be effective to clear our contamination's if our hypothesis on the nature of our contamination's was true.

Additionally, in order to fully test the effectiveness of the Meta3C method we applied the MetaTOR pipeline to published data (enriched LB sample from Marbouty et al. 2014 [1] as described above). Here, we compare its publication [2], GitHub tutorial and manual.

Chapter 2

Results

2.1 The laboratory protocol

As described in the introduction, the Meta3C protocol as described by Marbouty et al. 2014 [1] shows promise for use in metagenomic analysis of microbiomes, since it outperformed traditional WMS [2] methods. Specifically, it did this using less sample than was used to obtain the WMS data. This shows promise to decrease overall costs for metagenomic analysis of microbiomes. Additionally, the ability to obtain the genomic information of the individual genomes present in the complex sample and that the authors were able to reconstruct three dimensional reconstructions of genomic conformations [1] and detect host-virus interactions [2] makes this protocol promising to apply in our research into the microbiome content with relation to ISR inducing agents.

In order to apply the protocol to our samples, as described by Marbouty et al. [1] [24] it is important to understand the individual steps that are taken in order to obtain a paired-end sequencing library that can be used for computational analysis (as indicated in figure 1.2).

In table 2.1 these steps from figure 1.2 and those explained in section 4.2 are combined in a tabular overview. As can be seen in table 2.1, the steps described in figure 1.2 are global descriptions of the process. In practice these steps can be split into 11 individual steps:

First, the complete samples are fixed with Formaldehyde. This is done prior to lysis, to prevent any bias introduced in processing of the sample and thus capturing the conformation of the DNA as close to reality as possible. Formaldehyde creates a covalent bond between structures, mostly DNA and proteins, which keeps those structures that were interacting in nature, close to each other when processing the sample.

Second and third, lysis is done mechanically by beating the sample on a high frequency with glass beads. This is followed by a chemical lysis with sodium dodecyl sulfate (or SDS) to break cell walls that were not ruptured by the mechanical lysis.

Fourth, the DNA in the sample is digested with a specific restriction enzyme, HpaII. This enzyme recognizes and cuts the 4 bp region of C-CGG. We specifically use this enzyme because of its short restriction region. When one uses a restriction enzyme that restricts on a 6bp region, the resulting fragments are larger, because it would, statistically speaking, restrict less often than a 4bp restriction protein would. This might decrease the amount of interactions detected in the sample. This is why Marbouty et al. advice to use a restriction enzyme that recognizes a smaller restriction site. However, HpaII has an other bias: GC content. Baudry et al.[2] rightfully advise to create two libraries cut with different enzymes and mix them in for sequencing to circumvent this bias and any influences it might have on the overall process. We were not able to implement this in our research due to time constraints.

Fifth, the sample is diluted. This step is crucial to the success of the Meta3C protocol, because it ensures that sequences that were in interaction with each other in nature (the interactions fixed in by Formaldehyde) are ligated together, and prevents that sequences that are not in interaction ligate. The latter, when sequences that are not interacting in nature are ligated together, create a false interaction in the final data set, which could prevent correct reconstruction of three dimensional conformation of the genome, or create false links in the network that would complicate clustering of the network in computational analysis. Sixth, ligation occurs with no specific ligation enzyme since dilution specifies which sequences are ligated.

In step seven, we incubate the sample with glycine, overnight at a higher temperature to reverse the Formaldehyde fixation. This removes any potential proteins from ligated DNA complexes. Additionally, the sample is incubated with proteinase K to digest those removed proteins, improving sample quality.

Step eight through ten are steps to reduce the sample volume and extract the DNA from the soil-DNA mixture.

The final laboratory step of protein and RNA digestion ensure a pure DNA sample that is clean of inhibitors for downstream steps. In our case the sequencing steps.

During optimization, various of these steps were omitted to suit our needs, as is explained below and indicated as Alteration 1 - 3 in table 2.1.

Introductory figure	**Meta3C laboratory steps**	**Alteration 1**	**Alteration 2**	**Alteration 3**
1. Fixation	Formaldehyde fixation	-	-	Formaldehyde fixation
2. Cell lysis	Iterative bead beating lysis (Precellys beads)	Iterative bead beating lysis (Precellys beads)	Iterative bead beating lysis (Precellys beads)	Iterative bead beating lysis (Precellys beads)
	Chemical lysis	Chemical lysis	Chemical lysis	Chemical lysis
3. Digestion	DNA digestion with HpaII	-	-	-
4. Ligation	Sample dilution	-	Sample dilution	Sample dilution
	DNA proximity ligation	-	-	-
5. Purification	Formaldehyde fixation reversal in 65°C	-	-	Formaldehyde fixation reversal in 65°C
	Precipitation 1	-	Precipitation 1	Precipitation 1
	Phenol Chloroform extraction	Phenol Chloroform extraction	Phenol Chloroform extraction	Correct Phenol Chloroform extraction
	Precipitation 2	Precipitation 2	Precipitation 2	Precipitation 2
	Protein digestion and RNA digestion	Protein digestion and RNA digestion	Protein digestion and RNA digestion	Protein digestion and RNA digestion

Table 2.1: An overview of the laboratory steps taken in the Meta3C protocol as described in the first section of this report and adapted from Marbouty et al.[1]. The first column represents the steps as indicated in introductory figure 1.2. The second column shows the individual steps taken in the laboratory protocol as published[1]. The other columns indicate which steps were omitted during protocol optimization for direct soil application. For full explanation of individual steps in column 2 see materials section 4.2.

2.1.1 Soil complicates Meta3C protocol application

The above described protocol was initially developed using a mixture of three bacterial cultures. The authors continued to apply the protocol to a mixture of fungi and a sample of river sediment, enriched in Luria Broth (LB) overnight. There were no experiments available where this protocol is directly applied to soil in this publication, when we hypothesize that this would decrease biases introduced by enrichment steps or cultivation to such an extent that one would obtain a more accurate representation of the microbiome when Meta3C would be applied (assuming multiple restriction enzymes are used to prevent a bias there as well). To confirm this hypothesis and to investigate the earlier mentioned promise that the Meta3C method is described to have, we applied the above mentioned protocol to our soil samples. Development of the protocol for direct soil application was done on Rheyerskamp soil, whereas we aimed to apply the final protocol on rhizosphere samples related to previous research into the effect of ISR associated coumarins on the microbiome [8].

Initially, we applied the protocol to such a rhizosphere sample guided by the authors of Marbouty et al. 2014, as well as guided by their expertise in this protocol and use of their equipment in this method.

This resulted in a conformation that the application of the protocol to soil was possible, as some DNA was visible in the upper genomic size region in the gel shown in figure 2.1b.

Figure 2.1a is an experiment of the authors (not published) where they applied Meta3C on a pure cell culture, when comparing the level of the band shown here, this confirms our mention that our results in figure 2.1b was successful to some extent.

However, figure 2.1b also shows that this rhizosphere sample created a highly contaminated Meta3C library, as the area that shows a large area of the gel indicated contamination. Additionally, the final library was visibly brown and application of this sample to sequencing library preparation was unsuccessful.

Nevertheless, the DNA yield in this gel (figure 2.1) is promising and so we aim to replicate this yield in our laboratory.

(a) (b)

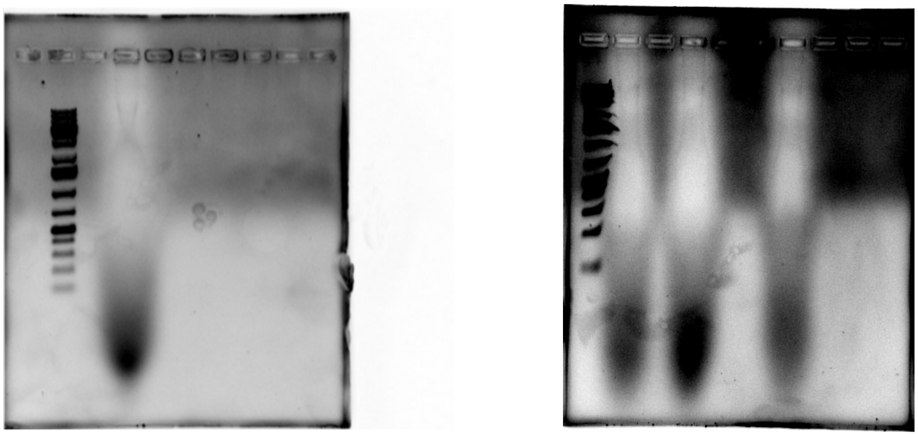
Figure 2.1: Two images from one agarose gel showing (**a**) The DNA resulting from a pure cell culture treated with the Meta3C protocol, as obtained in and by the Koszul lab during my supervisors visit.(**b**) Part of the agarose gel with the sample DNA obtained by my daily supervisor in the Koszul lab using a bulk soil sample. Note that the soil contaminates the sample, which results in the lower part of the sample run to take up a lot of the staining creating a smear that (a) does not contain. To visualise the small amount of DNA present in the sample (b) the agarose gel was incubated an additional hour in Ethidium Bromide.

2.1.2 The many steps of the Meta3C protocol complicates problem source identification.

The results of the direct replication of the Meta3C protocol described above, in our own lab and Rheyerskamp soil sample, resulted in figure 2.2ba. During the application of the protocol however, the sample did increase in volume after Phenol Chloroform extraction. It was split in order to obtain all DNA in the sample by precipitation and pooled into a single sample.

The DNA yield visible in gel (figure 2.2ba) is very low, especially, since an additional incubation of ethidium bromide was done on the gel in order to visualise the small amount of DNA in the upper part of the gel run. The lower part of this gel run resembles the shape and intensity of the result shown in figure 2.1b.

The low amount of DNA obtained in figure 2.2b and its resemblance to earlier attempts is promising. Thus, the next step in optimization was done. Here we applied the smaller beads used in the Power Soil DNA Isolation Kit (described in the introduction) as an initial attempt to increase DNA yield. In this experiment as well, the sample volume had increased but samples were not pooled to investigate the distribution of DNA in the sub-samples. This resulted in figure 2.2bb. The shape is again comparable to figure 2.2ba and 2.1b, however the DNA yield was much lower than both these previous figures. The image contrast was altered in this image to investigate the possibility of DNA in the upper part of the gel, but it appears there is little to no DNA here.



(a)

(b)

Figure 2.2: Two agarose gells resulting from initial attempts at replicating 2.1b. (**a**) A Rheyerskamp soil sample that went through the full Meta3C protocol as described in table 2.1. The shape of the sample (the contamination) on the gel represents that of figure 2.1b. Promising is also the small bit of DNA present int he upper part of this gel run. (**b** A repeat of the experiment that resulted in the gel shown in (a) with an altered lysis step (not shown in table 2.1) of using the lysis of PowerSoilDNA Isolation kit was applied here using 0.1mm beads for beating and a longer incubation time compared to Precellys beads in the Meta3C protocol. During application of this method, the phenol chloroform extraction was problematic: The volumes of the sample had increased. Because the following precipitation step requires a certain amount of sample, the increased volume sample was split to precipitate all of the sample available. This is why the different bands on this gel represent different sub-samples of the first original sample.

Due to the high number of steps described in the Meta3C protocol, it remains unclear which steps are responsible for the lack of DNA yield in our experiments. It is possible that the Phenol Chloroform extraction volume increase dilutes the final sample, thereby decreasing DNA yield on the gels. Other explanations include that soil contamination or any other individual steps in the protocol could be creating complications, preventing us from obtaining high DNA yields from our soil samples.

In order to determine which of these are the actual cause of a decreased DNA yield, we created a new protocol (Alteration 1 in table 2.1) in order to minimize resources during optimization and to decrease protocol complexity.

2.1.3 Protocol Alteration 1

This first alteration (Alteration 1 in table 2.1) omits Formaldehyde fixation, digestion and ligation from the initial Meta3C protocol. As can be clearly seen in table 2.1, the steps of formaldehyde fixation reversal and precipitation were omitted as well. This was done, because no decreased sample volume would be needed when no sample dilution was done, and no formaldehyde fixation reversal would be needed when no fixation was done.

Parallel to application of Alteration 1 on our soil samples, several experiments were combined to result in figure 2.4. Those experiments included:

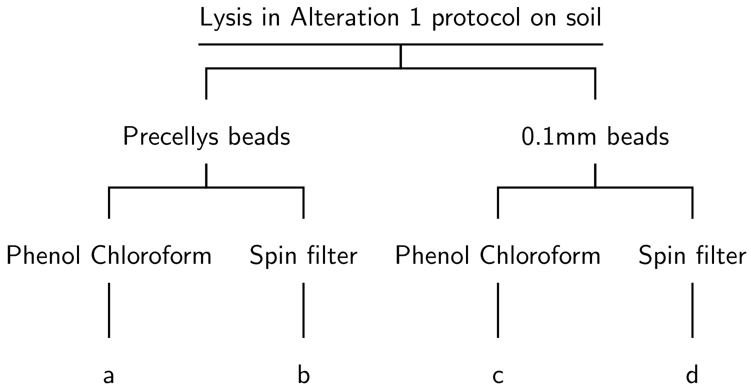


Figure 2.3: Schematic representation of the experimental setup to test the optimal lysis and purification technique to use comparing the Meta3C protocol and the PowerSoil DNA isolation Kit protocol. The 0.1mm beads and purification column were used by the Kit, whereas the Precellys beads and phenol-chloroform extraction were used by the Meta3C protocol.

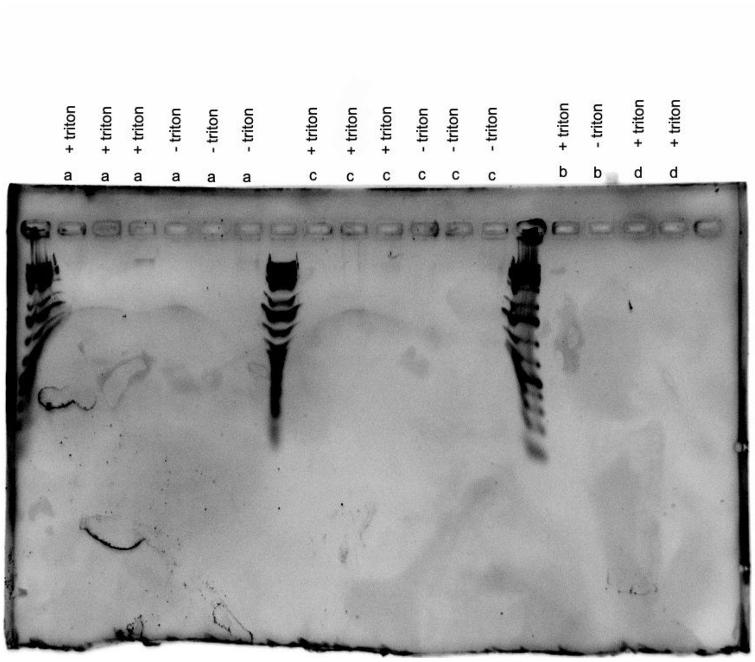


Figure 2.4: An agarose gel of Rhyerskamp soil samples treated with Alteration protocol 1 (table 2.1) which all showed a contamination of white precipitate after precipitation step in the Alteration 1 protocol (figure S4). The labels from figure 2.3 a, b, c and d correspond to the different lysis and DNA purification methods tested: (**a**) Lysis using Precellys beads, and phenol chloroform extraction (as Meta3C protocol) (**b**) lysis using Precellys beads, and spin column purification (**c**) lysis using 0.1mm beads, and phenol chloroform DNA extraction (**d**) lysis using 0.1mm beads, and spin column purification. The **+ or - triton** labels correspond to the addition of Triton to the mixture after chemical lysis with SDS in alteration protocol 1. Due to a volume increase during phenol chloroform extraction (like figure 2.2b) and the white precipitant, samples were split and not pooled after completion of the protocol. This created the multiple 'replicates' of, for example, 'a - triton'. Please note the contamination visible in the DNA ladders, black spots and general smears on this gel as well.

The addition or omission of Triton after lysis. We were unsure whether this should be done, as it was either a step after chemical lysis, necessary to halt lysis in the experiment, or a preparation for digestion and ligation. The latter were omitted in Alteration protocol 1 and so the Triton addition would be omitted with those steps. Specifically, the Meta3C protocol mentions to add ($500\mu l$) Triton 10% to the sample (of $1ml$). Figure 2.4 shows how half of the samples contain (+ triton) and half do not contain (- triton) as a step in the alteration protocol.

Additionally, in order to continue the lysis optimization started in the previous section, different lysis techniques were applied to the samples. These were combinations of the lysis and purification techniques described in the Meta3C protocol and the Power Soil DNA Isolation kit. Figure 2.3 shows how these steps using Precellys beads and Phenol Chloroform extraction from the Meta3C protocol, and the 0.1mm beads and spin filter from the kit to result in samples a, b, c and d respectively(see figure 2.3).

During the application of these experiments, the samples again increased in volume during Phenol Chloroform extraction. Samples were again split, resulting in, for example, multiple 'a + triton' samples in figure 2.4.

The resulting figure 2.4 of these tests was contaminated by a white precipitant visible after precipitation of the samples (see supplementary figure S4). It appears this precipitant has contaminated the gel as well, the ladders are deformed, black spots are visible, as well as a line throughout all samples. This contamination is likely due to the white precipitate, because it is seen over the gel in uniformity and was not present in previous experiments.

However, the band visible in the one 'a - Triton' sample is promising, since it shows DNA yield regardless of contamination and volume increases in the sample.

Nevertheless, it was necessary to clear the samples of this white precipitate in order to continue with this Alteration 1 protocol.

2.1.4 Solving protocol alteration complications

After some small tests (results not shown) it appeared that SDS and sodium acetate create a precipitate when mixed in sufficient amounts. The original protocol had a SDS concentration of 0.14316%, whereas the current Altered protocol was at 0.34%, approximately 2.4 times as high. This resulted in the formation of a precipitate of SDS and sodium acetate.

Initially, in the application of the protocol this white precipitate had formed during the first precipitation step, which indicates that this higher SDS concentration leads to the creation of a white precipitate, whereas this is not created when lower concentrations of SDS are used.

This led to the addition of a dilution step in Alteration protocol 2 (see table 2.1) where 10mL of pure grade water is added prior to precipitation. This dilutes the sample 10 times, which is more than sufficient to reduce the SDS concentration back to original concentrations, and brings the volume of the sample in this step to that of the original protocol.

As mentioned above, the previously obtained results in figure 2.4 were unreliable due to this contamination, except for the one 'a- triton' sample which did have a higher DNA yield. To confirm this finding and to confirm that Triton does not add to the DNA yield, the lysis with Precellys beads and phenol chloroform extraction without Triton (the a - triton from figure 2.4) was repeated in figure 2.5b + and -. Here, sample b+ shows that addition of Triton after lysis does increase DNA yield. However, due to the volume increase in this experiment (samples were pooled this time) the DNA yields can not be fully trusted in this figure.

For this reason, b+ was replicated in figure S3. Here, the yield of added Triton was comparable to that of figure 2.5 b+, indicating that the method of DNA extraction and lysis in samples b+ figure 2.5 are sufficient to obtain a DNA yield comparable to figure 2.1a. The method of Precellys beads and Phenol Chloroform extraction were for this reason applied in the further experiments for lysis and purification in our experiments.

2.1.5 Formaldehyde fixation affects DNA extraction from soil

Because the use of the Alteration 2 protocol does yield DNA we hypothesize that we can add a step of the original protocol to this alteration. Since an important part of the Meta3C method is fixating interactions to capture them, we decided to add the Formaldehyde fixation step to the Alteration 2 protocol.

The result of adding this step can be seen in figure 2.5 a+ and a- (+ and - again indicate the addition of Triton after lysis). These results show that there is no DNA extracted when the Formaldehyde fixation step is added. This indicates an effect of formaldehyde fixation on the DNA isolation from soil. It also shows that formaldehyde fixation makes it impossible to obtain DNA from the sample. Whether this was caused by the steps taken in Formaldehyde fixation, or by the nature of our sample, remained unclear.

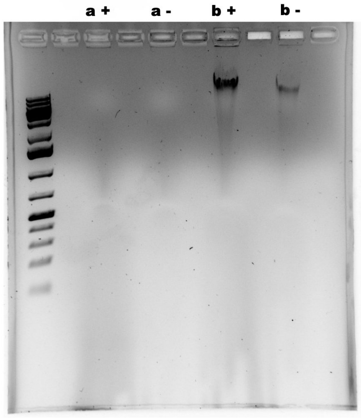


Figure 2.5: An agarose gel showing the DNA extracted by use of the second alteration protocol in table 2.1 with the addition **(+)** or the absence **(-)** of Triton using (**a**) formaldehyde fixed bulk soil sample and a (**b**) non-fixed soil sample.

2.1.6 Formaldehyde fixation affects DNA extraction from pure cell culture

In order to determine the origin of the absence of extracted DNA when Formaldehyde fixation was added to Alteration 2 and to confirm our methods, we decided to apply the Alteration 2 protocol with Formaldehyde fixation (as done in a+ in figure 2.5) to a pure cell culture sample. Additionally, this decision was based on the original protocol development by Marbouty et al. 2014 [1] where pure cell culture was initially used as well.

Our initial applications of this method to pure cell culture were complicated by the prescribed cell mass needed, when the resulting amount of DNA was too small to detect with the human eye after the precipitation step.

To solve this, we tested pure cell cultures of different sizes (25mL and 2mL instead of the prescribed 1mL of OD_{600} 1 cell culture) to obtain a cell pellet for fixation. Additionally, as a positive control a soil sample that was not fixed with formaldehyde and a negative control of Formaldehyde fixed soil sample were added to the experiment. Alteration protocol 2 was applied to these samples, resulting in figure 2.6.

Interestingly, the fixed cell culture samples do not show any DNA (figure 2.6a and b). This indicates that the effect of Formaldehyde fixation on the ability to extract DNA from a sample is not dependent on the nature of the sample, since both Formaldehyde fixed soil (d) and cell culture samples (a and b) do not show any DNA in figure 2.6.

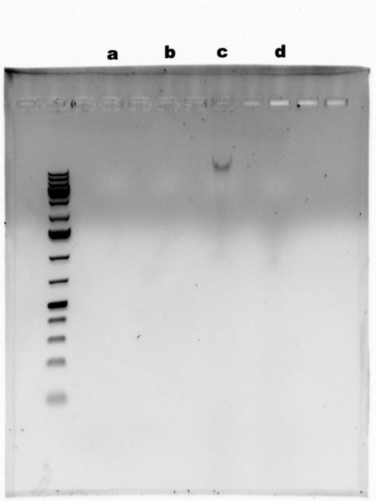


Figure 2.6: An agarose gel of (**a**) a formaldehyde fixed pure cell sample *Pseudomonas* WCS417r, 25mL of cell culture pelleted prior to formaldehyde fixation prior to application of Alteration 2 in table 2.1. (**b**)a formaldehyde fixed pure cell sample 1mL of cell culture pelleted prior to formaldehyde fixation, prior to application of Alteration 2 in table 2.1. (**c**) a positive control of application of Alteration protocol 2 (see table 2.1) to a non-fixed soil sample (**d**)a formaldehyde fixed soil sample used in application of the Alteration 2 protocol.

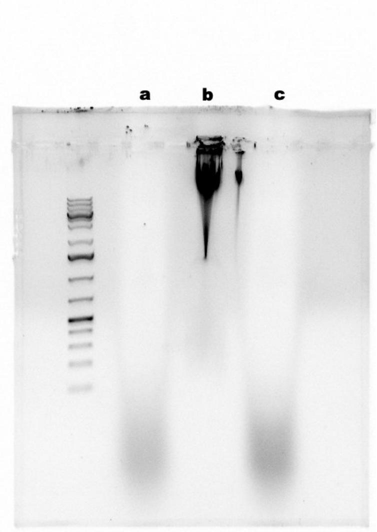


Figure 2.7: An agarose gel of (**a**) non-fixed soil sample with Alteration 2 protocol from table 2.1 used (majority of sample was lost during protocol application, but shown for it was unsure whether all sample was lost. The empty positive control in this case confirms that all sample was lost) (**b**) a large cell culture formaldehyde fixed and (**c**) a formaldehyde fixed soil sample. All had undergone the second alteration of table 2.1 with the addition of an overnight formaldehyde fixation reversal. The clear DNA isolated in **b** shows that this step is necessary in obtaining DNA from phenol chloroform extraction.

2.1.7 Formaldehyde fixation reversal is necessary for DNA extraction

We hypothesize that the omission of formaldehyde fixation reversal prevents DNA extraction during Phenol Chloroform extraction. The complexes that are fixed in Formaldehyde fixation would still be complexes during extraction, where the chemical properties of Phenol-Chloroform lead to extraction of negative charged DNA from the sample. Since the fixed complexes change the overall charges of DNA, it is likely that the majority of DNA is not extracted. Thus, we hypothesize that that addition of the Formaldehyde fixation reversal step to the current methods is necessary to extract DNA from any Formaldehyde fixed sample.

To confirm this hypothesis we created the Alteration 3 protocol (see table 2.1) where Formaldehyde fixation and Formaldehyde fixation reversal are added. This was then applied to the pelleted pure cell culture of 25mL (see figure 2.7b).

A non-fixed soil sample (a) and fixed soil sample (b) were added as a positive and negative control(see figure 2.7). The positive control (a) did not result in any DNA on this gel. We believe that this can be explained by a loss of sample during fixation, where most (or all) of the DNA in the sample was lost. Nevertheless, the result of the Alteration 3 protocol on the pure cell culture (figure 2.7b) confirms that Formaldehyde fixation reversal is necessary in order to obtain DNA from any Formaldehyde fixed sample.

The next step in application of the Meta3C protocol, was to add the digestion and ligation steps, returning to the original protocol. Additionally, we added a positive control of a pure cell culture that was used in an application of the Alteration 3 protocol. This resulted in figure 2.8a and b. The height of the band of sample (a) shown here are comparable to that of figure 2.1a, which also was a pure cell culture that was treated with the original Meta3C protocol. However, the nanodrop concentrations (table S3) of our sample, and the shape of the band in figure 2.8a indicate that the sample remains contaminated. Nevertheless, DNA yield of the Meta3C protocol on pure cell culture was of a sufficient level to continue with an application of this method on soil, as we can conclude that our methods are correct and other complications that would occur can be related to the nature of the sample.

Further, in a comparison of sample a and b in this experiment we can conclude that their differences in shape indicate a successful digestion of the sample, because the samples (a) and (b) solely differ in the presence or absence of the digestion and ligation steps. Whether this can be concluded about ligation remains unclear, as we are not able to predict the size of the sequences that would result from proximity ligation, when sizes of digested sequences are unknown. A biological replicate of sample (a) can be found in figure 2.12 where it was also used as a positive control.

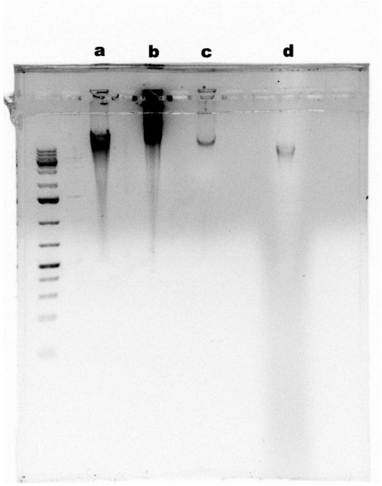


Figure 2.8: An agarose gel of samples (**a**) a Formaldehyde fixed pure cell culture that had undergone the full Meta3C protocol. (**b**) a Formaldehyde fixed pure cell culture that underwent the Alteration 3 protocol from table 2.1. (**c**) a Formaldehyde fixed soil sample used in application of the full Meta3C protocol, adding C2 and C3 buffers and steps associated with it from the Power Soil DNA Isolation Kit after lysis, to clean the soil from HS. (**d**) a soil sample underwent Alteration protocol 3 without Formaldehyde fixation. For accompanying nanodrop results see table S3

In order to fully confirm the success of the pure cell culture experiments, we aimed to sequence the pure cell culture sample. To do so in a reproductive manner, we aimed to obtain a DNA yield similar to that obtained by Marbouty et al. originally.

In figure 2.9 the Meta3C protocol was applied to a pure cell culture of 6mL (a) and 1mL(b) of 1 OD_{600} culture, the latter being the amount amount prescribed by Marbouty et al. 2014. [1]

Unfortunately, the contamination's clearly seen in the nanodrop concentrations associated with these samples in table S7 indicate a low quality of sample. This was confirmed when attempting sequencing library preparations, where no amplified sample could be obtained.

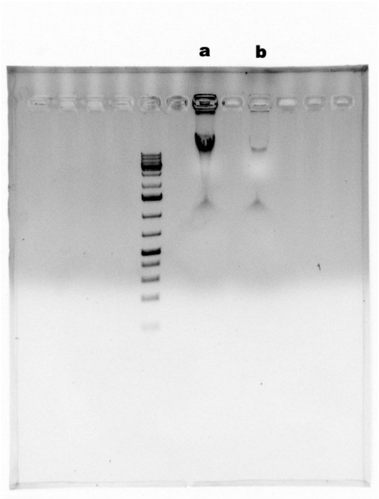


Figure 2.9: (**a**) fixed pure cell culture that has undergone full original Meta3C protocol. The pellet used contained 6 times the amount of cell mass as prescribed by the original protocol. (**b**) Fixed pure cell culture that has undergone full original Meta3C protocol. The pellet used is the amount used by the original protocol and therefor comes closest to the replication of figure 2.1. For accompanying nanodrop results see table S7

2.1.8 Comparison of PowerSoil DNA Isolation Kit and Meta3C methods

In parallel to above described pure cell culture experiments, using the original Meta3C method as well as the Alteration 3 protocol and additional cleaning steps, sample c in figure 2.8 was applied.

These cleaning steps were obtained from the Power Soil DNA Isolation Kit, alike the earlier described lysis alterations (figure 2.4 and 2.3). Comparison of the Kit and Meta3C methods in table 2.2 indicates a possibility of protocol optimization for soil samples. After lysis, this protocol applies two buffers (C2 and C3) and centrifugation steps to clean a sample.

The C2 and C3 buffers contain an 'Inhibitor Removal Technology', which according to the Kit protocol contains substances to remove humic substances. In order to confirm that these buffers are necessary to clean a soil sample from contaminants, we tested the Kit protocol with and without these steps (see figure S1 and S2). We confirm that these buffers interact in cleaning a soil sample from contaminants, and added these buffers and accompanying centrifugation steps to the Meta3C protocol. However, the location of the cleaning step remained to be determined prior to application of the cleaned Meta3C protocol. This was tested in a preliminary experiment where soil was treated with the full Meta3C protocol and cleaning steps with C2 and C3 buffers were added before lysis (a) and after lysis (b) (figure 2.10). Clearly, the application of C2 and C3 after lysis (figure 2.10b) yields more DNA than before lysis (a) and thus we applied the cleaning buffers C2 and C3 and steps after lysis to the experiments described above (figure 2.8c).

Additionally, a positive control was added of a soil sample that underwent the Alteration 3 protocol (figure 2.8d). From the result of figure 2.8c, cleaning the Meta3C protocol application to soil with C2 and C3 after lysis, we can conclude that these cleaning steps indeed successfully clear the soil sample of contaminants. However, the shape of this band does not represent that of figure 2.1a, which could indicate an aberration in our experiment. To confirm this indication we aimed to replicate the Meta3C protocol application with cleaning steps post-lysis on a soil and rhizosphere sample, resulting in figure 2.11. In both cases, there appears to be no DNA at the expected height as shown in figure 2.8c and the shape of the band resembles that of figure 2.2.

Introductory figure	**Meta3C laboratory steps**	**PowerSoil DNA Isolation Kit**
1. Fixation	Formaldehyde fixation	-
2. Cell lysis	Iterative bead beating lysis (Precellys beads)	Bead beating lysis (0.1mmbeads) and chemical lysis (bead beating buffer and C1 buffer
	Chemical lysis	-
3. Digestion	DNA digestion with HpaII	-
4. Ligation	Sample dilution	-
	DNA proximity ligation	-
5. Purification	Formaldehyde fixation reversal in 65 C	**Add C2 buffer, incubate in ice, add C3 buffer, incubate in ice**
	Precipitation 1	Load sample onto column
	Phenol Chloroform extraction	Spin column DNA extraction with buffer C5 and C6
	Precipitation 2	-
	Protein digestion and RNA digestion	-

Table 2.2: A comparison between the Meta3C method [1] and the procedure described by the PowerSoil DNA Isolation Kit [29]. Note the purification in the Kit includes buffers C2 and C3 (step indicated in bold) which contain the 'Inhibitor Removal Technology' according to the manufacturer and confirmed with preliminary tests (see figures S2, S1 and S1)

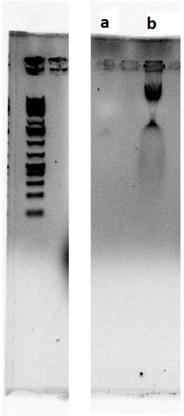
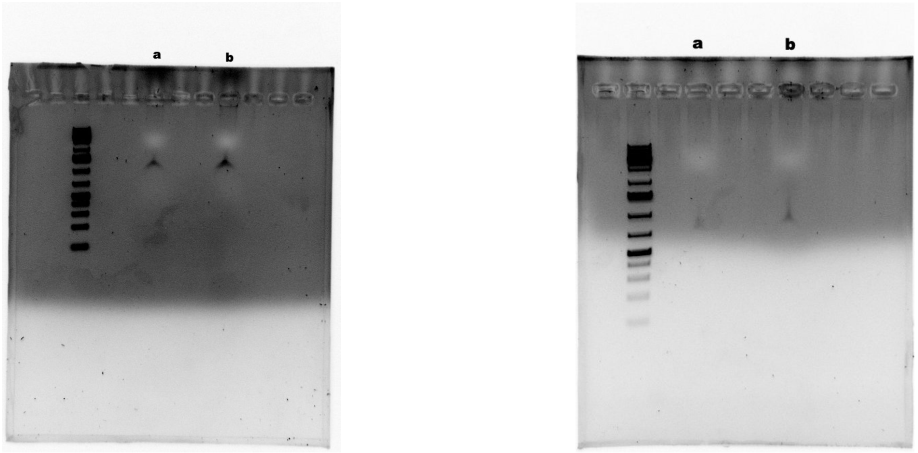


Figure 2.10: An agarose gel of two non-fixed soil samples that where only lysis and Phenol-Chloroform DNA extraction were applied. Sample (**a**) was cleaned before lysis and sample (**b**) was cleaned after lysis with C2 and C3 buffers. The samples differ in a cleaning of buffers C2 and C3 before (**a**) or after (**b**) lysis.



(a) (b)

Figure 2.11: Two replicates of (**a**)a rhizosphere sample and (**b**) a bulk soil sample were both fixed and treated with the full Meta3C including the cleaning steps with buffers C2 and C3 of the PowerSoil DNA Isolation Kit. These were attempts to replicate the results figure 2.8 sample c. For accompanying nanodrop results see table S5 and S6

Cleaning after formaldehyde fixation reversal is partially effective

We hypothesize that application of the cleaning steps after Formaldehyde fixation reversal could improve on the methods. However, when applying the Meta3C protocol with cleaning using C2 and C3 shortly before Phenol Chloroform extraction(figure 2.12b) we were not able to obtain an amount of DNA comparable to that obtained when applying Meta3C to pure cell culture in figure 2.12a. Nevertheless, the cleaning of the sample after formaldehyde fixation reversal did clean the sample when comparing it to figure 2.1b.

Since the C2 and C3 buffers are able to clean the sample from contaminants to some extent, we hypothesize that applying them in different concentrations might influence their effectiveness in our experiments. However, the contents of C2 and C3 remained undetermined. Thus, we explored the contents of C2 and C3 buffers by means of the manufacturers description and patent application [30] [31].

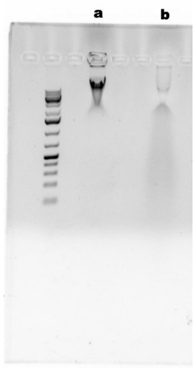


Figure 2.12: (**a**) replicate of a in figure 2.8, a full meta3C protocol on fixed pure cell culture. (**b**) A fixed soil sample that has been treated by the full meta3C protocol and cleaning between TE and chloroform extraction (so after fixation reversal) (**c**) full meta3C and cleaning after fixation reversal and DNA extraction using the spin filter. For accompanying nanodrop results see table S4

Inhibitor Removal Technology®

The Inhibitor Removal Technology, was described by the manufacture per buffer. Buffer C2 was described to contain : 'a reagent to precipitate non-DNA organic and inorganic material including humic substances, cell debris, and proteins.' And buffer C3 was described to be :' a second reagent to precipitate additional non-DNA organic and inorganic material including humic acid, cell debris, and proteins.' [30]

There is no mention of which chemicals are present in these buffers. However, this is described in the patent application of the Kit [31]. Here, the contents of C2 are described to be 133 mM Ammonium acetate and the contents of C3 are described as 120 mM aluminum ammonium sulfate dodecahydrate. Assuming these are solved in pure grade water.

In our familiarity with the protocol, the ammonium acetate in C2 can be determined as a chemical to help precipitation. The chemical in C3 (aluminium ammonium sulfate) is described in literature to be used in a method used for cleaning, for example, cleaning drinking water of contaminants: flocculation.

In flocculation, aluminum (or alum) is used to precipitate the humic substances and other debris from a water sample. [32] Based on the precipitating nature of flocculation in combination with ammonium acetate, we hypothesize that it is highly likely that our in our fixed soil samples, complexes of DNA, protein and HS are being precipitated and cleared from our sample.

Creating HS cleaning buffers from scratch

In an experiment to clear soil from HS. Cheng et al. [32] describe the use of a buffer containing aluminum sulfate. In contrast to our earlier findings of cleaning after lysis, the authors also advise to clean the samples prior to lysis to prevent the isolation of 'dead' soil DNA.

In an altered protocol where we use the cleaning with their buffers, lysis and phenol chloroform extraction described by Cheng et al. we can conclude that the statement to use cleaning pre-lysis in or samples is not applicable. Our results indicate the opposite, where cleaning before lysis (figure 2.13a) did not yield DNA, and where cleaning after lysis (figure 2.13b) showed more of a result, however no DNA in the height we expect based on previous isolated DNA from soil.

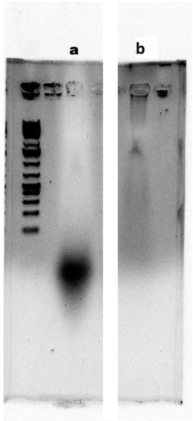


Figure 2.13: An agarose of samples that were tested for the optimal location to use C2 and C3 Inhibitor Removal Technology buffers from the PowerSoil DNA Isolation Kit. All samples were non-fixed Reijerskamp soil samples. **a** Replication of Cheng et al. [32] before lysis **b** technical replicate of (a) **c** C2 and C3 added pre-lysis **d** C2 and C3 added post lysis **e** Replication of Cheng et al. [32] after lysis. For accommodating nanodrop results see table S2.

Unfortunately, due to time constraints, we were not able to continue further optimization of the use of aluminum phosphate buffers. Since no sequencing library could be prepared from pure cell culture and DNA yields from soil were insufficient for sequencing, the computational analysis below was executed using a published data set from the Marbouty et al. 2014 publication.

2.2 Computational analysis of Metagenomic Interaction Data

2.2.1 The MetaTOR pipeline

In 2019 Baudry et al. published a paper in which they executed the Meta3C protocol on 20 mice feces samples (some resulting figures shown in figure 2.15). These samples were used to identify to what extent the

MetaTOR pipeline is able to obtain high-quality CCs (or metagenome-assembled genomes (MAGs)). The 82 high-quality CCs obtained by MetaTOR was compared to the amount of CCs obtained by using traditional binning methods MetaBAT and CONCOCT. MetaBAT and CONCOCT combine DNA characteristics like GC content to help contig binning. Their performance was lower than MetaTOR , thus MetaTOR in combination with Meta3C can therefor be considered a superior method for analysing metagenomic data. Naturally, this promising outcome emphasizes that our analysis should be done using MetaTOR as well. Additionally, MetaTOR was published by the same laboratory group as the data we are using here and so increases the likelihood that we are able to reproduce the results obtained by the authors of Marbouty et al. 2014 [1] [2].

Our analysis was initiated by use of the tutorial mentioned on the MetaTOR GitHub page [33] which explains the pipeline shortly and its commands would be ready to run locally after correct installation. More information on the functionalities of the individual commands were found in the paper of Baudry et al. [2] publication as well as the metaTOR manual [34].

The three resources for this one pipeline had some differences in application of commands and use of command flags. In order to determine the effect of these differences and to fully understand the steps taken by the pipeline in analysis of Meta3C interaction data we compared the commands as suggested by the different resources.

To have a frame of reference this was done on published data from Marbouty et al. 2014, specifically the data from their LB enriched river sediment experiment, as this is closest to our direct soil application described above. Please note that our analysis used this data resulting in figures 2.16, 2.18, 2.17 and 2.19, whereas the data and figures from figure 2.15 were created by the creators of MetaTOR by use of different data (the above described 20 mice species data from the Baudry et al. 2019 publication [2]) but are used for determining the success of individual pipeline steps.

Of the three resources, the manual explains the commands and flags in more detail. The tutorial creates a short overview of the pipeline and mentions that the [**metator pipeline**] command could be used directly for a fast run of the pipeline. The defaults that this command uses were found in the **pipeline.sh** file of the MetaTOR repository and are compared to the instructions of the tutorial and explanation in the manual.Additionally, the other commands of the tutorial that guide the reader through the pipeline step-by-step were followed, as described in the supplementary bash script. The initial metagenomic assembly that was created is described in this file as well. Lastly, the third resource: the publication associated with MetaTOR was used to obtain a more general idea of the functionality of the pipeline.

If any problems are found during replication of the current work, the reader is referred to this manual and the email addresses mentioned there for further help and explanation of individual pipeline steps. Additionally, the reader is referred to the bash script in supplements listing 6.1 for a full overview of the arguments used in the step-by-step tutorial run of the MetaTOR pipeline.

Installation

Prior to running the MetaTOR pipeline, the [**metator dependencies**] command was run in order to determine whether all dependencies of the pipeline were installed. Next, the [**metator deploy**] command was run from **root** to initiate the correct environment for the pipeline and determine configuration files and pipeline version. However, as shown in the supplemental bash script, the installation of SRA tools (for download of the data), IDBA-UD and CheckM tool kits are necessary prior to pipeline execution as well. The use for the latter two tool kits will become clear below.

The pipeline command

As mentioned above, the **metetor pipeline** command was run as a representation of all defaults used by the MetaTOR pipeline. The command was executed as follows:

```
metator pipeline -1 reads_forward.fastq -2 reads_reverse.fastq -a assembly.fa
```

Here, **-1** and **-2** indicate the forward and reverse FASTQ sequencing read files respectively. The **-a** flag indicates the pre-made metagenomic assembly that the pipeline needs to run the initial alignment step. The output of this command can be discussed and compared with the step-by-step run of the Github tutorial and will be discussed below along with the comparison of its defaults per individual step.

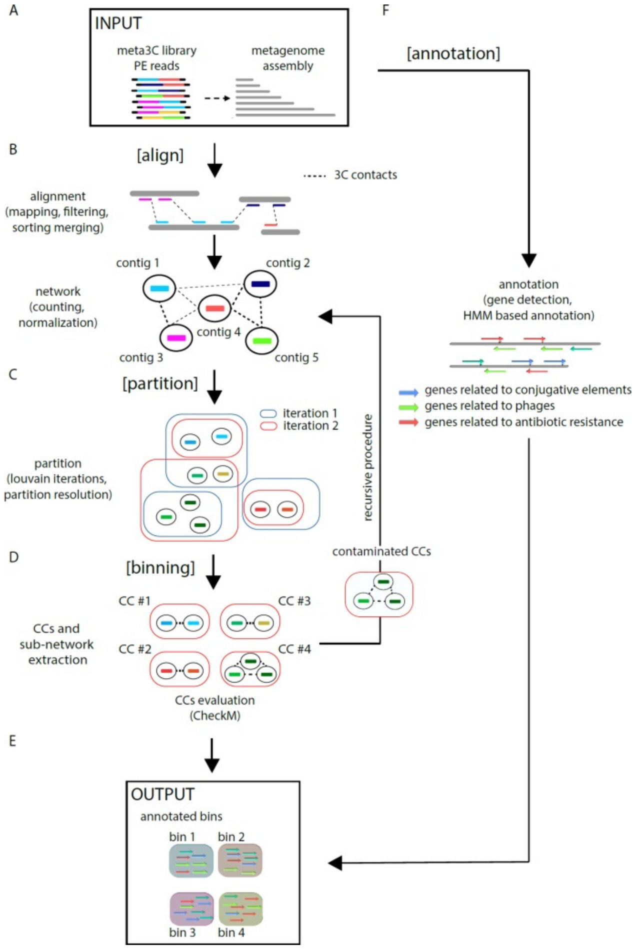


Figure 2.14: Schematic representation of the metaTOR pipeline, adapted from Baudry et al., 2019 [2]. (**A**) The input data consists of paired end (PE) reads resulting from NGS sequencing of a Meta3C library as well as a pre-made metagenomic assembly made with IDBA-UD where reads were considered single to detect long range interactions. Next, (**B**) the align command in the pipeline will align the PE reads to the assembly and create a weighted network of interactions where the contigs are nodes and their interaction frequency are the weight of interactions between these nodes. (**C**) The partition command will iterate the Louvain algorithm [3] over the network to create clusters. (**D**) The binning command will then identify core communities (CC's) within previous created clusters and subtract them from the network. (**E**) The pipeline output consists of bins corresponding to the CC's identified by the pipeline. (**F**) The metagenomic assembly and CC's are annotated with, among others, HMM based annotation. Note: the recursive procedure indicates where CC's that appear to hold multiple genomes (high contamination and high completion levels) can be partitioned and binned again in order to identify individuals within these communities. [2]

The align command

The first step in the MetaTOR pipeline is the alignment of the individual reads to the initial metagenomic assembly (see figure 1.2 step 8 and (**B**) in figure 2.14B). This way the read pairs can be used to create a network. In this network, the frequency of an interaction is calculated by counting how often a contig (or DNA chunk, for less size bias, explained below) is in contact with another contig, being how many reads that mapped to contig A also map to contig B. Contig A and contig B will be a node in the network respectively, and the interaction frequency the weight of the edge that connects the nodes in the network.

The [**metator align**] command takes an initial pre-made metagenomic assembly (the **-a** flag), forward FASTQ reads (the **-1** flag) and reverse FASTQ sequencing reads(the **-2** flag). It then uses Bowtie2 to align these individual reads to the pre-made assembly. Contrary to single genome assembly, no maximum mapping distance between read pairs is set, because of the long genomic distance that some interactions might have, since 3D interactions are obtained in Meta3C. In MetaTOR this is implemented by separately aligning the two read files to the assembly and then merging the assembly. [2][34][33]

First, the step-by-step tutorial states that the following command should be used:

```
metator align -1 forward.fastq -2 reverse.fastq -a assembly.fa -C 100000000 \
-Q 10 --clean-up -p example_project
```

When run in our local environment this command output an error while running Bowtie2. We were able to alter the command as well as re-run the [**metator dependencies**] command to make sure that all installations were correct. The command used here, as taken from the supplementary bash script, was:

```
align -1 ${FASTQ_FORWARD} -2 ${FASTQ_REVERSE} -a \
./contig.fa -c 10000 -C 10000 -Q 20 --clean-up -p ./metator_out
```

Which ran successfully, for as far as we could determine in the present, as there is no visual output of this step.

Interestingly, the argument used by the [**pipeline**] command and used by Marbourty et al. 2017 [24][34], is different from that proposed by the tutorial:

```
metator align -1 reads_forward.fastq -2 reads_reverse.fastq -a assembly.fa\
-Q 10 -c 500 -C 1000 -p another_new_project
```

Which has different values for the **-Q**, **-c** and **-C** values used in our independent run. It also does not have the –**clean-up** variable.

The **-Q** flag ensures that the quality of the assembly is high enough (based on the Phred quality score provided in the FASTQ metadata). Whether it is high enough is arbitrary and was therefore set to 20 in our step-by-step run, however the **metator pipeline** command takes a FASTQ quality of 10 as a minimum of reads to stay in the network later on.

Next, the **-c** and **-C** flags refer to the size of chunks made when creating a network. DNA chunks, to be specific, are made to prevent any size bias that contigs of different lengths create in the level of interaction frequencies. A larger contig will have a higher contact frequency, as there is more space for it to connect with other sequences, whereas smaller contigs will have less contact frequencies. When this is not accounted for in the means of 'chunking' the assembly, this means a size bias is present in the resulting network, making the binning, which is based on the interaction frequency, highly biased. The **-c** and **-C** flags in the command are the lower and upper boundaries respectively, that sequence chunks are allowed to be. The tutorial command here shows a very high upper boundary (**-C 100000000**), which , as mentioned in the tutorial description, relates to no chunking. It is unclear as to why this option is given, because this creates such a high bias in downstream steps that the resulting network can not be trusted to show correct contact frequencies. The tutorial explains to 'tweak the **-C** and **-Q** parameters', however alignment of these large data sets is very time consuming and so it would be wise to give a wider explanation as to how to tweak these to reasonable sizes in the tutorial for easier understanding of the pipeline, even when quick setup is in order. This is done in the pipeline command default and explained in the manual as well and therefor the blind pipeline command run would be better compared to individual runs of the pipeline steps.

There is one part of the tutorial command however, that is not mentioned in the manual, the publication and not explained in the tutorial: –**clean-up**. There is also **no –help flag** available to explain individual arguments in the pipeline either. As this argument is not used in the pipeline command we must assume there is no actual importance to use this command.

After these commands are run, the align command will have output some directories and files of which two files are of main importance. The first file, **network.txt** shows the nodes in the first two columns and the weight of these nodes in the third column. The manual and publication of the pipeline explain that this format was used for easy analysis with third party applications. However, when applying the network file to Gephy and Cytoscape, two commonly used software packages for network analysis, the .txt file was not accepted. This decreased the ability to inspect the network at this point in the analysis. We were only able to inspect the size of the output files and their raw format to obtain any information on whether this step was successful.

The second file that is output here is the **idx_contig_hit_size_cov.txt** file, which links the numbers in the **network.txt** file to the assembly and reads for use in later steps of the pipeline.

The partition command

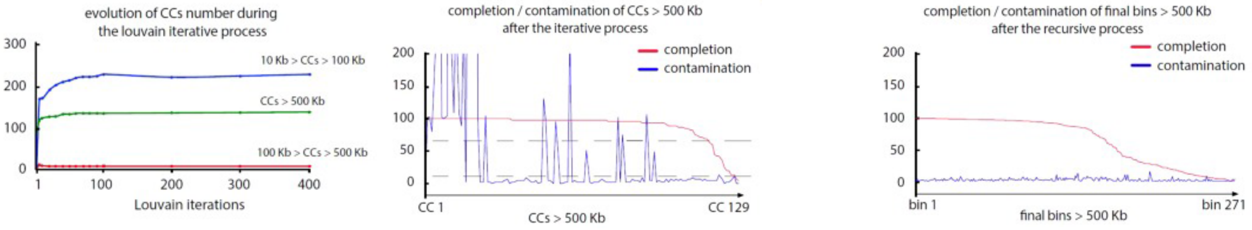
The second command executed in the MetaTOR pipeline is the partition of the network created in the (previous) align step. In this step, the network is clustered into bins (not yet core communities) by use of the Louvain algorithm (see figure 1.2 step 9 and (**C**) in figure 2.14). The manual of the MetaTOR pipeline states that the implementation in the pipeline was as the Louvain algorithm was originally published. [3]

In short, the Louvain algorithm will iterate over the network to create clusters that have a maximum amount of interactions within the clusters and a minimal amount of interactions between clusters (see also figure 1.1). For further detail on the specifics of the algorithm, the reader is referred to the original publication. [3]

The Louvain algorithm iterates over the network for a certain amount of times. All main pipeline resources (being the tutorial, manual [34] and publication[2]) mention how the partition command will output figures to determine the amount of iterations needed.

The published Meta3C protocol work from Marbouty et al. 2014 and 2017 used this pipeline, or early versions of this pipeline. In their 2014 publication the Louvain algorithm was iterated 100 times and the manual for MetaTOR mentions how the same was done in their 2017 publication.

In order to determine how many times the algorithm should be iterated over the data the pipeline outputs a figure comparing number of clusters found in the data, over the amount of iterations the algorithm has made. At the point of a plateau in the figure, it is superfluous to iterate more over your data, as the tutorial mentions. The 'metaTOR publication' by Baudry et al. mentions how the nodes that systematically cluster together over the selected 100 iterations are extracted and pooled into CCs, which was done in the Marbouty et al. [24] publication as well. The main text here simply states that the amount of CCs does not increase further after 100 iterations (like the tutorial) and therefor it was chosen to take 100. It is understandable to say that once the maximum amount of CCs is obtained one can say that the clustering is complete.



(a)

(b)

(c)

Figure 2.15: (**a**) Evolution of the number of CCs, ordered by size categories, during 400 Louvain (**b**) Completion (red) and contamination (blue) of the 129 CCs containing more than 500 kb after 100 Louvain iterations. Dashed lines: thresholds used to process CCs through a recursive procedure (completion threshold: upper 70%; contamination threshold: upper 10%) iterations for assembly nr. 3 (20 samples). Color represents the amount of DNA in a given CC. Blue: 10 to 100 kb. Red: 100 to 500 kb. Green: >500 kb. (**c**) Completion and contamination of the 269 CCs and sub-CCs bigger than 500 kb defined after the whole procedure. Red: completion. Blue: contamination.

In the tutorial the following command is used to do the partition step of the pipeline:

```
metator partition --iterations 300
```

Note that, due to alterations described below, the bash script includes the altered version of this command:

```
metator partition --iterations 100
```

Which appears to be a somewhat incorrect way to obtain the correct bins from the network as discussed below.

The **pipeline command** and manual however use a command with an additional argument (besides a specification of the output directory). As can be seen here:

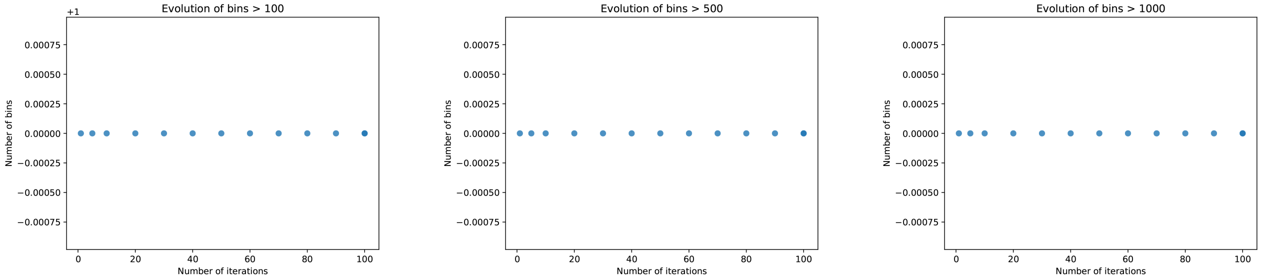
```
metator partition ---iterations 300 \--iter 100 \-p new_project
```

This additional argument is the [**-iter 100**], the manual explains that this selects the 100 bins (as the publication explained as well) to extract as CCs from the network. Whether doing iterations 100 times and selecting all bins, instead of 300 iterations and selecting 100 bins from the resulting clusters is better, is unclear and not elaborated in any of the resources.

The output of the **metator partition** step is the figure described above, that compares number of clusters to the amount of iterations. The publication of Baudry et al. 2019 [2] shows such a figure (see figure 2.15a) where the bins are sorted on size and for each bin size the number of bins is shown, mapped to the number of iterations. However, the pipeline does not output a figure where the three lines are accumulated into one figure like in figure 2.15a. Instead, it outputs individual plots of the individual lines in a 'regression plot'. The name of this plot is not explained in any of the resources.

The figures created by the pipeline in our analysis of the step-by-step tutorial run, shown in figure 2.16 are quite different than those created by the **metator pipeline** command in figure 2.17. Clearly, something was incorrect in running this command by using the step-by-step tutorial, since the number of bins/clusters remains at zero throughout all cluster sizes and iterations.

At this point, the resources as well as the command-line interface of the pipeline did not give information as to where this step had gone wrong in figure 2.16, therefor we continued to the next step of the pipeline to observe the effects of this silent error.

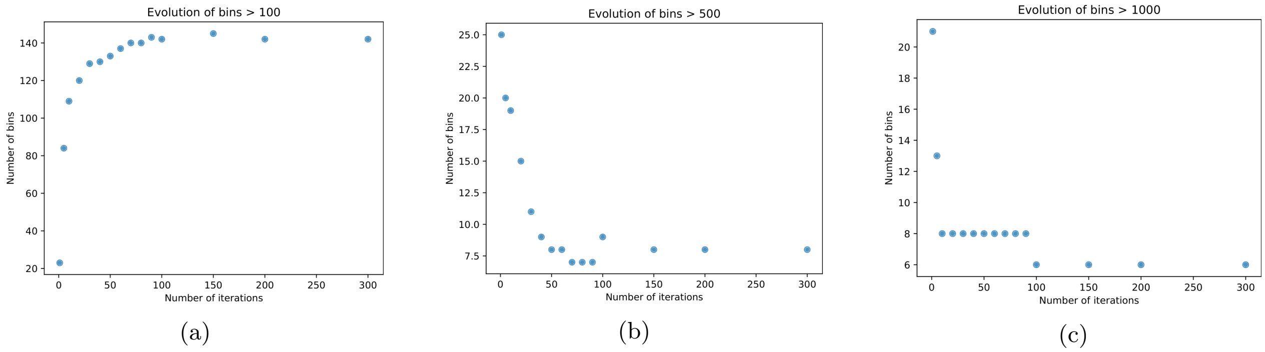


(a) (b) (c)

Figure 2.16: The three figures resulting from the [metator partition] command as prescribed by the step-by-step tutorial from the MetaTOR GitHub repository. The figures represent the evolution of bins/clusters in the network that have a size larger then (**a**) 100Kb (**b**) 500Kb and (**c**) 1000 Kb. Note that these lines should resemble, to some extent, the lines shown in figure 2.15a.

The execution of the **metator pipeline** command however, does seem to reach a plateau in figure 2.17a. This is comparable to the published figure 2.15a upper blue line shape, and therefore appears to be more successful than the step-by-step tutorial run.

Besides the visual output, the manual elaborates that the pipeline also outputs files in which the identities of chunks per bin are ordered to genomic size as well as number of chunks are present. Judging from the 2014 and 2017 paper it is this information that can be used to eventually obtain 3D reconstructions of the individual CCs which holds information on the natural state of genomes in these samples as well as information on the possible presence of viral genome, intergenomic or extragenomc. [1] [24] [34] We were not able to investigate this data in our runs, due to time constraints.



(a) (b) (c)

Figure 2.17: The three figures resulting from the pipeline command executing the partition command as explained by the MetaTOR manual. [34] The figures represent the evolution of bins/cluters in the network that have a size larger then (**a**) 100Kb (**b**) 500Kb and (**c**) 1000 Kb. Note that these lines should resemble, to some extent, the lines shown in figure 2.15a

The annotation command

The manual and tutorial describe how the annotation command can be run separately from the pipeline as the annotation is done on the initial assembly. This has been visualised in figure 2.14 as well. For this reason, the supplemental bash script executes this command after the 'fourth' binning step described below. However, in the **metator pipeline** command as well as the publication this step is described as the third step.

The manual states that the genes that are found on the assembly are mapped onto public hidden Markov model (HMM) databases. The elements that are predicted include conjugative elements, virus orthologous groups, single gene copies and essential genes. [34]

The command used by the tutorial here is simply:

```
metator annotation
```

However, this has to be run from the same environment as previous arguments in order to take the right files from the directories. The supplemental bash script has in this case not altered anything to this command.

The command used by the partition command and described by the manual however is:

```
metator annotation --evalue 1e-4 --size-contig-threshold 0\
-p another_new_project
```

The manual explains how the **–evalue** can be set to 0.1 or 0.001 or anything reasonable, relating to the statistical certainty that a certain element is predicted on a certain piece of the assembly. Additionally, there is the option to eliminate any small contigs that would be considered too small to represent communities in the metagenomic sample (using **–size-contig-threshold**). Here, it was set to 0 to annotate all bins. It is important to note however, that the tutorial does not mention the existence of these parameters, nor does it explain the defaults taken if one simply executes the command as shown above.

Output of this step can be inspected using CheckM, as described below.

The binning command

The final part of the MetaTOR pipeline is called 'binning', which should not be confused with partition, because no actual clustering is done in this step. Instead, the clustered network is used to extract bins. It uses the previously made annotations to annotate the bins as well, because the bins are linked to the original (now annotated) metagenomic assembly. [2].

The tutorial mentions to use the following command:

```
metator binning --n-bins 100
```

Where –**n-bins** determines the number of largest bins that are extracted from the network. The tutorial informs that increasing the number of bins is likely to extract small uninformative sequences, which is why this should be avoided. There were no alterations made to this command in the supplemental bash script.

The command used in the pipeline command and explained by the manual was:

```
metator binning --n-bins 100 --iter 300 -p new_project
```

Interestingly, it appears to be necessary to indicate the number of iterations again with 300, where this was done in the partition step as well. There is no explanation available as to why these were necessary to add again or why this was not done in the tutorial step.

The output of this step and thereby of the MetaTOR pipeline is the annotated bins within a respected folder per chunk in FASTA format, a folder with merged chunks and the sub network within the bin. [34] This last information, the sub network, is important for the step also shown in figure 2.14 where additional iteration can be done on the created contigs when contamination and completion levels of certain bins are high, for example when completion above 70% and contamination is above 10%, like done by Baudry et al. 2019[2]. Due to time constraints, we did not attempt any recursive binning and so we can not elaborate on the code presented in the tutorial and manual.

Core community quality determination

CheckM is a publicly available toolkit for assessing genome quality. It is a standard toolkit used when determining genome quality of individual genomes and meta genomes as well. [35]

The way to use CheckM to determine contamination and completion levels of the bins after running the MetaTOR pipeline is using the code snippets that follow. For more elaborate explanations on these commands the reader is referred to the publication and (online) documentation of the CheckM toolkit. [35]

First, some variables are initiated for readability of the commands that follow. The first variable refers to the location of the bins containing the merged contigs from the network. The second is the output directory for the results created by the CheckM toolkit. This directory is then created if it did not exist already (the **-p** flag).

```
fasta_dir="output/example_project/partition/iteration300/fasta_merged"
export checkm_dir="checkm_validation"
mkdir -p $checkm_dir
```

Next, **checkm tree** command is used on the merged bins as mentioned above. Here, the bins are placed in CheckMs reference genome tree. This tree is pre-made from a concatenation of 43 conserved marker genes with largely congruent phylogenetic histories. In other words, CheckM holds a genome tree that is used as reference for the bins thereby being able to later determine to what extent the bins are complete and/or contaminated. [35] According to the CheckM documentation of version 1.0.18 the **-x** flag simply assigns all files in the **fasta_dir** to be the only files considered in the analysis.

```
checkm tree -x fa $fasta_dir $checkm_dir
```

The third step is calling **checkm tree_qa** will determine the location of the bins within the previous mentioned reference genome tree in more detail. This is put into a summary file called **checkM_results.txt**. The **-o 2** flag specifies what information is included in this file. The exact code used is:

```
checkm tree_qa $checkm_dir -o 2 -f $checkm_dir/checkM_results.txt
```

Fourth, **checkm lineage_set** creates the **checkM_output_marker.txt** file, which determines the marker sets that the reference tree shows to be needed for the bins. So a set of marker genes that are required within the bins in order for them to be lineage correct/complete.

```
checkm lineage_set $checkm_dir $checkm_dir/checkM_output_marker.txt
```

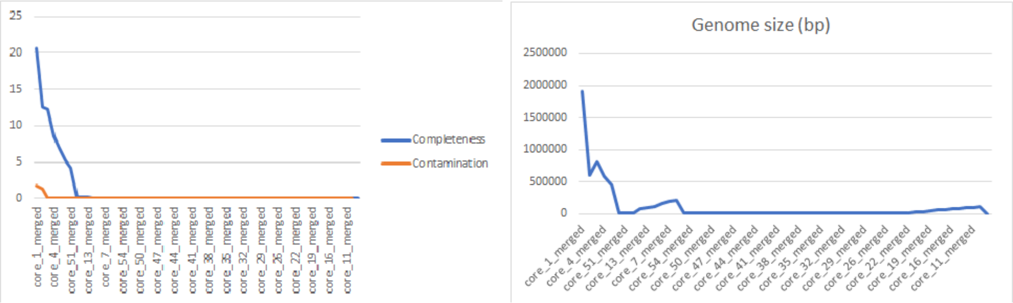
Fifth, Checkm analyze will take the previous marker sets and compare them to the merged bins and determine which of these markers are present in the bin.

```
checkm analyze -x fa $checkm_dir/checkM_output_marker.txt $fasta_dir\
$checkm_dir
```

This is then converted into completeness and contamination levels and stored in the **checkM_results_complete** file with again the **-o 2** flag specifying the information that is put in this summary file. For our specific purpose of recreating figure 2.15b and c, **-o 1** would have sufficed as well, as this simply gives contamination and completeness levels.

```
checkm qa -t 8 $checkm_dir/checkM_output_marker.txt ${checkm_dir} -o 2\
> $checkm_dir/checkM_results_complete.txt &
```

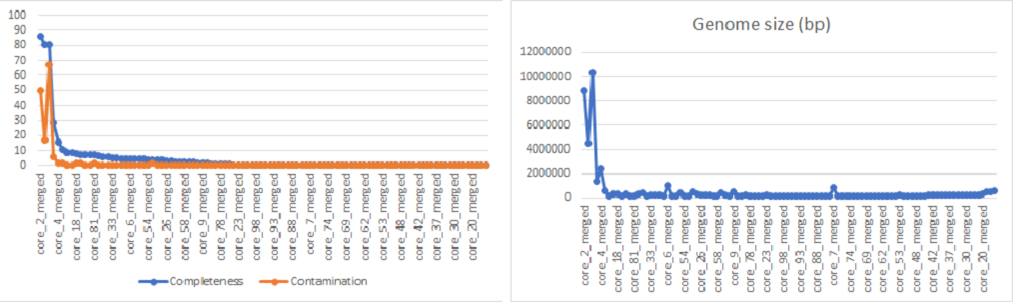
From this resulting summary table, **checkM_results_complete.txt** we can create the figures from Baudry et al. and compare how good the quality of our CCs are.



(a) (b)

Figure 2.18: Two graphs created from the output table of CheckM run on the MetaTOR pipeline output created by running the MetaTOR pipeline as suggested by its Github repository tutorial [33], specifically CheckM was run on the annotated bins it creates. **(a)** a comparison of the completeness and contamination in percentage (y-axis) of the bins (x-axis),**(b)** a graph representing the genome size in bp per bin.

For the tutorial based step-by-step run, the results of this figure, created in excel, are shown in figure 2.18. Note how the completion of the largest bin, **core_1_merged** is a little over 20% and drops to zero at the third largest core, **core_51_merged** in figure 2.18a. This is highly incomplete when one compares this to 2.15. There is the main difference that this published figure was based on an ensemble of 20 samples and the data used by us was not, however this should simply result in less bins, not necessarily bins that are this incomplete and a mere 3 where the published figures have an average of 82/20 = 4,1 high quality bins per sample. The genome size of the first few shown in figure 2.18b only show a level above 1500Kb base pairs for the first merged core, which indicates that this would be the only core to have an amount of bp that could correspond to an actual genome. This data, compared with the flat lines of bin evolution in figure 2.16 indicate that in this run, some complications have had an effect on the final CCs' quality.



(a) (b)

Figure 2.19: Two graphs created from the output table of CheckM run on the MetaTOR pipeline output created by running the MetaTOR pipeline in full defaults, using the **metator pipeline** command, specifically CheckM was run on the annotated bins it creates. **(a)** a comparison of the completeness and contamination in percentage (y-axis) of the bins (x-axis),**(b)** a graph representing the genome size in bp per bin.

The **metator pipeline** command run however, shows more promising results. Figure 2.19a shows how the first two bins are highly complete. However, they are also highly contaminated and like described before these would be bins suitable for recursive binning. The shape of this figure is comparable to that of the published run and because of the high contamination of the **core_4_merged CC**, it is likely to yield a comparable amount of CCs when one notes that only a single library was used for our data. The same can be said for figure 2.19b, where these bins have a genome size above 4000Kb even 20,000Kb indicating that these CCs are not individual CCs yet and recursive binning would be advised.

Taxonomic identity

Another property summarized by CheckM is the taxonomic classification of the individual bins. We have counted the amount of bins that are associated with a taxonomic identity or code by CheckM. The results of which can be seen in figure and figure 2.21 2.20.

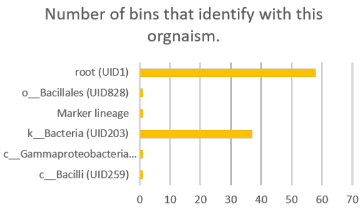


Figure 2.20: A graph created by counting the amount of times a certain taxonomic classifier was assigned to a bin in the CheckM analysis of the data created by execution of **metator pipeline** command using the MetaTOR pipeline on LB enriched river sediment data from Marbouty et al. 2019.

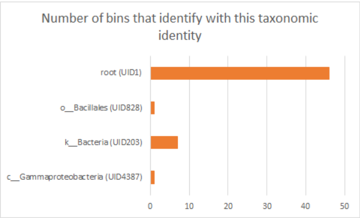


Figure 2.21: A graph created by counting the amount of times a certain taxonomic classifier was assigned to a bin in the CheckM analysis. This analysis was done on the data created by following the MetaTOR Github step-by-step tutorial as described in the supplementary bash script using the MetaTOR pipeline on LB enriched river sediment data from Marbouty et al. 2019.

The publication of the LB enriched river sediment sample analysis from Marbouty et al. 2014 describes the output of their analysis to be as described in table 2.3. The majority of species they found in their sample were Firmicutes and gamma proteobacteria. Firmicutes were not found in our analysis of this sample, as can be seen in both figures 2.21 2.20 Because these figures both resemble each other, which implies that this way to visualize and judge the output of the Meta3C and MetaTOR method is limited.

Additionally, the resulting figures 2.16 and 2.18 from the step-by-step tutorial run also raised suspicion as to the quality of the output bins when comparing them to 2.15, which indicates that it is easier to be critical to our results when using these figures than the taxonomic identities alone.

Species found in Marbouty et al. 2014 river sediment	Part of taxonomic classifier	Amount found
Aeromonas veronii	G-proteobacteria	marjority
Exiquobacterium sp.	Firmicutes	majority
-	Bacilli	several
-	Enterobacteria	several

Table 2.3: An overview of the assigned taxonomic classifications to bins obtained from analysis by Marbouty et al. 2014 published analysis of data obtained from LB enriched river sediment using Meta3C.

Chapter 3

Conclusion and Discussion

3.1 Key steps in the Meta3C protocol

The Metagenomic Chromosome Conformation Capture protocol is a challenging protocol to execute. Because of the number of steps within the protocol, including several key steps of formaldehyde fixation, formaldehyde fixation reversal and phenol-chloroform DNA extraction, it was hard to determine the origin of several complications in reproduction of earlier published results. [1] (figure 2.2).

In Meta3C, when samples are fixed, the formaldehyde fixation reversal is key in preparation to DNA extraction. Because the effect of formaldehyde fixation on DNA extraction was shown in soil (figure 2.5) and in pure cell culture (figure 2.7) it is likely that our hypothesis on Humic Substance interaction with the formaldehyde is only partly correct. The effect of formaldehyde fixation on pure cell culture indicates that the DNA-Protein complexes created from native source are already enough to prevent successful DNA extraction. [28]

The importance of the formaldehyde fixation reversal step in the protocol has to be stressed. Omitting this from the protocol inhibits DNA extraction. The Meta3C protocol can therefore not be executed in a time shorter than two days.

Further, phenol-chloroform extraction created a complication during practical applications of the protocol. The complications it created further stresses its importance in the overall process to obtain high DNA concentrations. We urge the reader to be aware of the phases present in the phenol-chloroform stock used and that solely the lower phase of the stock solution should be used to prevent sample dilution (volume increase) during DNA extraction.

Our attempts at using spin-filter DNA extraction instead of Phenol Chloroform extraction did not improve DNA yields. However, the gel that resulted from these tests was contaminated by the SDS-sodium acetate precipitation and should therefore be repeated to confirm this statement (see figure 2.4).

Important to note is that digestion and subsequent proximity ligation steps are necessary to make the protocol a Chromosome Conformation Capture protocol and must therefore never be skipped in experiments that are to be sequenced. The proximity ligation is what captures the 3D genomic information that will improve metagenomic binning as well as *de novo* genome assembly from mixed samples.

Due to time constraints I was unable to fully replicate the protocol of Marbouty et al. [1][26]. Replication was done successfully on a pure cell culture, as far as we can judge based on gel results. Illumina sequencing library preparation was attempted but did not yield any amplified DNA and therefor could not be sequenced.

Application of the protocol including cleaning steps using C2 and C3 buffer from the PowerSoil DNA Isolation Kit Inhibitor Removal technology on soil appeared successful in figure 2.8, but we were not able to replicate these results and therefore the protocol remains to be optimized for application on soil samples directly.

3.1.1 Future laboratory experiments

Because the replication of the pure cell culture was successful, a mixed cell culture or other clean sample would be suited for the protocol. As done by Marbouty et al. [1] by enriching their soil sample for eukaryotes with an overnight incubation in LB which still showed a large microbial diversity in the sample.

Another way to create a sample representative of the soil microbiome would be by using a firm wash of plant roots to obtain the rhizoplane (area around the root even closer than the rhizosphere). This protocol uses a Phosphate-buffered saline (PBS) buffer and vortexing to clean roots of the soil and other contaminants before shock freezing the roots (S. Poppeliers, not published). The resulting washed roots would be suitable for formaldehyde fixation and lysis as prescribed by the Meta3C protocol.

Once these replications are successful there are several options left for protocol optimization:

First, the creation of Inhibitor Removal Technology from scratch should not be neglected, because some complications in our experiments might have been made in creation of buffers for this experiment and some chemicals were not fully identical to those used in the publication, we used aluminium potassium sulfate dodecahydrate because it was readily available. However, the authors of the publication we based our methods on used aluminium sulfate [36]. It is unclear to what extent this difference has affected our results and should therefore be looked into in future research.

Second, Harkes et al. 2019 [37] describe improvements on this same technique for sandy soils that contain a range of organic matter. In their research, an protocol optimized for sandy soils was applied for successful DNA isolation in 104 samples. Their steps
”Add 2.5mL of 181mM disodium sulfate and 121mM guanidium thicyanate bead solution”
and
”Add 0.25mL of lysis buffer (150mM NaCl, 4% (w)̌SDS, 0.5M Tris) and 0.8mL of a 120mM ammonium aluminium sulfate dodecahydrate solution.”
Are promising to apply in our endeavours to clear the sample from soil contaminants as it is in alignment with the chemicals used by the Kit analysed in this work, as well as the publication described above. [37][36]

3.2 Computational analysis

Above described conclusions and suggestions refer to the laboratory protocol replications and optimization attempts of the Meta3C protocol. In theory, these practices would yield paired-end (Illumina) sequencing results. Because of the complexity of this sample it was of importance to understand the steps taken in analysis of the Meta3C paired end sequencing data.

Our experience with the MetaTOR pipeline indicate that this pipeline is already deprecated and gave several errors in python installation and a change in the source files of the program were needed in order to run the pipeline successfully.

Additionally, the tutorial for quick setup pipeline execution did not contain clear explanation of the parameters available and, if the parameters were available, failed to give indications on the right explanation of their function and/or what adjustments were needed on these parameters. For example, the explanation of the amount of Louvain iterations is too shallow. The main issue here is that the tutorial mentions how 100 iterations is enough, but this number could better be considered a minimal amount of iterations needed to identify the maximum amount of individual clusters present in the network.

This, mainly because the Louvain algorithm optimizes clustering based on the maximal amount of interactions inside a cluster and the least interactions between clusters. It will determine, by placing nodes in new clusters, how this influences these parameters and will find an optimal clustering there. More iterations that oscillate around the same number of clusters does therefor possibly change the contents of individual clusters. Simply stating that the reason to not iterate more is solely based on computational efforts is therefor lacking in explanation of the functionality of the additional iterations.

It might be useful to improve the networks by identifying the sequences that do not alter over all clusters when comparing results from runs of different iteration numbers. Thereby considering the option that these sequences are present in multiple clusters and therefore could be included in their final MAGs, perhaps improving assembly of repetitive sequences.

When identifying the individual steps of MetaTOR in the tutorial, their publication does elaborate by explaining the individual 4 main steps of the pipeline and their outputs. However, it is necessary to read the manual in order to fully understand how the sequences are treated, put into a network and clustered and binned into core communities that could be investigated in downstream analysis.

These statements are confirmed when comparing the results from the running tutorial commands and running the black box pipeline command. All figures resulting from the tutorial run (figure 2.16 and 2.18) do not represent the shapes suggested in the publication [2] (figure 2.15). This run was also not reproducible because a second run of the bash script in the supplemental gave new errors, which could not be solved within the time constraints of this project.

The black box **metator pipeline** command did appear to give somewhat correct results (figure 2.17 and 2.19 compared to published figure 2.15), however it mainly gave highly contaminated large bins that indicate that clustering was incomplete and recursive binning is suggested for future assessment of this data. It could also be possible that smaller chunks are needed since a large contig in bins could highly bias the clustering based on interaction frequencies and might have resulted in the one or two large clusters resulting from the pipeline command.

Analysis of the MetaTOR pipeline with CheckM was successful with regard to obtaining the contamination and completeness levels of the bins output by the pipeline. We also compared the taxonomic identities given by CheckM based on our less successful tutorial run and the **metator pipeline** command run. This comparison appears to be a low method to determining output quality of the bins and to what extent the pipeline could cluster the bins better than other methods. This is why we suggest future research to be done in comparing the identifications of CheckM based on a non-Meta3C data set from the same sample source to the identifications of CheckM on a Meta3C data set to fully obtain an assessment whether the promising results published by Baudry et al. 2019 [2] are reproducible.

When applying the MetaTOR pipeline to experimental Meta3C data we advise to use a combination of all three resources when running the pipeline as it is at the moment as none are completely self explanatory. Important to note here is how one argument, –**clean-up**, is missing from all sources and we can thus not determine its function.

For development of the pipeline we advised to explain parameters available in fast run tutorials as well, to prevent black box runs of the pipeline which limit the user of the pipeline in understanding the output of the pipeline fully.

Even though, the creators of MetaTOR state in the manual this pipeline is in development, the complications we encountered during the execution were very time consuming. This is why we can confirm that this pipeline is indeed still in development and can not be easily applied to the Meta3C data.

Nevertheless, I was able to investigate the individual steps necessary for analysis of Meta3C and refer the reader to figure 2.14 for an overview of the steps taken and to the manual for explanation of various parameters important in this analysis. [34] The methods of laboratory Meta3C and the use of MetaTOR in analysis do remain promising improvements on current metagenomic approaches and analysis, but many optimization and improvements remain until these methods can become a standard in the field of metagenomics.

Chapter 4

Materials and Methods

4.1 Samples

4.1.1 Soil source

Provided by the plant-microbe interactions group, rhizosphere samples and a bulk soil sample from previous research was used for the final optimized protocol[8]. For all intermediate steps bulk soil from the Reijerscamp nature reserve , was used.

4.1.2 Microbe cultivation

Pseudomonas simiae WCS417r was provided by my supervisor Dr. I. Stringlis. A sample was taken from -80°C and plated on KB agar plates overnight at 28°C. Next, a pure culture was selected and re-plated on fresh KB agar and grown overnight in 28°C.

For obtaining the pure cell cultures individual communities were grown in 25mL liquid Luria Broth (**LB**) medium overnight at 200rpm and 28°C. The KB agar plates were put at 4°C for reuse. The next day, 20 mL of stationary culture was diluted into 100mL final volume of liquid LB medium and grown to OD_{600} 1. As described in the results, culture volumes of 1, 2 and 25mL were spun down to create pure cell cultures. These were stored in -80°C until used for downstream analysis.

4.2 Metagenomic Chromosome Conformation Capture

For an overview of all reagents and equipment used in the laboratory approach of Meta3C see table S8 and table S9.

4.2.1 Formaldehyde fixation

The first part of the metagenomic chromosome conformation capture (Meta3C) protocol as described by Marbouty et al.,[1] was formaldehyde fixation.

For fixation, 25mL of Formaldehyde (Sigma-Aldrich) working stock was added to the sample and incubated for 30 minutes by shaking on max speed in the Multi Reax test tube shaker (see also table S9). The Formaldehyde working stock was prepared using 5 mL of 37% stock, diluted in 50mL H_2O to reach 3%. Samples were further incubated with a gentle shake at 4°C for 30 minutes.

Next, the fixation reaction was quenched using half the current sample volume (10mL) of Glycine (Sigma-Aldrich). From the stock, 93.84gr was added to 500mL H_2O to reach 2.5M working stock. This was followed by a 20 minute incubation at gentle shake at room temperature (RT). To pellet the sample, the centrifuge (Eppendorf) with a rotor that holds 50-mL Falcontubes, was set at 7830 rpm and 4 °C for 10 minutes. Finally, water was added to rinse the samples from Formaldehyde and Glycine. The samples were transferred to 5-mL tubes (Eppendorf) and the soil was pelleted at 4°C for another 10 minutes and stored in -80°C.

4.2.2 The original protocol: reproduction

The (fixated) samples all create one library each. The methods described apply to one library. First, the samples were defrosted in ice for 30 minutes and solved in 1x TE buffer (Promega) to fill a 2-mL tube.

Mechanical lysis

For lysis beads (Precellys) were added to the sample. This mixture was shaken horizontally for three cycles of 30 seconds at frequency of 30/s in the Tissue LyserII (Qiagen). In between each cycle, the samples were incubated in ice for 5 minutes. After the last 5 minutes, the beads and debris had settled down in the tube. The samples were not spun down, but up to 1mL of the solution was recovered and transferred to 1.5 mL tubes.

The beads and debris were then washed with 200 μL 1x TE 1x to obtain as much of the DNA present as possible. This solution was then added tot he previous mentioned 1.5mL tube.

Chemical lysis

Sixty μL 10% SDS (Invitorgen by Sigma-Aldrich) working stock was added to the sample and incubated for 10 minutes at RT. Working stock was prepared by diluting 5mL 20% stock in 5mL H_2O.

Digestion

The samples were then distributed over 5 tubes and 800μL digestion mix. The digestion mix contained 10% X-100 Triton for molecular biology (Sigma-Aldrich), 10x CutSmart digestion buffer (New England BioLabs) and 1000U restriction enzyme HpaII (R0171M NEB 50.000u/ml) solved in pure grade water of final volume 4020μL per library. The mixture of sample and digestion mix was split into 5 parts. The digestion incubation was done at 37°C, shaking at 200rpm for 3 hours.

Next, the samples were pelleted for 20 minutes in 4°C at 13.300rpm. The supernatant was removed with a pipette to prevent the pellets from breaking, as they were quite unstable at this point. These pellets were then re-suspended in 1 mL pure grade water and the 5 tubes are pooled into 2 50-mL Falcon tubes to contain equal amounts of the sample.

Ligation

For each tube, the ligation mixture contained 350μL 100mM Adenosine 5'-triphosphate disodium salt hydrate (ATP) (Sigma-Aldrich), and 250U T4 DNA Ligase (Thermo Scientific) in 14mL H_2O. One gram of 99% ATP stock was added to 14mL H_2O and 1.6mL 1M NaOH to reach 100mM pH 6.13 ATP working stock. This mixture was added to both Falcon tubes and incubated in a 4°C water bath for 16 hours.

Formaldehyde fixation reversal

Next, 0.5M EDTA (Merck), proteinase K solution (Bioline) and 10% SDSwere added to chelate ATP and digest proteins. Formaldehyde fixation was reversed by incubating this mixture overnight in 65°C stove [38]. Working stock EDTA was prepared by adding 2.8gr stock in 15mL H_2O, which was brought to pH 7 using NaOH and HCl. On the second day of the protocol, the sample was cooled at room temperature for 30 minutes and transferred to 2 new 50-mL Falcon tubes.

Precipitation 1

The samples were then precipitated using 1.6mL of 3M sodium acetate (Merck) and 16mL 2-propanol (Sigma-Aldrich). Sodium acetate working stock was prepared by adding 204.12gr stock in 400mL H_2O). This was followed by an incubation of 1 hour at -80°C. Next, the DNA was pelleted for 20 minutes in 4°C at 13.300rpm and the supernatant was removed. The pellets were air-dried at room temperature for 5 minutes and 900μL of 1x TE was added.

Phenol Chloroform DNA extraction

To solve the pellets, the tubes were placed in a 37°C incubator, shaking at 200rpm at an angle for 15 minutes after which phenol chloroform extraction as performed by using 900μL Phenol Chloroform (Sigma-Aldrich) and centrifuging at 13.300rpm for 5 minutes. The supernatant was taken and split, adding 200 μL per 1.5-mL tube.

Precipitation 2

To these tubes, 40μL of 3M sodium acetate (pre-made stock as mentioned in precipitation 1 above) and 1.1mL of cold 100% Ethanol (Sigma-Aldrich) was added. The samples were then incubated for a minimum of 30 minutes in -80°C.

Next, the now precipitated DNA was pelleted at 4 °C 13.3000 rpm for 10 minutes and the pellets were air-dried in 37°C for 5 to 10 minutes.

Protein digestion and purification

Last, 30μL of Tris buffer (pre-made stock using 1 M Trizma hydro chloride (Tris buffer) (Merck) and 2 μL 5% RNAse (GeneAid) was added to the mixture. In 500mL H_2O, 78 gr Tris was added and NaOH and HCl to adjust the pH to 7.5 were added to create the working stock. Additionally, 20 μL of stock RNAse was added to 400 μL Tris to create an overall working stock for this reaction. The reaction was then added to the sample and incubated for another 30 minutes in a heat block at 37°C. Final pellets were eluted in 30 μL Tris buffer.

4.2.3 Optimization alterations

For optimization of the metagenomic 3C protocol, subsets of the original Meta3C protocol were created. These subsets included versions that would or would not have specific steps of the original protocol as described above. The different alteration protocols have been visualized in table 2.1.

4.2.4 Humic substance and soil particle removal

In order to remove humic substances and soil particles from the fixated and non-fixated samples, the C2 and C3 buffer of the PowerLyser® PowerSoil® DNA Isolation Kit (Qiagen) were used.

Humic substance removal using the Power Soil DNA isolation Kit

Samples that were treated with C2 and C3 were also treated with the complementary steps from the manufacturers protocol: From C3, 250 μL was added to a sample of preferably 400-500 μL (or a sample split into these parts) and incubated for 5 minutes on ice. Next, samples were centrifuged at RT for 1 minute at 13.300 rpm. Next, 600μL was transferred to a new tube and 200μL C2 was added.

After an incubation of 5 minutes on ice followed by a centrifugation of 1 minute at 13.300 rpm the sample was used for downstream steps of the Meta3C protocol, depending on the location that the cleaning steps were added.

Recreation of C2 and C3 PowerSoil DNA Isolation Kit

As described by Cheng et al. [32] 300mg soil was dissolved in 300 μL 0.1M phosphate buffer. This was obtained by adding 7.2gr Na_2HPO_4 to 20mL H_2O to reach 1M, and adding 2.8gr NaH_2PO_4 to reach in 20mL H_2O to reach 1M. These were were combined to create a working stock phosphate buffer of pH 6.72. Next, 200μL 100mM aluminium potassium sulfate dodecahydrate (Merck) was added to the mixture to precipitate soil particles. By adding 23.7gr in 0.5L H_2O 100mM working stock was prepared.

The sample was then vortexed for 2 minutes at max speed and 250 μL 3M NaOH was added to adjust the pH to 8.0, thereby halting the precipitation. Next, 250μL of lysis buffer was added prior to adding beads. This buffer contained 100mM NaCl, 500 mM Tris (pH 8.0) and 10% SDS. The tubes were lysed in the Power lyser II (Qiagen) at a frequency of 30/s for 10 minutes. The tubes were then centrifuged at 13.300rpm for 30 seconds and the supernatant was transferred to a new tube.

Then phenol chloroform extraction and precipitation with ethanol and ammonium acetate was performed alike the Meta3C protocol described above. Final pellets were eluted in 30 μL Tris buffer.

4.3 Sequence library preparation

4.3.1 Covaris

To create fragments of 350bp the Covaris aparatus located at the Useq facilities was used. An amount of 100ng DNA was solved in 60uL low TE (prepared by PhD student Sanne Poppeliers) and samples were transfered to the appropriate Covaris tubes.

4.3.2 Adapter ligation and amplification

Further library preparation was performed using the NEBNext® Ultra ™ II DNA Library Prep Kit for Illumina®. The Adaptor used was from a kit numbered #E6612A, the USER enzyme from the #E6610A kit and the other kit consumables were labeled #E7374A.

Size selection options for 300-400bp were used and 15 cycles in the PCR amplification (in BioRad S1000 ™ Thermo cycler). Primers used in amplification can be found in table 4.1.

Primer	Sequence
E661A_P2_A2 (forward)	5'- CAA GCA GAA GAC GGC ATA CGA GAT AGG TCA CTG TGA CTG GAG TTC AGA CGT GTG CTC TTC CGA TC*T -3'
NEB_universal (reverse)	5' - AAT GAT ACG GCG ACC ACC GAG ATC TAC ACT CTT TCC CTA CAC GAC GCT CTT CCG ATC*T -3'

Table 4.1: Primers used in library preparation 26/7/2019

4.4 Computational analysis

For computational analysis all work was performed on the PMI server maintained by Dr. Ronnie de Jonge.

All steps taken in the analysis of the pipeline can be found in the bash script in supplements listing 1.

Chapter 5

Acknowledgements

A special thanks to Dr. Giannis I. Stringlis and Dr. Ronnie de Jonge for the daily supervision, lively discussions and critical feedback on this project. Also all appreciation to the Plant Microbe Interactions Group lead by Prof. Corné Pieterse at Utrecht University for providing a welcoming environment for my research. A last thanks to the Useq facilites for the use of their Covaris machine and to Sanne Poppeliers for helping me with sequencing library preparations in the lab.

Bibliography

[1] M. Marbouty et al. "Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms". In: *eLIFE* 3 (2014). DOI: 10.7554/eLife.03318.

[2] L. Baudry et al. "MetaTOR: A Computational Pipeline to Recover High-Quality Metagenomic Bins From Mammalian Gut Proximity-Ligation (meta3C) Libraries". In: *frontiers in Genetics* 10 (2019).

[3] V. D. Blondel et al. "Fast unfolding of communities in large networks". In: *Journal of Staistical Mechanics: Theory and Experiment* (2008). DOI: 10.1088/1742-5468/2008/10/P10008.

[4] L. Garmedia et al. "Metagenomics and antibiotics". In: *Clinical Microbiology and Indection* 18 (2012), pp. 27–31. DOI: 10.1111/j.1469-0691.2012.03868.x.

[5] C.M.J. Pieterse et al. "Induced Systemic Resistance by Beneficial Microbes". In: *Annual Reviews Phytopathology* 52 (2014), pp. 347–375. DOI: 10.1146/annurev-phyto-082712-102340.

[6] R.L. Berendsen, C.M. Pieterse, and P.A. Bakker. "The rhizosphere microbiome and plant health". In: *Trends in plant science* 17 (8 2012), pp. 487–486. DOI: 10.1016/j.tplants.2012.04.001.

[7] G.J. Benoit Cnonlonfin, A. Sanni, and L. Brimer. "Review Scopoletin - A Coumarin Phytoalexin with Medicinal Properties". In: *Critical Reviews in Plant Sciences* 31 (2012), pp. 47–56. DOI: 10.1080/07352689.2011.616039.

[8] I A. Stringlis et al. "MYB72-dependent coumarin exudation shapes root microbiome assembly to promote plant health". In: *PNAS* 115 (2018), pp. 5213–5222. DOI: 10.1073/pnas.1722335115/-/DCSupplemental.

[9] I.A. Stringlis, R. de Jonge, and C.M.J. Pieterse. "The Age of Coumarins in Plant-Microbe Interactions". In: *Plant and Cell Physiology* 60 (7 2019), pp. 1405–1419.

[10] R. van Peer, G.J. Niemann, and B. Schippers. "Induced resistance and phytoalexin accumulation in biological control of Fusarium wilt of carnation by Pseudomonas sp. strain WCS417r". In: *Phytopathology* 81 (1991), pp. 728–734.

[11] R.L. Berendsen et al. "Unearthing the genomes of plant-beneficial *Pseudomonas* model strains WCS358, WCS374 and WCS417". In: *BMC Genomics* 539 (2015).

[12] B. Lugtenberg and F. Kamilova. "Plant-Growth-Promoting-Rhizobacteria". In: *Annual Review of Phytopathology* 63 (2009), pp. 541–556. DOI: 10.1146/annurev.micro.62.081307.162918.

[13] C. Quince et al. "Shotgun metagenomics, from sampling to analysis". In: *nature biotechnology* 35 (2017). DOI: 10.1038/nbt.3935.

[14] J.D. Elsas van. et al. "The metagenomics of disease suppressive soils- experiences from the METACONTROL project". In: *Trends in Biotechnology* 26 (11 2008), pp. 591–601. DOI: 10.1016/j.tibtech.2008.07.004.

[15] M.J.E.E.E. Voges et al. "Plant-derived coumarins shape the composition of an *Arabidopsis* synthetic root microbiome". In: *PNAS* 116 (25 2019), pp. 12558–12565.

[16] S.S. Sandhu, A. Pourang, and R.K. Sivamani. "A review of next generation sequencing technologies used in the evaluation of the skin microbiome: what a time to be alive". In: *Dermatology Online Journal* 25 (7 2019).

[17] J. Kuczynski et al. “Experimental and analytical tools for studying the human microbiome”. In: *Nature Reviews Genetics* 13 (2012), pp. 47–59. DOI: `10.1038/nrg3129`.

[18] L.E. Kafetzopoulou et al. “Metagenomic sequencing at the epicenter of Nigeria 2018 Lassa fever outbreak”. In: *Science* 363 (6422 2019), pp. 74–77.

[19] E.K. Binga, R.S. Lasken, and J.D Neufeld. “Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology.” In: *The ISME journal* 2 (2008), pp. 233–241. DOI: `1751-7362/08`.

[20] J. Alneberg et al. “Binning metagenomic contigs by coverage and composition.” In: *Nature Methods* 11 (2014).

[21] D. Kang et al. “MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities”. In: *PeerJ* 3 (2015).

[22] P. Menzel, K.L Ng, and A. Krogh. “Fast and sensitive taxonomic classification for metagenomics with Kaiju.” In: *Nature communications* 7 (2016), p. 11257. DOI: `10.1038/ncomms11257`.

[23] A. Sczybra et al. “Critical Assessment of Metagenome Interpretation - a benchmark of metagenomics software”. In: *Nature Methods* 14 (2017), pp. 1063–1073. DOI: `10.1038/nmeth.4458`.

[24] M. Marbouty et al. “Scaffolding bacterial genomes and probing host-virus interactinos in gut microbiome by proximity ligation (chromosome capture) assay”. In: *Science Advances* 3 (2017).

[25] J. Dekker et al. “Capturing Chromosome Conformation”. In: *Science* 295 (2002), pp. 1306–1311.

[26] M. Marbouty et al. “Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay”. In: *Science Advances* 3 (2017).

[27] T. Foutel-Rodier et al. “Generation of a Metagenmeta3omics Proximity Ligation 3C Library of a Mammalian gut Microbiota”. In: *Methods in Enzymology* (2018). DOI: `1.1016/bs.mie.2018.08.001`.

[28] J. E. Tomaszewski, R. P. Schwarzenback, and M. Sander. “Protein Encapsulation by Humic Substances”. In: *Environmental Science and Technology* 45 (2011), pp. 6003–6010. DOI: `10.121/es.2663h`.

[29] MO BIO Laboratories. “Inhibitor Removal Technology®”. In: 02232016 ().

[30] MO BIO Laboratories. “PowerLyzer® PowerSoil® DNA isolation Kit”. In: *Instruction manual* 0727216 (2016).

[31] M.N. Brolaski, R.J. Venugopal, and D. Stolow. “Kits and processes for removing contaminants from nucleic acids in environmental and biological samples”. In: *United States Patent US7459,548* 45 (2008).

[32] W.P. Cheng, F.H. Chi, and R.F. Yu. “Effect of phosphate on removal of humic substances by aluminum sulfate coagulant”. In: *Journal of Colloid and Interface Science* 272 (1 2004), pp. 153–157.

[33] Koszul laboratory. *metaTOR Github page*. URL: `https://github.com/koszullab/metaTOR`. (accessed: 09.12.2019).

[34] L Baudry et al. “MetaTOR user manual”. In: (2018).

[35] D.H. Parks et al. “CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes.” In: *Genome Research* 25 (7 2015), pp. 1043–1055.

[36] D. Dong et al. “Removal of humic substances from soil DNA using aluminium sulfate”. In: *Journal of Microbiological Methods* 66 (2006), pp. 217–222. DOI: `1.116/j.mimet.2005.11.010`.

[37] P. Harkes et al. “Conventional and organic soil management as divergent drivers of resident and active fractions of major soil food web contsituents”. In: *Nature Scientific Reports* 9 (2019), p. 13521.

[38] E. A. Hoffman et al. “Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes”. In: *The journal of biological chemistry* 290 (2015), pp. 26404–26411. DOI: `10.1074/jbc.R115.651679`.

Chapter 6

Supplements

6.1 Figures

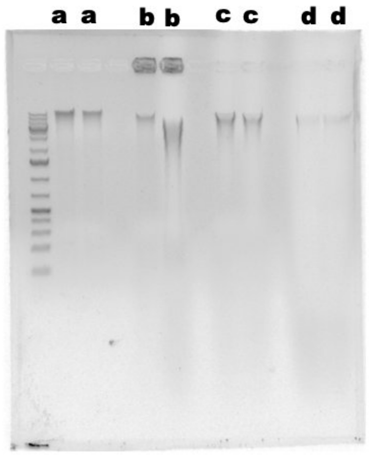


Figure S1: An agarose gel for samples (**a**): full protocol. (**b**) no C2 and 5+6 (**c**) no C3 **d**: no C2 and no C3. The best way to see the difference however is by looking at the tubes directly(figure S2, and the nanodrop qualities (table S1

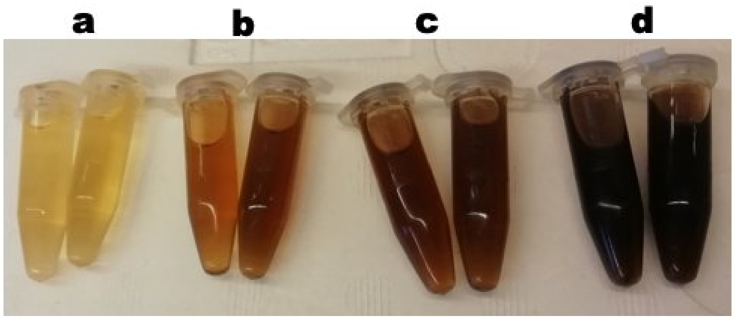


Figure S2: Eppendorf tubes containing samples that were processed with the PowerSoil DNA Isolation kit. Specifically, this was a test to identify the necessity of using C2 and C3 buffers in the protocol. Sample**a**: full PowerSoil protocol. (**b**) no C2, (**c**) no C3 and (**d**) no C2 and no C3.

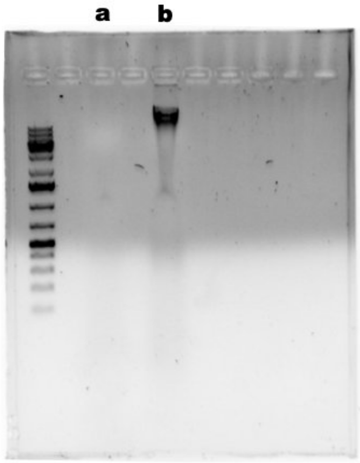


Figure S3: An agarose gel showing DNA extractions using the last alteration protocol form table 2.1 as well as added Triton after lysis. A comparison between the formaldehyde fixated sample (**a**) and non fixated sample (**b**) shows that formaldehyde has an effect on DNA extraction.

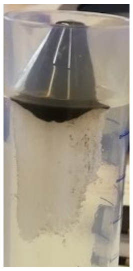


Figure S4: Soil sample b in figure 2.4 after precipitation and centrifugation when applying Alteration protocol 1. All 4 samples a,b,c and d of figure 2.4 showed this precipitate.

6.2 Tables

	a	**a**	**b**	**b**	**c**	**c**	**d**	**d**
ng/μL	27.8	22.8	478.2	1290.5	1081.6	1130.4	-19.2	2683.5
260/280	1.93	1.93	1.33	1.37	1.35	1.36	1.51	1.36
260/230	1. 33	1.44	0.74	0.79	0.75	0.76	0.63	0.76

Table S1: Nanodrop results from samples (**a**), (**b**), (**c**) and (**d**) as described in figure S1 and figure S2. Note the low 260/280 values for samples c and d

	a	**b**	**c**	**d**	**e**
ng/μL	2449.0	2302.8	304.3	323.1	1309.7
A260	48.979	46.056	6.085	6.463	26.193
A280	37.355	36.010	4.282	4.698	20.047
260/280	1.31	1.28	1.42	1.38	1.31
260/230	0.73	0.73	1.27	1.06	1.77

Table S2: Nanodrop results from samples (**a**), (**b**), (**c**) and (**d**) and (**e**) as described in figure 2.13

	a	**b**	**c**	**d**
ng/μL	14642.2	799.4	9471.0	19395.7
A260	29.843	15.987	189.420	387.915
A280	37.355	36.010	4.282	4.698
260/280	65.985	10.202	44.95	150.555
260/230	4.35	2.48	4.49	2.03

Table S3: Nanodrop results from samples (**a**), (**b**), (**c**) and (**d**) as described in figure 2.8

	a	**b**	**c**
ng/μL	17804.7	2720.5	317.3
A260	856.095	54.411	6.345
A280	79.811	38.104	1.173
260/280	4.46	1.43	5.41
260/230	4.14	1.66	4.06

Table S4: Nanodrop results from samples (**a**), (**b**) and (**c**) as described in figure 2.12

	0x diluted	**10x diluted**	**100x diluted**	**1000x diluted**
control				
ng/μL	-1.5	-2.5	-0.1	0.0
A260	-0.029	-0.051	-0.002	0.000
A280	-0.037	-0.024	0.005	0.018
A260/A280	0.79	2.07	-0.33	00.00
A260/A230	0.25	2.88	0.07	0.01
a				
ng/μL	9070.3	1833.3	179.7	22.0
A260	181.407	36.665	3.595	0.440
A280	37.129	6.726	0.682	0.097
A260/A280	4.89	5.45	5.27	4.56
A260/A230	4.82	5.20	5.32	4.87
b				
ng/μL	11346.3	2675.3	211.7	22.7
A260	226.925	53.446	4.235	0.455
A280	47.128	10.742	0.805	0.088
A260/A280	4.82	4.98	5.27	5.17
A260/A230	4.70	4.68	5.20	5.41

Table S5: Nanodrop results from dilutions of 2.11a to see if this has an effect on the values measured.Samples (**a**) and (**b**) are both from this figure as well. control was a measurement of the blank to see if this had a major influence on the end measurements.

	a	**b**
ng/μL	456.5	15457.7
A260	9.129	309.155
A280	4.918	66.327
A260/A280	1.86	4.66
A260/A230	3.49	4.43

Table S6: Nanodrop results from dilutions of (**a**) and (**b**) to see if this has an effect on the values measured.(**a**) and (**b**) as shown in figure 2.11b

	a	**b**
ng/μL	13968.2	19549.0
A260	279.363	390.980
A280	59.285	98.934
260/280	4.71	3.96
260/230	4.55	3.72

Table S7: Nanodrop results from samples (**a**), (**b**) and (**c**) as described in figure 2.9

Meta3C laboratory step	**Final Material**	**Preparation**	**Product number**	**Manufacturer**
Formaldehyde fixation	3% Formaldehyde solution	10x dilution of 36% stock (5 ml stock in 50ml final volume pure grade water)	50-00-0	Sigma-Aldrich
Formaldehyde fixation quenching	2.5M Glycine	93.84 gr Glycine in 500 mL final volume pure grade water	G8898	Sigma-Aldrich
sample preparation	1 x TE	-	V6232	Promega
Iterative bead beating lysis	Beads	-	KT03961-1-004.2	Precellys
Chemical lysis	10% SDS	2x dilution of 20% SDS stock	AM9820	Thermo Scientific
DNA digestion	1000 U HpaII	solved in final volume 4020μL digestion mixture	R0171M	New England BioLabs
	CutSmart digestion buffer	xxx amount in digestion mixture	B7204S	New England BioLabs
	10% Triton X-100 for molecular biology	-	X100	Sigma-Aldrich
Sample dilution	14mL pure grade water	added to ligation mixture in next step	-	-
DNA proximity ligation	1.6mL 1M NaOH	4gr stock in 100mL pure grade water	106498	Merck
	700μL of Adenosine 5'-triphosphate disodium salt hydrate (ATP) stock	1gr of 99% stock into 14mL pure grade water, NaOH used to alter pH to 6.13	A26209	Sigma-Aldrich
	250U T4 DNA ligase	xxxx in final ligation mixture	L6030-HC-L	Thermo Scientific

Formaldehyde fixation reversal in 65°C overnight	Proteinase K	amount 20mg/mL	BIO-37084	Eurobio
	0.5M EDTA	2.8gr in 15mL pure grade water brought to pH 7 with NaOH and HCl	E5134-500G	Merck
Precipitation 1	3M sodium acetate and 2-propanol	1.6mL 3M sodium acetate and 16mL 2-propanol	229873-10G, I9516	Sigma-Aldrich
	3M sodium acetate	204.12gr in 400mL pure grade water	229873-10G	Sigma-Aldrich
Phenol Chloroform DNA extraction	Phenol Chloroform	900μL pure Phenol-Chloroform per reaction	873359	Interchim
Precipitation 2	3M sodium acetate	(see precipitation 1)	-	-
	100% Ethanol	-	51976	Sigma-Aldrich
Protein digestion and RNA digestion	1M Trizma hydro chloride (Tris) buffer	78gr in 500mL pure grade water final volume, pH 7.5 (adjusted with NaOH and HCl)	33742-500G	Sigma-Aldrich
	5%RNAse	20μL RNAse in 400μL Tris buffer	RA50001500	Geneaid

Table S8: An overview of all reagents used in the Met3C laboratory protocol executions as described in this report. Please note that this is based on the original protocol as published by Marbouty et al. 2014 and 2017 [1] [24], alterations done in the present work are indicated in table 2.1 in this report.

Meta3C laboratory step	**Equipment**	**Manufacturer**
Formaldehyde fixation	cooling centrifuge (5mL eppendorf tube rotor)	Eppendorf
	Hei-MIX Multi Reax test tube shaker	Heidolph
Formaldehyde fixation quenching	slow rotation on roller mixer	STUART SRT9
Sample preparation	-	-
Iterative bead beating lysis and chemical lysis	TissueLyserII (shaker)	Qiagen
DNA digestion	38°C 200rpm incubator	Thermo Scientific
Sample dilution	-	-
DNA proximity ligation	water bath on 16°C in 4°C climate chamber	-
Formaldehyde fixation reversal	65°C stove	-
Precipitation 1	-80 freezer and centrifuge with rotor for 50mL tubes	Eppendorf
Phenol Chloroform DNA extraction	cooling centrifuge	Eppedorf
Precipitation 2	-80 or -20 freezer and cooling centrifuge	Eppendorf
Protein digestion and RNA digestion	heat block	Thermolyne type 16500 Dri-Bath

Table S9: An overview of all equipment used in the Met3C laboratory protocol executions as described in this report. Please note that this is based on the original protocol as published by Marbouty et al. 2014 and 2017 [1] [24], alterations done in the present work are indicated in table 2.1 in this report.

Listing 6.1: bash version

```
#!/bin/bash
set −e
set −u
set −o pipefail

echo ”IMPORTANT: ANALAYSIS IS TIME CONSUMING. \
   RUN WITHIN SCREEN TO PREVENT UNWANTED EARLY ABORTION OF PROGRAMME”
echo ”IMPORTANT: BUG FIXING TIPS INCLUDED IN SCRIPTFILE AS CODE COMMENTS”

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−INSTALLATIONS−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
HOME=”/media/vol3/giannis/jolien/pm/meta3C_script_jolien”

mkdir −p ./toolkits/ #−p tag: only make new if directory does not exist yet.

cd ./toolkits #for ease this script moves directory of executions

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−SRA
      TOOLKIT−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
# Intallation of ubunut linux 64bit architecture is used here.
# If run on a different machine please make sure to search NCBI SRA toolkit
# download online and download the toolkit according to your machine.

echo ”NOTIFICATION: checking neccesary packages”
echo ”NOTIFICATION: downloading SRA Toolkit version 2.9.6−1−ubuntu64”

wget −−output−document sratoolkit.tar.gz\
 http://ftp−trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current−ubuntu64.tar.gz
tar −vxzf sratoolkit.tar.gz #unpacking the sra toolkit file


SRA_BIN_PATH=${HOME}/toolkits/sratoolkit.2.9.6−1−ubuntu64/bin/

# A test to make sure the toolkit is installed as wanted by checking if fastq−
# dump can be found and is an executable
if test −x ${SRA_BIN_PATH}/fastq−dump
then
    echo ”CONFIRMATION: SRA toolkit installed successfully”
else
    echo ”NOTIFICATION: INSTALLATION NOT SUCCESSFULL. \
        Possible solution: go to analysis.sh and check\
        if you are using the right machine for our SRA file.”
    break
fi

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−IDBA−UD−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
# IDBA−UD is not used by metaTOR, but neede for making an initial metagenomic
# assembly.

echo ”NOTIFCATION: installing IDBA_UD”

git clone ”https://github.com/loneknightpy/idba” ./idba_ud
cd idba_ud/ # Current location is now $HOME/toolkits/idba_ud
bash build.sh #execute the installation

export PATH=$PATH:${HOME}/toolkits/idba_ud/bin
export PATH=${PATH}:/home/ronnie/bin/

echo ”NOTIFICATION: IDBA installed succesfully”

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−MetaTOR−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
# This installation is locally and we run it from bash script.
# Full installtion possible as well, requires some alterations in the script
# This is so I was able to alter some files of the pipeline to circumvent
# errors that required that.

echo ”NOTIFICATION: Downloading and installing metaTOR locally.”
```

```
cd ${HOME}/toolkits
mkdir −p ./metator_repository
git clone ”https://github.com/koszullab/metaTOR” ./metator_repository

METATOR_EXECUTABLE=”${HOME}/toolkits/metator_repository/metator/bin/metator.sh”

echo ”NOTIFACTION: running metaTOR dependencies to install final \
    packages needed. ”

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−MetaTOR required packages−−−−−−−−−−−−−−−−
# The metaTOR pipeline is dependent on a few other libraries and this command
# will make sure they are all available in the environment
# bash $METATOR_EXECUTABLE dependencies
# Additionally, make sure dependencies like python, R and bowtie2 are
# installed.
# These are not checked by the metaTOR pipeline:
# sudo apt install bowtie2 samtools hmmer prodigal

echo ”NOTIFICATION: if error below: Please run this script as root then run \
metator deploy from root, if not neccesary: skip command. ”
#bash $METATOR_EXECUTABLE deploy
bash $METATOR_EXECUTABLE dependencies

echo ”NOTIFICATION: Metator deploy ran successfully”

# Run ”bash $METATOR_EXECUTABLE version” to determine your intalled version...
METATOR_VERSION=” 0.1.7 ”

# Below: a fix for an error where the config.sh files from metator were put in
# the $HOME directory by the version command.
mv config.sh ${HOME}/toolkits/metator_repository/metator/bin
mv config_current.sh ${HOME}/toolkits/metator_repository/metator/bin

bash $METATOR_EXECUTABLE version

echo ”NOTIFCATION: the version used in J. Rietkerk’s report was: $METATOR_VERSION ”

#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−INSTALLATIONS
     END−−−−−−−−−−−−−−−−−−−−−−−−−−−−

echo ”NOTIFICATION: ALL PACKAGES NEEDED FOR METAGENOMIC CHROMOSOME \
CONFORMATION CAPTURE ANALYSIS ARE INSTALLED”


#−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−Analysis
     steps−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−−
# The analysis is made up of a initiating steps before metaTOR is called
# STEP 1 and 2: downloading the SRA file and creating the prelimiary
# metagenomic assembly with IDBA_UD.
# After this it is the 4 steps of the metaTOR pipeline.

#−−−−−−−−−−−−−STEP 1−−−−−−−−−−−−−−−−−−Downloading data from SRA database−−−−−−−−−−−
# For the report SRAtoolkit version 2.9.2 ubuntu64 was used.
# Experment SRX1166811 and reads SRR21190405 from the NCBI database were used
# for the first analysis tryout.
# Prefetch and fastq−dump are SRA toolkit functinos. Prefetch will download the
# .sra file if this was not previously downloaded. Fastq−dump will take this
# .sra file and get the forward and reverse fastq reads in separate files.

# IMPORTANT: to make sure the SRA download location is correct to your machine
# run ”∗/toolkits/sratoolkitx−/bin/vdb−config −i” and use this graphical
# interface to change the location used.

echo ”NOTIFICATION: Downloading SRA fastq files”
echo ”NOTIFICATION: See config_analysis.sh file for circumventing errors here”

SRA_OUTPUT_PATH=”/media/vol3/giannis/jolien/pm/meta3C_script_jolien/ncbi/public/sra”
SRA_IDENTIFIER=”SRR2190405”
```

```
prefetch ${SRA_IDENTIFIER}

cd $SRA_OUTPUT_PATH
fastq−dump −−split−files ${SRA_IDENTIFIER}.sra

# note to self: fastq−dump works but takes LONG as heck.
# will use the data I downloaded before.
# When publishing this thing in my report make sure to fixt the fastq
# forward and reverse locations! Also memory was exhaussted because
# I downloaded the second set..

FASTQ_FORWARD=”${SRA_OUTPUT_PATH}/${SRA_IDENTIFIER}_1.fastq”
FASTQ_REVERSE=”${SRA_OUTPUT_PATH}/${SRA_IDENTIFIER}_2.fastq”

echo ”SRA path ${SRA_BIN_PATH}”
echo ”SRA output path: ${SRA_OUTPUT_PATH}”
echo ” Identifier sra ${SRA_IDENTIFIER}”
echo ”Fastq forward filename ${FASTQ_FORWARD}”
echo ”Fastq reverse filename ${FASTQ_REVERSE}”


#−−−−−−−−−−−−−−−−STEP 2−−−−−−−−−−−−−−−−−−−−−−Initial assembly−−−−−−−−−−−−−−−−−−−−
# IDBA_UD version used in making of my report:
# Here we merge the previous made fastq read files into a single file ,
# because we want them to be assembled without being related to eachother
# As this might make the assembly more fragmented.
# After that we make the assembly with idba. I did not make a specific
# executable variable for this because I added the bin/ to our path above.
# If this errors, perhaps look for an IDBA installation outside of this script.

echo ”NOTIFICATION: Merging fastq files into a single fasta file \
for assembly with IDBA−UD”

FASTA_MERGED=”${SRA_IDENTIFIER}_12.fasta”
fq2fa −−merge $FASTQ_FORWARD $FASTQ_REVERSE ${FASTA_MERGED}

IDBA_UD_OUTPUT=”${SRA_OUTPUT_PATH}/idba_ud_out”
mkdir −p $IDBA_UD_OUTPUT

echo ”NOTIFICATION: Making initial assembly with IDBA_UD”

idba −r ${FASTA_MERGED} −−num_threads 20 −o $IDBA_UD_OUTPUT
INITIAL_ASSEMBLY_LOCATION=”${IDBA_UD_OUTPUT}”

echo ”NOTIFICATION: Initial assembly made”

#−−−−−−−−−−−−−−−−STEP 3−−−−−−−−−−−−−−−−−−−−−−Align reads to assembly−−−−−−−−−−−−−
# See report maintext for full explanation on individual steps of the
# MetaTOR pipeline.

# To circumvent an error in the path of the initial assembly
# I navigate to its location first , execute the command and
# navigate back to where we started.

cd ./idba_ud_out
mkdir −p ./metator_out


# Additionally, an error in the creation of an output file occured.
# I think the metator_out has to be in the running location as well..
# so we make a directory for that here and will put it there..
# running location during project was:
# /media/vol3/giannis/jolien/pm/ReplMarb2014/ncbi/public/sra/idba_ud_out/
# metator_out_testscriptrun

# Now run the command..
bash ${METATOR_EXECUTABLE} align −1 ${FASTQ_FORWARD} −2 ${FASTQ_REVERSE} −a \
./contig.fa -c 10000 -C 10000 -Q 20 --clean-up -p ./metator_out
#note that this command has to be run from the location of contig.fa


#----------------STEP 4----------------------Louvain iterations------------------
# See report maintext for full explanation on individual steps of the
# MetaTOR pipeline.

# We keep in the same execution location as before to prevent similar
# errors and to make it easier for the pipeline to locate the needed files ..


bash ${METATOR_EXECUTABLE} partition --iterations 100

# arguement works but got an error that it can't locate certain files
# probably because of the apart run instead of sequential run...

#----------------STEP
    5----------------------Binning-----------------------------
# See report maintext for full explanation on individual steps of the
# MetaTOR pipeline.

# match annotations to bins, extract bin genomes and subnetowrks,
# build bin-local and global contact maps.


bash ${METATOR_EXECUTABLE} binning --n-bins 100

# This bins the 100 biggest cores only. You may bin all of them, though
# that's not recommended as many of them are singleton sequences that
# aren't very informative.
# inspect the output/partition folder and look at the figures. If you've
# used our publicly available data sets, you should not ethe visible
# bump in virus orthologous groups (VOGs) among smaller bins.


#----------------STEP 6----------------------Annotating the bins-----------------
# See report maintext for full explanation on individual steps of the
# MetaTOR pipeline.

bash ${METATOR_EXECUTABLE} annotation



#--------------STEP 7-----------------------Additional figure creation----------
# Analysis after running as suggested by the MetaTOR github tutorial.
# This creates a CheckM library to analyse and create figures from. I used the
# output tables in excel to create figures.

export fasta_dir="${HOME}/partition/fasta_merged/iteration300"
export checkm_dir="checkm_validation"

mkdir -p $checkm_dir

checkm tree -x fa $fasta_dir $checkm_dir # Add -t 8 for multithreading etc.
checkm tree_qa $checkm_dir -o 2 -f $checkm_dir/checkM_results.txt
checkm lineage_set $checkm_dir $checkm_dir/checkM_output_marker.txt
checkm analyze -x fa $checkm_dir/checkM_output_marker.txt $fasta_dir $checkm_dir
checkm qa -t 8 $checkm_dir/checkM_output_marker.txt ${checkm_dir} -o 2\
> $checkm_dir/checkM_results_complete.txt &


#--------------------OUTPUT----------------------------------------------------
# for a full description of the output created by this script, users are referred to
# the main text of the report to which this script is supplemental.
```