„The genetics behind the evolution of a community of four bacterial species“

Bachelor thesis at the Biology Department

Science Faculty

University of Bern

Written by

Sabrina Michaela Rasch

April 25, 2022

Supervisor: Dr A. Kapopoulou, Institute of Ecology and Evolution, Bioinformatician Theoretical Ecology and Evolution Lab

Prof. Dr C. Bank, Institute of Ecology and Evolution, Head of the Division Theoretical Ecology and Evolution

Abstract

Microorganisms are not only a part of our daily lives but are also widely used in industry. For example, the co-cultured community of *Agrobacterium tumefaciens*, *Comamonas testosteroni*, *Microbacterium saperdae* and *Ochrobactrum anthropi* is very efficient and reliable for degrading/detoxifying substances used for cutting or grinding metals, such as metal working fluid (MWF). In this study, I consider the selective pressure on single-nucleotide polymorphisms (SNPs) as I investigate the genetics of this community. For this purpose, the microbes were allowed to grow in different combinations in the MWF for 44 weeks, and the genomes of the community were regularly sequenced for SNPs. 41 SNPs became fixed, meaning they had a frequency above 0.98 in the population. Of these fixed SNPs 17 are intergenic, and the remaining 24 belong to 16 different genes. The function of these 16 genes and their relevance in the degradation/detoxification of MWF are discussed. Furthermore, the SNPs found at the same position as a fixed SNP in another replicate of the experiment were examined. The community where the additional SNPs occurred, whether they were non-synonymous or synonymous, and a line graph showing SNP frequency at each time point of sequencing were investigated. Finally, the selection pressure exerted on the mutations was investigated by examining the ratio of non-synonymous to synonymous mutations.

1 Introduction

In ecology and evolutionary research, an important question is how microbial communities interact and evolve. In particular, studies focus on the negative and positive effects of the interactions between different species that make up these communities (De Roy et al., 2014; Mitri & Richard Foster, 2013; Widder et al., 2016). Positive, or facilitative, interactions are defined as those where at least one side benefits and the other side is not harmed (Friedman & Gore, 2017; Stachowicz, 2001). The environment in which the community is located is likely to determine the nature of the interaction (Chamberlain et al., 2014). The stress gradient hypothesis (SGH), proposed by Bertness and Callaway (1994), states that more positive interactions occur in a stressful, unfavourable environment.

Microorganisms are an essential part of our daily lives. They help break down food in our gut, cause disease or help treat them (Widder et al., 2016). Even in industrial waste processing, microbes play an important role (Van der Gast et al., 2002; Van Der Gast et al., 2004). Previously used disposal methods, such as evaporation, landfill or incineration, are either uneconomical, inefficient or ecologically unacceptable. Therefore, interest in biological degradation methods is increasing (Van Der Gast et al., 2004). One substance disposed of by such methods is metal working fluid (MWF), used for cutting or grinding metals. These highly alkaline and polluting fluids are very chemically complex. They contain both biocides that inhibit the activity of microorganisms and chemical compounds that are a rich source of nutrients (Piccardi et al., 2019). A microbial community has been found that efficiently and reliably degrades the toxic substances in MWF. This community consists of *Agrobacterium tumefaciens*, *Comamonas testosteroni*, *Microbacterium saperdae* and *Ochrobactrum anthropi* (Van Der Gast & Thompson, 2005, 2014; Van Der Gast et al., 2004).

Piccardi et al. (2019) studied the nature of the interactions in this community. They grew the community in different stressful environments and mainly found positive and facilitative interactions in the more stressful environments, supporting the SGH. However, the process of MWF degradation and the molecular background of the interactions is not yet fully understood. This makes it difficult to make predictions about evolutionary stability or how MWF degradation can be made more efficient. In the present study, the genetic background of the community is investigated. The microorganisms were grown in different combinations in MWF over 44 weeks, and the genomes of the community were sequenced regularly for single nucleotide polymorphisms (SNPs).

The obtained data was then analysed for SNPs that were fixed in the population. These SNPs significantly influence how the population adapts to different environments. The first question of this study is **(1) Which genes contain fixed mutations, and what is the function of the protein they code for?** In addition, it will be investigated whether the microbes have adapted more to each other or the environment. The second question of this study is **(2) Do we see the same mutations in the mono- and co-culture replicates?** Finally, I will also investigate which mutations are harmful and which mutations are more beneficial. The third question of this study is **(3) Which genes are under positive/negative selection?**

2 Methods

This study consists of two parts: The experiment and the analysis of the obtained data. The four microbes studied are *Agrobacterium tumefaciens*, *Comamonas testosteroni*, *Microbacterium saperdae*, and *Ochrobactrum anthropi*.

All calculations and graphs were done with R-Studio (Version 4.1.1, R Core Team, 2021). The packages *data.table* (Version 1.14.2, Dowle and Srinivasan, 2021), *tidyverse* (Version 1.3.1, Wickham et al., 2019) and *latex2exp* (Version 0.5.0, Meschiari, 2021) were used. The annotated code ("Code.Rmd") and the original data sets (subfolder: "/Data/Original") can be found in the project folder. In addition, some data sets were edited in a text editor in order to be able to manipulate them when imported into R-Studio. The project folder is stored at:
https://github.com/sabrinarasch/Microbial_Communities.git.

2.1 Experiment

The experiment was conducted in the lab of Prof. Dr Sara Mitri at the University of Lausanne and is based on the paper of Piccardi et al. (2019).

First, the four microbes were isolated from waste metal working fluid (MWF) based on their ability to survive and grow in it (Van Der Gast & Thompson, 2014). However, only *A. tumefaciens* and *C. testosteroni* survive as monocultures (Piccardi et al., 2019). The community was shown to degrade different MWF substrates efficiently and reliably (Van Der Gast & Thompson, 2005, 2014; Van Der Gast et al., 2004).

Before the experiment, all four microbes were grown on tryptic soy broth (TSB). There were four experimental set-ups, here referred to as microcosms, each of which was run in five replicates.

Microcosms 1 and 2 were monocultures of *A. tumefaciens* and *C. testosteroni*, respectively. In microcosm 1, *A. tumefaciens* died out in three of the five replicates. Therefore, they were not further investigated. Microcosm 3 was a co-culture of *A. tumefaciens*, *C. testosteroni* and *M. saperdae*. Finally, microcosm 4 was a co-culture of all four microbes (*A. tumefaciens*, *C. testosteroni*, *M. saperdae* and *O. anthropi*). In this microcosm, *M. saperdae* died out in two of the replicates (Table 1). For the experiment, 1% of the solution was added to a new bottle of fresh MWF each week for all 17 replicates, and colony forming units per millilitre (CFU/ml) were measured. This was done over 44 weeks, corresponding to about 600 generations (Figure 1). At the beginning of the experiment and on weeks 11, 22, 33 and 44, the genomes of the community were sequenced for single-nucleotide polymorphisms (SNPs) using Illumina.

The obtained data can be found under the name *"allSNPs_w_gene_position.tsv"*. It includes:

- "Cosme.Rep": In which microcosm and replicate the SNP was found
- "Micr.Genom": In which microbe and chromosome the SNP occurred
- "Place": The position of the SNP
- "Alt": The alternative nucleotide
- "Gene": Whether the SNP is in a gene (with a tag for the gene) or intergenic
- "NS/S": Whether the SNP is non-synonymous or synonymous
- "t0–t4": The frequency of the SNP at all time points of sequencing

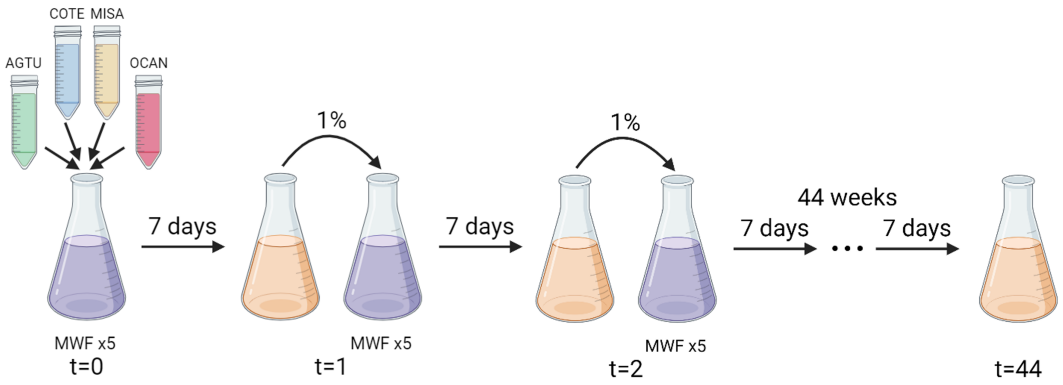

Figure 1: **Sketch of the Experiment.** The experiment was carried out in the Lab of Prof. Dr Sara Mitri at the University of Lausanne and is based on the paper of Piccardi et al. (2019). AGTU: *Agrobacterium tumefaciens*; COTE: *Comamonas testosteroni*; MISA: *Microbacterium saperdae*; OCAN: *Ochrobactrum anthropi*; MWF: Metal working fluid.

Microcosm	Organism(s)	No. Populations
1	*A. tumefaciens*	2 (3 went extinct)
2	*C. testosteroni*	5
3	*A. tumefaciens*, *C. testosteroni*, *M. saperdae*	5
4	*A. tumefaciens*, *C. testosteroni*, *M. saperdae*, *O. anthropi*	5 (in 2 *M. saperdae* went extinct)

Table 1: **Microcosms.** Four experimental set-ups were carried out, referred to as Microcosms. Microcosm 1 and 2 are monocultures, and microcosm 3 and 4 are co-cultures.

2.2 SNP localisation and Protein function

To answer the questions of which genes have fixed mutations and what the function of these genes are, a list of fixed SNPs was provided by the lab of Sara Mitri. The definition of a fixed SNP is when the frequency in the population is more than 0.98. This can be the case at the last sequencing time point, t = 44, or earlier. If it is earlier, the frequency must not drop below 0.98 again; only then has a SNP been defined as fixed. I carried out two steps with this data. First, I examined whether each SNP is located in a gene and, if so, in which one. Secondly, I looked up the function of the gene. The original file of the fixed SNPs can be found under the name *"FixedMutations_OverTime.txt"*. It includes:

- "Microbe.Replicon": In which microbe and chromosome the SNP was found
- "Microcosme.Replicate": In which microcosm and replicate the SNP occurred
- "Position": The position of the SNP
- "Ref" and "Alt": The reference and alternative nucleotide
- "Number": A number I assigned to the SNP for identification

SNP localisation. For the first step, I have used three data sets that consist of the whole genome data of the four microbes.
One consists of information about all genes in the microbes. It can be found under the name *"alltogether_modified.lst.txt"* and includes, among other things:

- "Start" and "End": The start and end nucleotide of a gene
- "Gene": The name of the gene
- "Organism": To which microbe and chromosome the gene belongs to

- "UniProtID": An UniProt ID
- "Protein": The name of the protein for which this gene codes

The other two data sets consist of the sequence data of the genes. They can be found under the names *"alltogether_modified.gen.fasta.gen"* and *"alltogether_modified.prt"* and include, among other things:

- "alltogether_modified.gen.fasta.gen"
  - "DNA": The nucleotide sequence of the gene
- "alltogether_modified.prt"
  - "AA": The amino acid sequence of the protein
- Both
  - "Start" and "End": The start and end nucleotide of a gene
  - "Gene": The name of the gene
  - "Organism": To which microbe and chromosome the gene belongs to
  - "UniProtID": A UniProt ID
  - "Protein": The protein name for which this gene codes

To find out whether a SNP was in a gene, I compared the position of the fixed SNPs with the coordinates of the genes. If a SNP was in a gene, it was added to a list, including:

- "Organism": In which microbe and chromosome the SNP is
- "Replicate": In which microcosm and replicate the SNP occurred
- "Number": The assigned number of the SNP
- "Position": The position of the SNP
- "Ref" and "Alt": The reference and alternative nucleotide
- "Gene": The name of the gene
- "Start" and "End": The start and end nucleotide of a gene
- "Protein": The protein name for which this gene codes.
- "UniProtID": A UniProt ID

Afterwards, I used the data sets with the nucleotide and amino acid sequences to find the sequences corresponding to the genes I found. For this I used the gene name.

Protein function. In the second step, I used the UniProt ID to search for the protein functions in the online data bank UniProt (The UniProt Consortium, 2018). In the entries of this data bank, I could find a QuickGo ID, where the function was further explained (Binns et al., 2009). Additionally, there was an external link leading to the InterPro data bank (Blum et al., 2021). This data bank contains information about the protein families. Genes without a UniProt ID (hypothetical genes) were further analysed using the programme BLAST. This programme is used to find similar sequences to the ones entered into the programme in different species. The nucleotide and amino acid sequences of the hypothetical genes were entered into the BLAST programme of the UniProt data bank (The UniProt Consortium, 2018).

2.3 Mono- vs co-culture

The second question of this study is whether the same mutations were found in mono and co-cultures. It has been shown that *M. saperdae* and *O. anthropi* cannot survive as monocultures in MWF (Piccardi et al., 2019). Therefore, only *A. tumefaciens* and *C. testosteroni* were grown as monocultures, and therefore only these data were used for this investigation.

Since the function of the proteins of the fixed SNPs was studied in this paper, only these SNPs were used. As already mentioned, a SNP is fixed if the frequency in the population is more than 0.98 and does not decrease again. These fixed SNPs occurred in certain microcosms/replicates. Further, some SNPs with intermediate frequency occurred in the same position as a fixed SNP in other microcosms/replicates. I looked at these additional SNPs and examined in which microcosm and replicate they occurred, whether they were non-synonymous or synonymous. Furthermore, a line plot was made showing the frequency of the SNPs at each time point of the sequencing (Figures 2 to 15).

2.4 SNPs under selective pressure

To determine which genes are under positive selection and which under negative selection, a dN/dS ratio was calculated. This is the ratio of non-synonymous to synonymous mutations. Since there are only 20 different amino acids but 64 possible codons, some mutations do not change the amino acids; these are synonymous mutations. In a non-synonymous mutation, the nucleotide is

change so that the codon corresponds to another amino acid. The data set of all SNPs was used to estimate the dN/dS. First, all SNPs had to be found located in one of the genes containing at least one fixed SNP. The dN/dS was then calculated by grouping the SNPs of the genes by microcosm. The number of non-synonymous and synonymous SNPs was counted and divided. If only non-synonymous or only synonymous SNPs were present, *NA* (for missing values) was the output for dN/dS.

3 Results

In the following part of the report, the results will be described. The three research questions are: **(1)** Which genes contain fixed mutations, and what is the function of the protein they code for? **(2)** Do we see the same mutations in the mono- and co-culture replicates? **(3)** Which genes are under positive/negative selection?

3.1 SNP localisation and protein function

A provided list of 41 fixed SNPs was examined for the first question. As well as the whole genome data of the four study organisms to find the reference sequence and the function of the genes the fixed SNPs belong to. Of the 41 fixed SNPs, 17 were intergenic, meaning not belonging to a gene. Of the remaining 24, two were assigned to genes coding for hypothetical proteins and 22 to 14 different genes. Whereby only two of the genes have multiple fixed SNPs. The gene nasR of *Comamonas testosteroni* has six fixed SNPs, and the gene yjiA of *Ochrobactrum anthropi* has four. The function of the 14 genes was looked up in the data banks UniProt (The UniProt Consortium, 2018), QuickGo (Binns et al., 2009) and InterPro (Blum et al., 2021).

3.1.1 Hypothetical genes

For the two genes with hypothetical proteins, both the nucleotide and amino-acid sequences were used to find a similar sequence by running a BLAST similarity search on the UniProt website (The UniProt Consortium, 2018).

SNP 6 belongs to the gene AGTU_01326 (Table 3). The similarity search for the nucleotide sequence of this gene resulted in a maximum match of 62.1% (2 matches). Furthermore, the similarity search for the amino acid sequence yielded at most a match of 46.3% (2 matches) (Table 4). The function of this proteins is unknown.

SNP 18 belongs to the gene COTE_04126 (Table 3). The similarity search for the nucleotide sequence of this gene resulted in a maximum match of 85.2% (1 match). Furthermore, the similarity search for the amino acid sequence yielded the same match (Table 4). The function of this protein is unknown.

3.1.2 SNPs of *Agrobacterium tumefaciens*

SNP 1 belongs to the gene ropA_2 (Table 3). The provided UniProt ID (Q05811) leads to the ropA gene of *Rhizobium leguminosarum bv. Viciae*. This gene codes for the outer membrane protein IIIA, which has a porin activity and acts in ion transport (The UniProt Consortium, 2018). It is located in the outer membrane and can transport substances smaller than 1,000 Da. The transporter consists of beta-strands that form a beta-barrel (Binns et al., 2009).

SNP 3 belongs to the gene fadD (Table 3). The provided UniProt ID (P69451) leads to the fadD (oldD) gene of *Escherichia coli*. This gene codes for a long-chain-fatty-acid–CoA ligase, which catalyses the esterification of exogenous long-chain fatty acids. This produces metabolically active CoA thioesters, which can then be degraded or incorporated into phospholipids. In addition, this protein is involved in the aerobic beta-oxidative degradation of fatty acids. This enables the aerobic growth of *E. coli* on fatty acids as a sole carbon and energy source (The UniProt Consortium, 2018).

SNP 4 belongs to the gene AGTU_00650 (Table 3). The provided UniProt ID (B1WVN5) leads to the rfr32 gene of *Crocosphaera subtropica*. This gene codes for the pentapeptide repeat protein Rfr32 (The UniProt Consortium, 2018). The function of the pentapeptide repeat protein family is uncertain. Cyanobacteria have many such proteins, suggesting that they are involved in a metabolism specific to these bacteria, such as nitrogen fixation or photosynthesis. An enzymatic function is unlikely, whereas a structural function could be the case (Bateman et al., 1998).

SNP 5 belongs to the gene yjgN (Table 3). The provided UniProt ID (P39338) leads to the yjgN gene of *Escherichia coli*. This gene codes for the inner membrane protein YjgN (The UniProt Consortium, 2018), which belongs to the family DUF898 of bacterial proteins. The function of most of the proteins in this family is unknown. (Blum et al., 2021).

SNP 7 belongs to the gene secA (Table 3). The provided UniProt ID (P52966) leads to the secA gene of *Rhodobacter capsulatus*. This gene codes for the protein SecA, which is part of the

Sec protein translocase complex (The UniProt Consortium, 2018). The cytosolic chaperone SecB mediates the targeting of the precursor to the membrane-attached SecA. SecA is the peripheral part of the preprotein translocase and acts as an ATPase. The integral membrane proteins SecY/E/G and SecD/F of this translocase are ultimately responsible for the translocation of the preprotein (Helde et al., 1997).

SNP 8 belongs to the gene frmR (Table 3). The provided UniProt ID (P0AAP3) leads to the frmR (yaiN) gene of *Escherichia coli*. This gene codes for the transcriptional repressor FrmR of the frmRAB operon (The UniProt Consortium, 2018). The operon is responsible for perceiving and detoxifying formaldehyde. The FrmR protein is inactivated in the presence of formaldehyde, which allows expression of the formaldehyde detoxification machinery (Denby et al., 2016). It belongs to the protein family of metal-sensitive transcriptional repressors that bind either copper, nickel or cobalt ions and thus repress transcription (Blum et al., 2021).

3.1.3 SNPs of *Comamonas testosteroni*

SNPs 11, 12, 15, 19, 24, and 27 belong to the gene nasR (Table 3). The provided UniProt ID (Q48468) leads to the nasR gene of *Klebsiella oxytoca*. This gene codes for a nitrate regulatory protein that regulates the expression of the nasFEDCBA operon (The UniProt Consortium, 2018). This operon is required for nitrate and nitrite assimilation. The nasFED genes encode a nitrate and nitrite transporter, the nasCA genes encode an assimilatory nitrate reductase, and the nasB gene encodes an assimilatory nitrite reductase. The nasR gene is located upstream of nasF and encodes a positive regulator of its expression that responds to nitrate and nitrite (Wu et al., 1999).

SNP 16 belongs to the gene trkH (Table 3). The provided UniProt ID (E1V6C5) leads to the trkH gene of *Halomonas elongata*. This gene codes for the Trk system potassium uptake protein TrkH, which is a potassium chloride symporter (Binns et al., 2009; The UniProt Consortium, 2018). Together with the regulatory NAD-binding peripheral membrane protein TrkA, TrkH forms the Trk system, which is responsible for the uptake of potassium (Blum et al., 2021).

3.1.4 SNPs of *Microbacterium saperdae*

SNP 30 belongs to the gene MISA_01549 (Table 3). The provided UniProt ID (P9WIA3) leads to the Rv2368c gene of *Mycobacterium tuberculosis*. This gene codes for a PhoH-like protein (The UniProt Consortium, 2018). The PhoH Protein of *Escherichia coli* has an ATP-binding activity

and belongs to the pho regulon. This regulon is activated when phosphate is limited, and the products are involved in the transport and utilisation of various forms of combined phosphates and free phosphate (Kim et al., 1993).

SNP 31 belongs to the gene puuP_3 (Table 3). The provided UniProt ID (P76037) leads to the puuP (ycjJ) gene of *Escherichia coli*. This gene codes for the Putrescine importer PuuP, which helps in the uptake of putrescine. Putrescine is used as an energy source when glucose is not available (The UniProt Consortium, 2018). Putrescine (1,4-diaminobutane) is a polyamine that serves as the precursor of spermidine and spermine. It is formed by decarboxylation of ornithine (Binns et al., 2009).

SNP 32 belongs to the gene ftsH_1 (Table 3). The provided UniProt ID (P9WQN3) leads to the ftsH gene of *Mycobacterium tuberculosis*. This gene codes for the ATP-dependent zinc metalloprotease FtsH (The UniProt Consortium, 2018). It acts as a processive, ATP-dependent zinc metallopeptidase that catalyses the hydrolysis of peptide bonds. Water acts as a nucleophile, metal ions hold the water molecule, and a charged amino acid side chain serves as a ligand for the metal ions (Binns et al., 2009). The main function of this protein is the selective degradation of non-assembled, incompletely assembled and/or damaged membrane-anchored proteins. It is also involved in the processing of pre-proteins, the dislocation of membrane proteins and the degradation of regulatory proteins (Blum et al., 2021).

SNP 33 belongs to the gene dedA_1 (Table 3). The provided UniProt ID (P0ABP6) leads to the dedA gene of *Escherichia coli*. This gene codes for the Protein DedA (The UniProt Consortium, 2018), which belongs to a family of uncharacterised proteins, further including inner membrane proteins YghB and YqjA (Blum et al., 2021).

SNP 34 belongs to the gene srlR (Table 3). The provided UniProt ID (P15082) leads to the srlR (gutR) gene of *Escherichia coli*. This gene codes for a Glucitol operon repressor, which represses the phosphoenolpyruvate-dependent sugar phosphotransferase system. This system is responsible for the uptake and phosphorylation of specific carbohydrates from the extracellular environment (Binns et al., 2009).

3.1.5 SNPs of *Ochrobactrum anthropi*

SNPs 38, 39, 40, and 41 belongs to the gene yjiA (Table 3). The provided UniProt ID (P24203) leads to the yjiA gene of *Escherichia coli*. This gene codes for the P-loop guanosine triphosphatase YjiA, which has a GTP-dependent regulatory function (The UniProt Consortium, 2018). It can be induced in response to DNA damage but is not an essential gene. Nevertheless, its function appears to be important for cell survival (Khil et al., 2004).

3.2 Mono- vs co-culture

Piccardi et al. (2019) showed that only *Agrobacterium tumefaciens* and *Comamonas testosteroni* can survive as monocultures in MWF. Because of this, only the SNPs of these two microbes were used to compare whether the same mutations occure in mono- and co-cultures. Four different cultures were grown, referred to as microcosms. Microcosm 1 is the monoculture of *A. tumefaciens*, and microcosm 2 is the monoculture of *C. testosteroni*. Microcosm 3 is a co-culture of *A. tumefaciens*, *C. testosteroni* and *Microbacterium saperdae* and in microcosm 4, *Ochrobactrum anthropi* was introduced into the co-culture of *A. tumefaciens*, *C. testosteroni* and *M. saperdae* (Table 1). I only compared the SNPs in the same position as a fixed SNP, because the function of the protein of the fixed SNPs was also studied in this paper. Therefore 15 SNPs were looked at namely SNP 1, 3, 4, 5, 6, 7 and 8 of *A. tumefaciens* and SNP 11, 12, 15, 16, 18, 19, 24 and 27 of *C. testosteroni*.

For seven SNPs, no additional SNP was found in the same position. Of these seven SNPs, five are in *A. tumefaciens*. These five are SNP 4 in gene AGTU_00650 (rfr32) (Position 677536; microcosm.replicate 3.2; Figure 4; Table S3), SNP 5 in gene yjgN (Position 2863371; microcosm.replicate 3.4; Figure 5; Table S4), SNP 6 in gene AGTU_01326 (Position, 1372439; microcosm.replicate 4.5; Figure 6; Table S5), SNP 7 in gene secA (Position 555455; microcosm.replicate 4.1; Figure 7; Table S6), and SNP 8 in gene frmR (yaiN) (Position 181352; microcosm.replicate 4.4; Figure 8; Table S7). All these SNPs occurred in the co-cultures. SNP 4 and 5 are in microcosm 3, and SNP 6, 7 and 8 are in microcosm 4. All are non-synonymous except SNP 6. The remaining two SNPs are in *C. testosteroni*. These are SNP 16 in gene trkH (Position 5643748; microcosm.replicate 2.3; Figure 11; Table S10) and SNP 18 in gene COTE_04126 (Position 4436504; microcosm.replicate 2.5; Figure 12; Table S11). Both occurred in the monoculture (microcosm 2) and are non-synonymous.

For two SNPs, one additional SNP was found in the same position. One of these is SNP 1 in *A. tumefaciens* in the gene ropA_2 (Position 703340; microcosm.replicate 1.1; Figure 2; Table S1). The fixed SNP occurred in the monoculture, and the additional SNP occurred in the co-culture of microcosm 3. Both are non-synonymous. The other is SNP 19 in *C. testosteroni* in the gene nasR (Position 5466537; microcosm 2.5; Figure 13; Table S12). The fixed SNP occurred in the monoculture, and the additional SNP occurred in the co-culture of microcosm 4. Both are non-synonymous.

For two SNPs, two SNPs each were found at the same position. Both are in *C. testosteroni* in the gene nasR and occurred in the co-culture of microcosm 3. One is the SNP 24 (Position 5466905; microcosm.replicate 3.3; Figure 14; Table S13) which occurred in the co-culture of microcosm 3, as did one of the additional ones. The third SNP in this position occurred in the co-culture of microcosm 4. The other fixed is SNP 27 (Position 5466552; microcosm.replicate 3.4; Figure 15; Table S14) which occurred in the co-culture of microcosm 3. Both SNPs in the same position as this SNP are found in the co-culture of microcosm 4. All those SNPs are non-synonymous.

There are two fixed SNPs in *C. testosteroni* that are already in the same position. These are SNP 12 and 15 (Position 5466897; microcosm.replicate 2.2 & 2.3; Figure 10; Table S9) in the gene nasR. Corresponding to this position, there are four more SNPs. Both fixed SNPs occurred in replicates of the monoculture, as did two of the additional SNPs. The remaining two occurred each in one of the co-cultures. All of them are synonymous.

There is one SNP in *C. testosteroni*, where six SNPs occurred in the same position. This is SNP 11 in gene nasR (Position 5466695; microcosm.replicate 2.1; Figure 9; Table S8). The fixed SNP occurred in the monoculture, as did one of the additional ones. Furthermore, one additional SNP occurred in the co-culture of microcosm 3, and the remaining four occurred in the co-culture of microcosm 4. In microcosm 4 replicate 5, there are two SNPs, one of which is synonymous. All other SNPs are non-synonymous.

There is one SNP in *A. tumefaciens*, where eight SNPs occurred in the same position. This is SNP 3 in gene fadD (Position 83564; microcosm.replicate 1.2; Figure 3; Table S2). The fixed SNP occurred in the monoculture, as did one of the additional ones. Furthermore, five additional SNP occurred in the co-culture of microcosm 3, and the remaining two occurred in the co-culture of microcosm 4. All of them are non-synonymous.

3.3 SNPs under selective pressure

For the third question, the dN/dS ratio was calculated by grouping the SNPs of all replicates of a microcosm together. For all 16 genes, there are 24 fixed SNPs and 113 not fixed SNPs. 72 of the not fixed SNPs are non-synonymous (NS), and 41 of the not fixed SNPs are synonymous (S). Of the fixed SNPs, only three are S (SNP 6, 12, 15) the rest is NS. All results can be found in Table 5.

For nine of the genes, no dN/dS could be calculated in any microcosms because only NS or S SNPs were present.

For gene ropA_2, one NS SNP in microcosm 1 and one NS SNP in microcosm 3 were found.

For gene fadD, two NS SNPs in microcosm 1 and 4 and five NS SNPs in microcosm 3 were found.

For gene AGTU_00650, one NS SNP in microcosm 3 and two NS SNPs in microcosm 4 were found.

For gene secA, one NS SNP was found in microcosm 4.

For gene COTE_04126, one NS SNP in microcosm 2 and one S SNP in microcosm 4 were found.

For gene MISA_01549, one NS SNP was found in microcosm 3.

For gene ftsH_1, two NS SNPs in microcosm 3 and four NS SNPs in microcosm 4 were found.

For gene dedA_1, two S SNPs in microcosm 3 and three NS SNPs in microcosm 4 were found.

For gene yjiA, five NS SNPs in microcosm 4 were found.

For five of the seven remaining genes, a dN/dS could be calculated. However, in some microcosms still, only NS or S SNPs occurred.

For gene yjgN, one S SNP in microcosm 1 and five S SNPs in microcosm 4 were found. In microcosm 3, one NS SNP and four S SNPs were found, resulting in a dN/dS of 0.25.

For gene frmR, three S SNPs in microcosm 1 and two S SNPs in microcosm 3 were found. In microcosm 4, four NS SNP and two S SNPs were found, resulting in a dN/dS of 2.

For gene trkH, two NS SNPs in microcosm 4 were found. In microcosm 2 one NS SNP and one S SNP were found, resulting in a dN/dS of 1.

For gene puuP_3, two NS SNP in microcosm 4 were found. In microcosm 3 ten NS SNPs and eight S SNPs were found, resulting in a dN/dS of 1.25.

For gene srlR, two NS SNP in microcosm 4 were found. In microcosm 3 one NS SNP and one S SNP were found, resulting in a dN/dS of 1.

Finally, a dN/dS could be calculated in two genes in all microcosms.
For gene AGTU_01326 in microcosm 4 five NS SNPs and two S SNPs were found, resulting in a dN/dS of 2.5.
For gene nasR in microcosm 2 four NS SNPs and four S SNPs were found, resulting in a dN/dS of 1. In microcosm 3 13 NS SNPs and four S SNPs were found, resulting in a dN/dS of 3.25. Furthermore, in microcosm 4 17 NS SNPs and four S SNPs were found, resulting in a dN/dS of 4.25.

4 Discussion

This study examined three questions regarding the genetics of a microbial community that can degrade/detoxify MWF. First, it was asked which genes contain a fixed mutation and what the function of the protein is they encode for. Second, a comparison between the mutations occurring in mono- and co-cultures was made. Lastly, whether a gene is under positive or negative selection was examined.

When we look at all SNPs that occur at the same position as a fixed one, my results show that more SNPs were found in the co-cultures than in the monocultures. However, for *Agrobacterium tumefaciens* we compare two replicates of monocultures with ten replicates of co-cultures, and for *Comamonas testosteroni* we compare five replicates of monocultures with ten replicates of co-cultures. Therefore, there should be some normalisation to make the mono- and co-cultures comparable, which was not part of this analysis. Despite that, when comparing the co-cultures individually with the monoculture, we still see that in *A. tumefaciens* in the monoculture, the least SNPs occurred. Between the co-cultures, most SNPs occur in the co-culture without *Ochrobactrum anthropi*. In *Comamonas testosteroni*, the same number of SNPs occurred in the monoculture and the co-culture of all microbes, and the least in the co-culture without *O. anthropi*. Furthermore, it turned out that overall, more non-synonymous SNPs occurred than synonymous SNPs. One reason for this could be that I have not considered the rate of transitions (T $\leftrightarrow$ C and A $\leftrightarrow$ G) and transversions (T, C $\leftrightarrow$ A, G). At the third (wobble) position in a codon, transitions are more likely to be synonymous than transversions, which leads to an underestimation of synonymous and an overestimation of non-synonymous mutations (Yang & Bielawski, 2000). Another obstacle when interpreting the results is that not all microbes can be grown as monocultures, and some combinations of co-cultures were not performed.

In this study, I calculated the dN/dS ratio by simply counting the number of non-synonymous and synonymous mutations and dividing these two numbers. This is not a very accurate way to do this calculation. Only about 25% of all possible mutations are indeed synonymous, which should be corrected for (Hurst, 2002). However, this has been omitted because I do not have the quantity of data available for these corrections. I hope these findings might indicate a trend, but I recognise that more data is necessary for conclusive analysis. Furthermore, several substitutions may have occurred at one position, especially if a long time has elapsed since the beginning of the divergence of the two sequences, which has to do with the mutation rate. In addition, the degradation/detoxification process of MWF by microbes is not yet fully understood. This makes it increasingly challenging to say why a gene is under a particular selection and whether the adaptation is to the environment or other microbes.

Originally, the dN/dS ratio was developed for analyses of genetic sequences of divergent species, where the differences represent fixation events. Two simplifying assumptions were often used: the sites of mutation are independent and do not influence each other, and no more than two mutations segregate in the population at a single site (Kryazhimskiy & Plotkin, 2008). However, the dN/dS ratio is often used to analyse samples from a single population, even though, in these cases, the differences do not represent fixation events but polymorphisms in the population. Kryazhimskiy and Plotkin (2008) investigated the expected dN/dS ratio under selection for exactly such samples. They found that the dN/dS ratio is relatively insensitive to the selection coefficient in this context. Furthermore, the relationship between selection and dN/dS does not follow a monotonic function. These two results mean that $dN/dS<1$ can occur under both negative and positive selection, making it impossible to infer selection pressure from dN/dS. Nonetheless, while the number dN/dS is not conclusive, I use it to support my speculations about the selection pressure on the fixed mutations. I hope this will reveal interesting targets for future studies and more informative methods.

Although the dN/dS ratio can only give limited information about the type of selection, we can infer a trend that if more non-synonymous SNPs occur, positive or neutral selection happens and if more synonymous SNPs occur, negative or neutral selection happens. For me, the most promising genes in terms of how a mutation affects evolution in MWF are fadD, frmR, nasR and yjiA. Therefore, these are addressed here, and the rest are discussed in detail in the appendix section C.

- **fadD/SNP 3.** I assume this gene is under positive or neutral selection because only NS SNPs occurred. The protein encoded by this gene, a long-chain-fatty-acid–CoA ligase, enables aerobic growth on fatty acids as a sole carbon and energy source. The substances in MWF are not degraded uniformly by microbes. This means that some substances are degraded faster than others. Rabenstein et al. (2009) has shown that fatty alcohol ethoxylates, alkyl amides, and fatty acids are degraded most rapidly. In addition, Cheng et al. (2005) found from previous studies that aerobic degradation of waste MWF is more efficient than anaerobic degradation. From this, I conclude that the long-chain fatty acid CoA ligase of this gene is important for the aerobic growth of the microbe and the efficient degradation of MWF. The fact that there is a mutation in this gene could indicate that it has been adapted to the conditions in MWF. Further, the fixed SNP occurs in monoculture and a SNP in this position was found in both replicates of the monoculture that survived. Therefore, I assume that a mutation of this gene represents an adaptation to the environment.

- **frmR/SNP 8.** I assume this gene is under negative or neutral selection because overall, more S SNPs were found. Nevertheless, a dN/dS ratio of 2 in microcosm 4 was calculated. However, the fixed SNP is NS and occurs in a co-culture. The protein is a repressor of the frmRAB operon, necessary for detoxifying formaldehyde. When formaldehyde is present, the protein is inactivated, and thus it can be detoxified. Formaldehyde releasing biocides are commonly used in MWF. Selvaraju et al. (2005) showed that biocides releasing formaldehyde are the most effective and that representatives of the genus *Pseudomonas* showed resistance to them up to a particular concentration. This could be because they are able to transform formaldehyde into formic acid. Therefore, it is important that this gene is not altered and is under negative selection.

- **nasR/SNP 11, 12, 15, 19, 24, 27.** I assume this gene is under positive selection because, in microcosm 2, a dN/dS ratio of 1, in microcosm 3, a dN/dS ratio of 3.25, and in microcosm 4, a dN/dS ratio of 4.25 were calculated. Additionally, of the fixed SNPs, only two are S the rest is NS. Monoethanolamine (MEA) is often a component of MWF and can be used by microbes. MEA probably converses into ammonia, which is then oxidised to nitrite and nitrate. Representatives of the genus *Pseudomonas* can use MEA as their sole carbon source, but only after a certain adaptation period. Rabenstein et al. (2009) found that the nitrate concentration in the emulsion he used increased at the beginning of the experiment and decreased later. When the nitrate and nitrites concentration was again very low, he found

close relatives of *C. testosteroni* and *Pseudomonas putida*. *C. testosteroni* can reduce nitrates under aerobic conditions. The nasR gene encodes a positive regulator of the mechanism required for nitrate and nitrite assimilation, which is induced by nitrate/nitrite (Wu et al., 1999). However, when the nasR gene is mutated, nitrate and nitrite induction of the nasF operon expression is stopped (Goldman et al., 1994). This speaks against an adaptation of the gene. The results available in this study can be explained by the fact that the number of NS mutations is overestimated, as mentioned above. However, four fixed SNPs occur in the monoculture and the remaining three in a co-culture. I assume that a mutation of this gene probably does represent an adaptation to the environment.

- **yjiA/SNP 38, 39, 40, 41.** I assume this gene is under positive or neutral selection because only NS SNPs occurred. MWF is susceptible to physical, chemical and microbial contamination, with microbes often being potential pathogens and/or deteriorgens. The most detrimental effects on MWF from microbes are that the stability of the MWF can be altered or the corrosion rate increased, shortening the lifespan of tools. Therefore, biocides are usually used to disinfect and keep microbial contamination under control. An alternative method is irradiation by UV light, which effectively inactivates microbes. The UV radiation causes photochemical damage to the nucleic acids, i.e., DNA and RNA. This prevents replication and cell division, eventually leading to cell death. It has already been shown that UV can reduce the microbial load in water, air, and surfaces. The main advantage over biocides is that no by-products are produced, and it does not require additional storage or disposal (Saha & Donofrio, 2012). The protein encoded by the gene yjiA is induced in response to DNA damage. Therefore, an adaptation of this protein might help to reduce DNA damage. Additionally, all fixed SNPs occur in a co-culture. However, I assume that a mutation of this gene is an adaption to the environment.

- **ropA_2/SNP 1.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the environment.

- **AGTU_00650/SNP 4.** I assume this gene is under positive or neutral selection. The function of the protein is uncertain. Therefore, I do not make an assumption about why there needs to be an adaptation of this protein.

- **yjgN/SNP 5.** I assume this gene is under negative selection. The function of the protein is unknown. Therefore, I do not make an assumption about why this gene has a mutation

- **AGTU_01326/SNP 6.** I assume this gene is under positive selection. It codes for an uncharacterised protein. Therefore, I do not make an assumption about why there needs to be an adaptation of this protein.
- **secA/SNP 7.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the other microbes.
- **trkH/SNP 16.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the environment.
- **COTE_04126/SNP 18.** I assume this gene is under neutral selection. It codes for an uncharacterised protein. Therefore, I do not make an assumption about why there needs to be an adaptation of this protein.
- **MISA_01549/SNP 30.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the environment.
- **puuP_3/SNP 31**. I assume this gene is under positive selection and that a mutation of this gene is an adaption to the environment.
- **ftsH_1/SNP 32.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the other microbes.
- **dedA_1/SNP 33.** I assume this gene is under neutral selection. It codes for an uncharacterised protein. Therefore, I do not make an assumption about why there needs to be an adaptation of this protein.
- **srlR/SNP 34.** I assume this gene is under positive or neutral selection and that a mutation of this gene is an adaption to the environment.

For future studies, I would first try to understand the degradation/detoxification of MWF, and the function of the proteins should be studied more precisely. Additionally, we should study the function of the proteins with yet unknown function. Lastly, I would suggest using a method to calculate dN/dS that accounts for all the concerns mentioned above.

5 Acknowledgements

First and foremost, I would like to thank my supervisor, Adamandia (Mado) Kapopoulou, for trusting me with this project and answering all my questions. Her feedback on the draft of this report was constructive and greatly appreciated. She provided me with moral support, and it was a pleasure to work with her.

I would also like to thank Claudia Bank and the THEE lab for making this work possible. They gave me a warm welcome, and the discussions were very helpful.

Also, I would like to thank my family and friends for their support and all their feedback. Without that support, I could not have succeeded in completing this project.

References

Bateman, A., Murzin, A. G., & Teichmann, S. A. (1998). Structure and distribution of pentapeptide repeats in bacteria. *Protein Science*, *7*(6), 1477–1480. https://doi.org/10.1002/pro.5560070625

Bertness, M. D., & Callaway, R. (1994). Positive interactions in communities. *Trends in Ecology and Evolution*, *9*, 191–193. https://doi.org/10.1016/0169-5347(94)90088-4

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). Quickgo: A web–based tool for gene ontology searching. *Bioinformatics*, *25*(22), 3045–3046. https://doi.org/10.1093/bioinformatics/btp536

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., ... Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research*, *49*(D1), D344–D354. https://doi.org/10.1093/nar/gkaa977

Chamberlain, S. A., Bronstein, J. L., & Rudgers, J. A. (2014). How context dependent are species interactions? *Ecology Letters*, *17*(7), 881–890. https://doi.org/10.1111/ele.12279

Cheng, C., Phipps, D., & Alkhaddar, R. M. (2005). Treatment of spent metalworking fluids. *Water Research*, *39*(17), 4051–4063. https://doi.org/10.1016/j.watres.2005.07.012

De Roy, K., Marzorati, M., Van den Abbeele, P., Van de Wiele, T., & Boon, N. (2014). Synthetic microbial ecosystems: an exciting tool to understand and apply microbial communities. *Environmental microbiology*, *16*(6), 1472–1481. https://doi.org/10.1111/1462-2920.12343

Denby, K. J., Iwig, J., Bisson, C., Westwood, J., Rolfe, M. D., Sedelnikova, S. E., Higgins, K., Maroney, M. J., Baker, P. J., Chivers, P. T., & Green, J. (2016). The mechanism of a formaldehyde-sensing transcriptional regulator. *Scientific Reports*, *6*(November), 1–15. https://doi.org/10.1038/srep38879

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. Version 1.14.2. https://CRAN.R-project.org/package=data.table

Friedman, J., & Gore, J. (2017). Ecological systems biology: The dynamics of interacting populations. *Current Opinion in Systems Biology*, *1*, 114–121. https://doi.org/10.1016/j.coisb.2016.12.001

Goldman, B. S., Lin, J. T., & Stewart, V. (1994). Identification and structure of the nasR gene encoding a nitrate- and nitrite-responsive positive regulator of nasFEDCBA (nitrate as-

similation) operon expression in Klebsiella pneumoniae M5al. *Journal of Bacteriology*, *176*(16), 5077–5085. https://doi.org/10.1128/jb.176.16.5077-5085.1994

Helde, R., Wieseler, B., Wachter, E., Neubüser, A., Hoffschulte, H. K., Hengelage, T., Schimz, K. L., Stuart, R. A., & Müller, M. (1997). Comparative characterization of secA from the α-subclass purple bacterium Rhodobacter capsulatus and Escherichia coli reveals differences in membrane and precursor specificity. *Journal of Bacteriology*, *179*(12), 4003–4012. https://doi.org/10.1128/jb.179.12.4003-4012.1997

Hurst, L. D. (2002). The Ka/Ks ratio: diagnosing the form of sequence evolution. *TRENDS in Genetics*, *18*(9), 486–487. https://doi.org/10.1016/S0168-9525(02)02722-1

Khil, P. P., Obmolova, G., Teplyakov, A., Howard, A. J., Gilliland, G. L., & Camerini-Otero, R. D. (2004). Crystal Structure of the Escherichia Coli YjiA Protein Suggests a GTP-Dependent Regulatory Function. *Proteins: Structure, Function and Genetics*, *54*(2), 371–374. https://doi.org/10.1002/prot.10430

Kim, S. K., Makino, K., Amemura, M., Shinagawa, H., & Nakata, A. (1993). Molecular analysis of the phoH gene, belonging to the phosphate regulon in Escherichia coli. *Journal of Bacteriology*, *175*(5), 1316–1324. https://doi.org/10.1128/jb.175.5.1316-1324.1993

Kryazhimskiy, S., & Plotkin, J. B. (2008). The population genetics of dN/dS. *PLoS Genetics*, *4*(12). https://doi.org/10.1371/journal.pgen.1000304

Meschiari, S. (2021). *Latex2exp: Use latex expressions in plots*. Version 0.5.0. https://CRAN.R-project.org/package=latex2exp

Mitri, S., & Richard Foster, K. (2013). The genotypic view of social interactions in microbial communities. *Annual Review of Genetics*, *47*, 247–273. https://doi.org/10.1146/annurev-genet-111212-133307

Piccardi, P., Vessman, B., & Mitri, S. (2019). Toxicity drives facilitation between 4 bacterial species. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(32), 15979–15984. https://doi.org/10.1073/pnas.1906172116

R Core Team. (2021). *R: A language and environment for statistical computing*. Version 4.1.1. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rabenstein, A., Koch, T., Remesch, M., Brinksmeier, E., & Kuever, J. (2009). Microbial degradation of water miscible metal working fluids. *International Biodeterioration and Biodegradation*, *63*(8), 1023–1029. https://doi.org/10.1016/j.ibiod.2009.07.005

Saha, R., & Donofrio, R. S. (2012). The microbiology of metalworking fluids. *Applied Microbiology and Biotechnology*, *94*(5), 1119–1130. https://doi.org/10.1007/s00253-012-4055-7

Selvaraju, S. B., Khan, I. U., & Yadav, J. S. (2005). Biocidal activity of formaldehyde and nonformaldehyde biocides toward Mycobacterium immunogenum and Pseudomonas fluorescens in pure and mixed suspensions in synthetic metalworking fluid and saline. *Applied and Environmental Microbiology*, *71*(1), 542–546. https://doi.org/10.1128/AEM.71.1.542-546.2005

Stachowicz, J. J. (2001). Mutualism, facilitation, and the structure of ecological communities. *BioScience*, *51*(3), 235–246. https://doi.org/10.1641/0006-3568(2001)051[0235:MFATSO]2.0.CO;2

The UniProt Consortium. (2018). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, *47*(D1), D506–D515. https://doi.org/10.1093/nar/gky1049

Van der Gast, C. J., Knowles, C. J., Starkey, M., & Thompson, I. P. (2002). Selection of microbial consortia for treating metal-working fluids. *Journal of Industrial Microbiology and Biotechnology*, *29*, 20–27. https://doi.org/10.1038/sj.jim.7000271

Van Der Gast, C. J., & Thompson, I. P. (2005). Effects of pH amendment on metal working fluid wastewater biological treatment using a defined bacterial consortium. *Biotechnology and Bioengineering*, *89*(3), 357–366. https://doi.org/10.1002/bit.20351

Van Der Gast, C. J., & Thompson, I. P. (2014). *Bioremediation*. (U.S. Patent 8,703,475 B2).

Van Der Gast, C. J., Whiteley, A. S., & Thompson, I. P. (2004). Temporal dynamics and degradation activity of an bacterial inoculum for treating waste metal-working fluid. *Environmental Microbiology*, *6*(3), 254–263. https://doi.org/10.1111/j.1462-2920.2004.00566.x

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). *Welcome to the tidyverse*. Version 1.3.1. https://doi.org/10.21105/joss.01686

Widder, S., Allen, R., Pfeiffer, T., Curtis, P., Wiuf, C., Sloan, W., Cordero, O., Brown, S., Momeni, B., Shou, W., Kettle, H., Flint, H., Haas, A., Laroche, B., Kreft, J., Rainey, P., Freilich, S., Schuster, S., Milferstedt, K., ... Soyer, O. (2016). Challenges in microbial ecology: Building predictive understanding of community function and dynamics. *ISME Journal*, *10*(11), 2557–2568. https://doi.org/10.1038/ismej.2016.45

Wu, S. Q., Chai, W., Lin, J. T., & Stewart, V. (1999). General nitrogen regulation of nitrate assimilation regulatory gene nasR expression in Klebsiella oxytoca M5al. *Journal of Bacteriology*, *181*(23), 7274–7284. https://doi.org/10.1128/jb.181.23.7274-7284.1999

Yang, Z., & Bielawski, J. R. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution*, *15*(12), 496–503. https://doi.org/10.1016/S0169-5347(00)01994-7

6 Figures and tables

SNP	Org.	Rep.	Pos.	Ref.	Alt.
1	AGTU.c01	1.1	703340	C	T
2		1.1	2351899	T	G
3		1.2	83654	C	T
4		3.2	677536	G	C
5		3.4	2863371	G	T
6		4.5	1372439	C	A
7	AGTU.c02	4.1	555455	G	T
8	AGTU.p02	4.4	181352	G	A
9	COTE.c01	2.1	574130	*NA*	C
10		2.1	574142	*NA*	G
11		2.1	5466695	A	C
12		2.2	5466897	A	C
13		2.3	574130	*NA*	C
14		2.3	574142	*NA*	G
15		2.3	5466897	A	C
16		2.3	5643748	G	C
17		2.4	3711132	A	C
18		2.5	4436504	T	G
19		2.5	5466537	T	G
20		3.1	574130	*NA*	C
21		3.1	574142	*NA*	G
22		3.3	574130	*NA*	C
23		3.3	574142	*NA*	G
24		3.3	5466905	G	T
25		3.4	574130	*NA*	C
26		3.4	574142	*NA*	G
27		3.4	5466552	G	A
28		4.5	574130	*NA*	C
29		4.5	574142	*NA*	G
30	MISA.c01	3.3	1628511	T	G
31		3.5	2587322	A	C
32		4.3	699272	C	T
33		4.5	1281972	T	G
34		4.5	3643534	A	C
35	OCAN.c01	4.1	1584485	C	T
36		4.3	1584485	C	T
37		4.4	1584485	C	T
38	OCAN.c02	4.1	2116572	G	C
39		4.3	2116572	G	C
40		4.4	2116572	G	C
41		4.5	2116572	G	C

Table 2: **All fixed SNPs.** Of all SNPs that occurred, 41 of them got fixed. The definition of a fixed SNP is when the frequency in the population is more than 0.98. This can be the case at the last time point of sequencing, t = 44, or earlier. If it is earlier, the frequency must not fall below 0.98 again; only then has a SNP been defined as fixed. SNP: a number I assigned to the SNP for identification;
Org.: organism and chromosome; Rep.: microcosm and replicate; Pos.: position on the chromosome; Ref.: reference nucleotide; Alt.: alternative nucleotide.

Org.	Rep.	SNP	Pos.	Ref.	Alt.	Gene	Start	End	Protein	ID
AGTU.c01	1.1	1	703340	C	T	ropA_2	703145	704191	Outer membrane protein IIIA	Q05811
	1.2	3	83654	C	T	fadD	82768	84477	Long-chain-fatty-acid–CoA ligase	P69451
	3.2	4	677536	G	C	AGTU_00650	677068	677808	Pentapeptide repeat protein Rfr32	B1WVN5
	3.4	5	2863371	G	T	yjgN	2863278	2864348	Inner membrane protein YjgN	P39338
	4.5	6	1372439	C	A	AGTU_01326	1372185	1372541	*hypothetical protein*	*NA*
AGTU.c02	4.1	7	555455	G	T	secA	553939	556656	Protein translocase subunit SecA	P52966
AGTU.p02	4.4	8	181352	G	A	frmR	181281	181562	Transcriptional repressor FrmR	P0AAP3
COTE.c01	2.1	11	5466695	A	C	nasR	5465708	5466982	Nitrate regulatory protein	Q48468
	2.2	12	5466897	A	C					
	2.3	15	5466897	A	C					
	2.5	19	5466537	T	G					
	3.3	24	5466905	G	T					
	3.4	27	5466552	G	A					
COTE.c01	2.3	16	5643748	G	C	trkH	5643413	5644879	Trk system potassium uptake protein TrkH	E1V6C5
	2.5	18	4436504	T	G	COTE_04126	4436374	4436670	*hypothetical protein*	*NA*
MISA.c01	3.3	30	1628511	T	G	MISA_01549	1628109	1629221	PhoH-like protein	P9WIA3
	3.5	31	2587322	A	C	puuP_3	2587129	2588556	Putrescine importer PuuP	P76037
	4.3	32	699272	C	T	ftsH_1	697714	699714	ATP-dependent zinc metalloprotease FtsH	P9WQN3
	4.5	33	1281972	T	G	dedA_1	1281677	1282456	Protein DedA	P0ABP6
	4.5	34	3643534	A	C	srlR	3643263	3644012	Glucitol operon repressor	P15082
OCAN.c02	4.1	38	2116572	G	C	yjiA	2115807	2116997	putative GTP-binding protein YjiA	P24203
	4.3	39								
	4.4	40								
	4.5	41								

Table 3: **Function of the protein.** List of the fixed SNPs belonging to a gene. Of the 41 fixed SNPs, 17 are intergenic, and the remaining 24 belong to 16 different genes. Org.: organism and chromosome; Rep.: microcosm and replicate; SNP: a number I assigned to the SNP for identification; Pos.: position on the chromosome; Ref.: reference nucleotide; Alt.: alternative nucleotide; Gene: name of the gene; Start/End: number of the nucleotide where the gene starts and ends; Protein: name of the protein the gene codes for; ID: UniProt ID.

UniProt ID	Gene	Protein	Organism	BLAST Ref.
A0A267EQ85	BOX15_Mlig020150g1	Uncharacterized protein	*Macrostomum lignano*	AGTU_01326_DNA
A0A267FH13	BOX15_Mlig030971g1	Uncharacterized protein	*Macrostomum lignano*	AGTU_01326_DNA
A0A1R4JP40	FM119_08580	Ribosomal-protein-alanine acetyltransferase	*Mycetocola reblochoni*	AGTU_01326_AA
A0A399J8P6	DWB68_15525	Uncharacterized protein	*Galactobacter valinip*	AGTU_01326_AA
A0A4S5BXX3	E8K88_02485	DUF2829 domain-containing protein	*Lampropedia aestuarii*	COTE_04126_DNA / COTE_04126_AA

Table 4: **BLAST results.** These are the results of the similarity search with BLAST. All information is taken from the UniProt data bank (The UniProt Consortium, 2018). BLAST Ref.: The reference sequence used for the similarity search. AGTU_01326_DNA/COTE_04126_DNA: nucleotide sequence of the gene; AGTU_01326_AA/COTE_04126_AA: amino acid sequence of the gene.

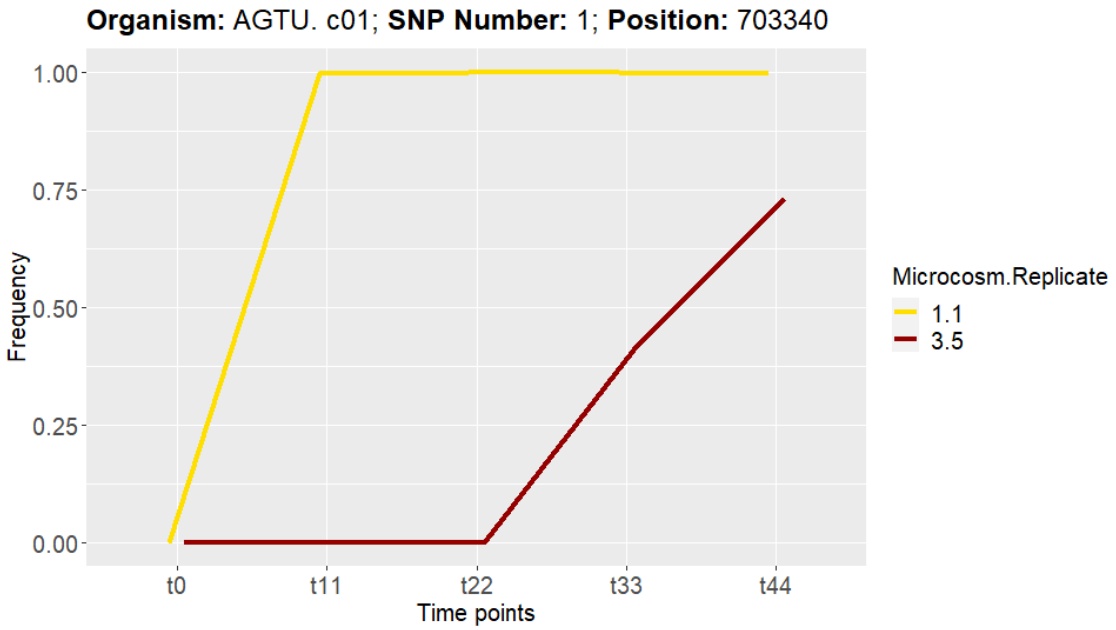

Figure 2: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. Both SNPs are in position 703340 of the first chromosome of *A. tumefaciens*. The fixed SNP (SNP 1) occurs in the monoculture (microcosm.replicate 1.1) and the additional SNP in the co-culture with *C. testosteroni* and *M. saperdae* (3.5). Both SNPs are non-synonymous and belong to the gene ropA_2.

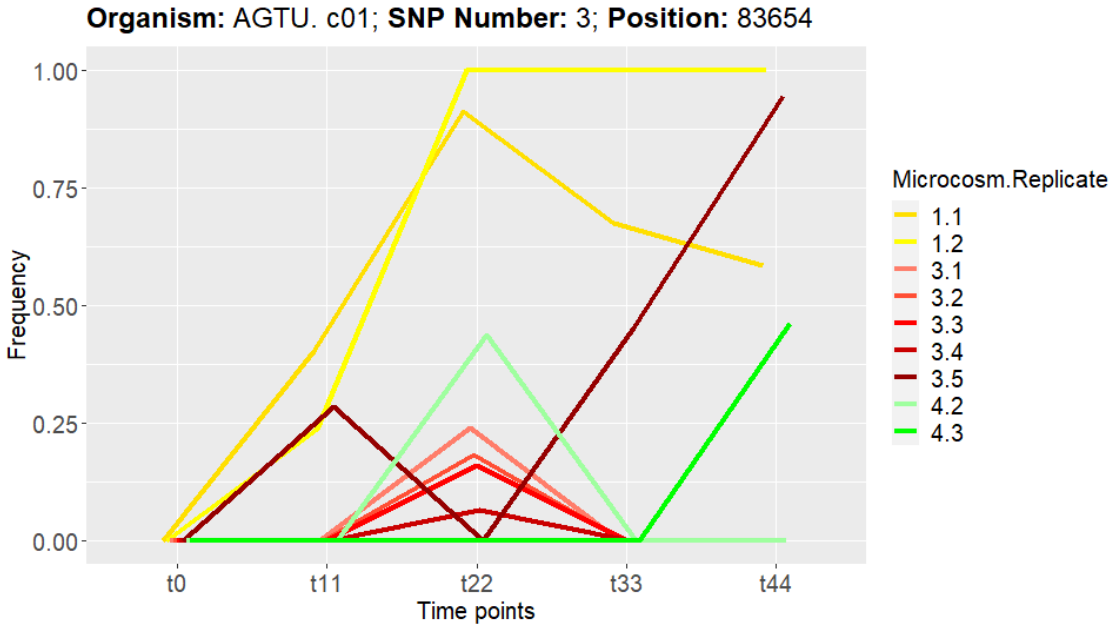

Figure 3: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. All SNPs are in position 83654 of the first chromosome of *A. tumefaciens*. The fixed SNP (SNP 3) occurs in the monoculture (microcosm.replicate 1.2). The additional eight SNPs are in the mono- (1.2) and co-cultures (3.1–3.5; 4.2, 4.3). All SNPs are non-synonymous and belong to the gene fadD.

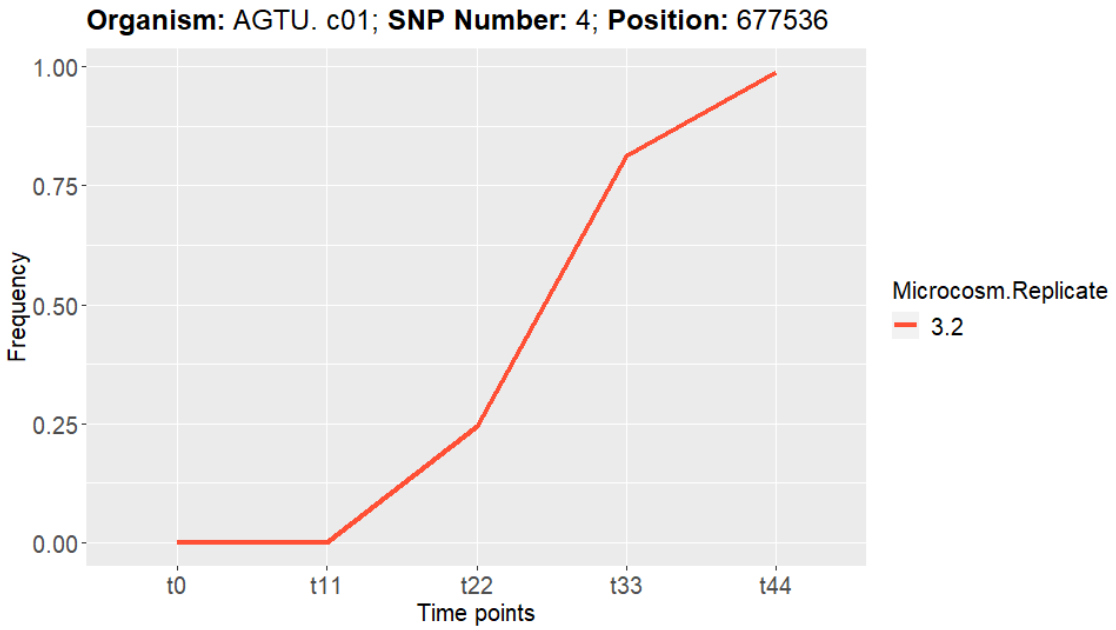

Figure 4: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 4 is in position 677536 of the first chromosome of *A. tumefaciens*. It occurs in the co-culture with *C. testosteroni* and *M. saperdae* (microcosm.replicate 3.2). The SNP is non-synonymous and belongs to the gene AGTU_00650.

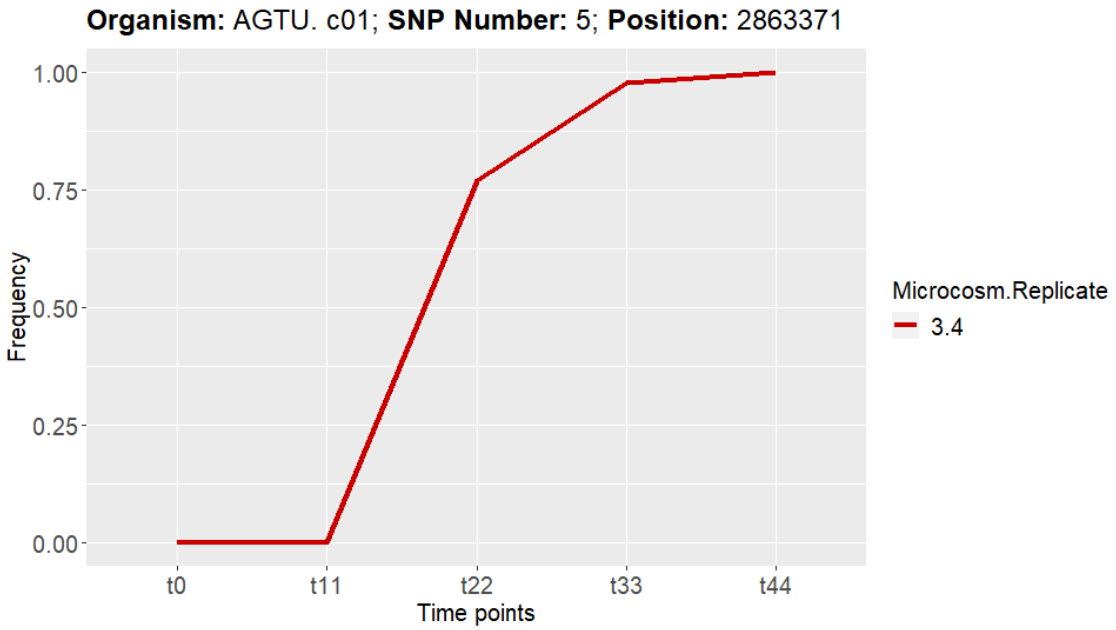

Figure 5: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 5 is in position 2863371 of the first chromosome of *A. tumefaciens*. It occurs in the co-culture with *C. testosteroni* and *M. saperdae* (microcosm.replicate 3.4). The SNP is non-synonymous and belongs to the gene yjgN.

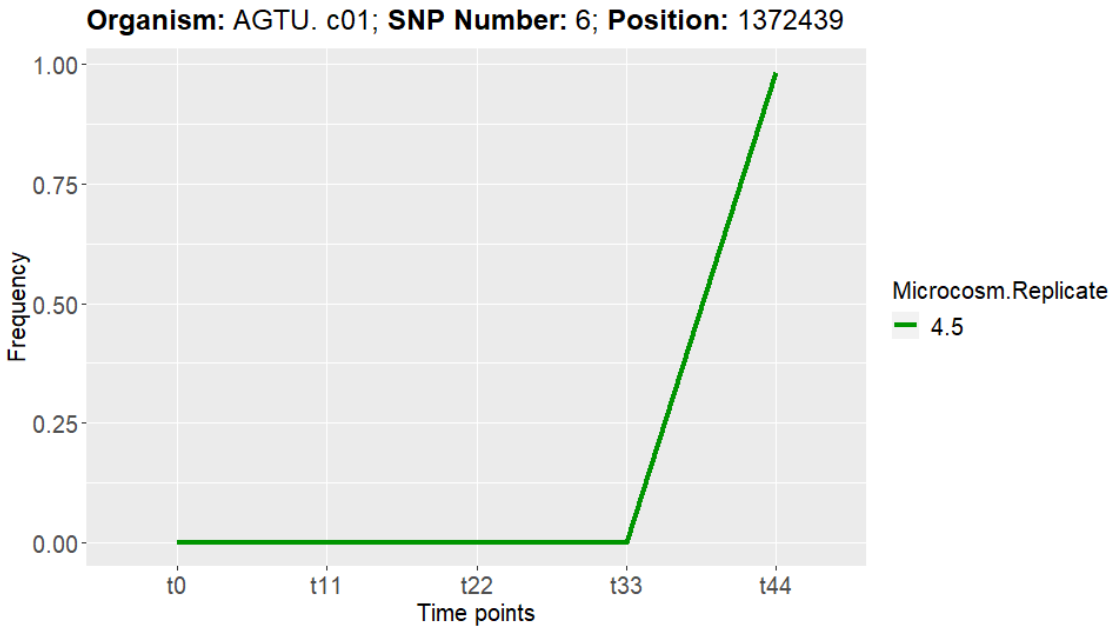

Figure 6: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 6 is in position 1372439 of the first chromosome of *A. tumefaciens*. It occurs in the co-culture with *C. testosteroni*, *M. saperdae* and *O. anthropi* (microcosm.replicate 4.5). The SNP is synonymous and belongs to the gene AGTU_01326.

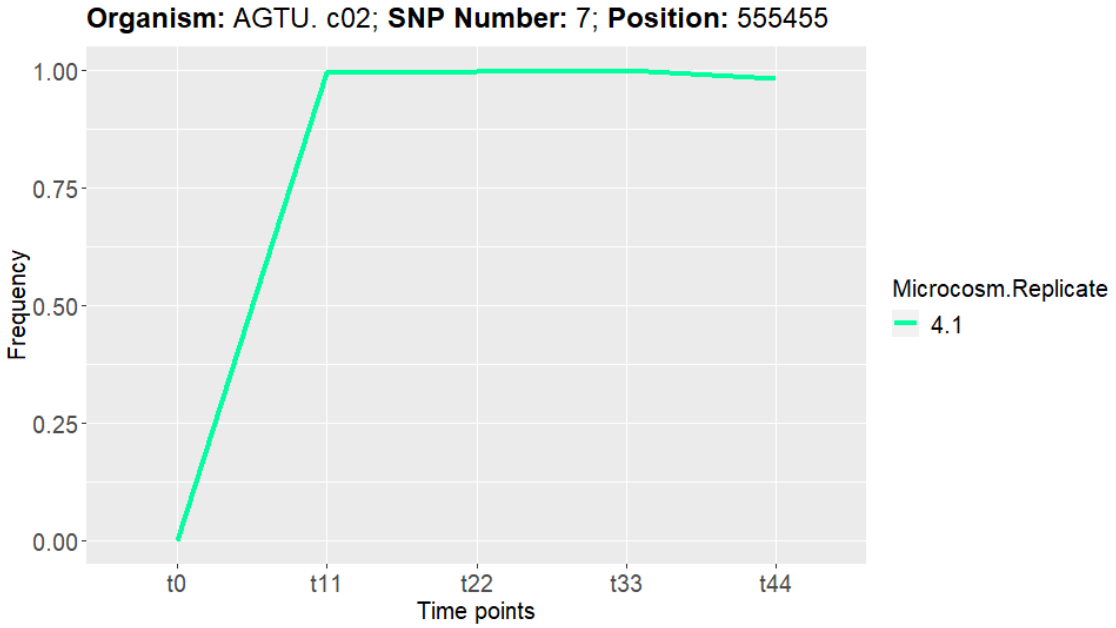

Figure 7: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 7 is in position 555455 of the second chromosome of *A. tumefaciens*. It occurs in the co-culture with *C. testosteroni*, *M. saperdae* and *O. anthropi* (microcosm.replicate 4.1). The SNP is non-synonymous and belongs to the gene secA.

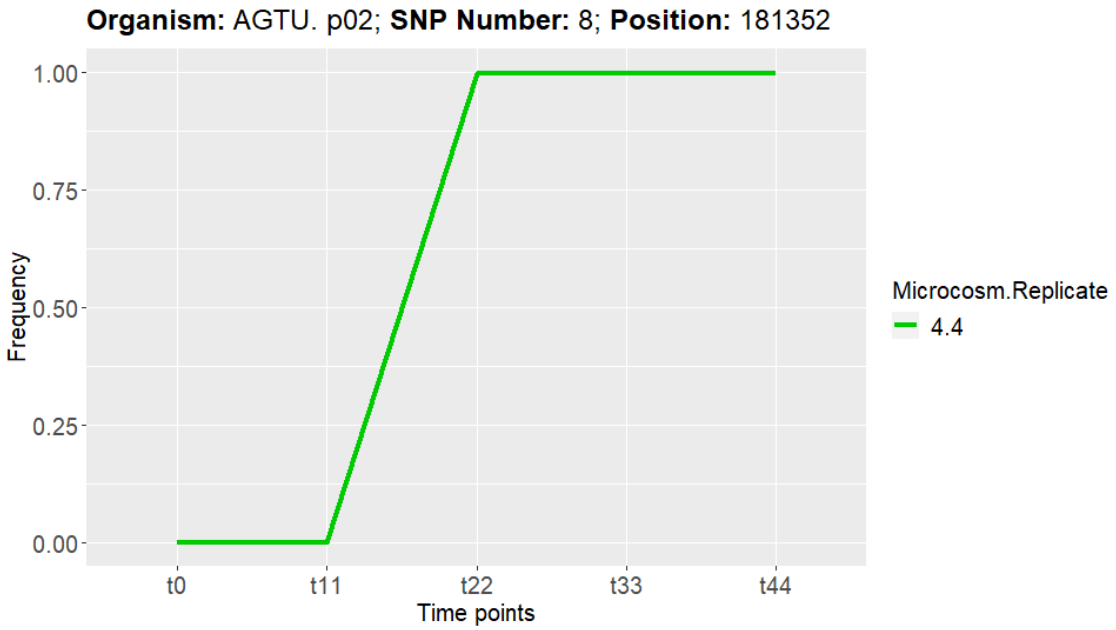

Figure 8: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 8 is in position 181352 of the second plasmid of *A. tumefaciens*. It occurs in the co-culture with *C. testosteroni*, *M. saperdae* and *O. anthropi* (microcosm.replicate 4.4). The SNP is non-synonymous and belongs to the gene frmR.

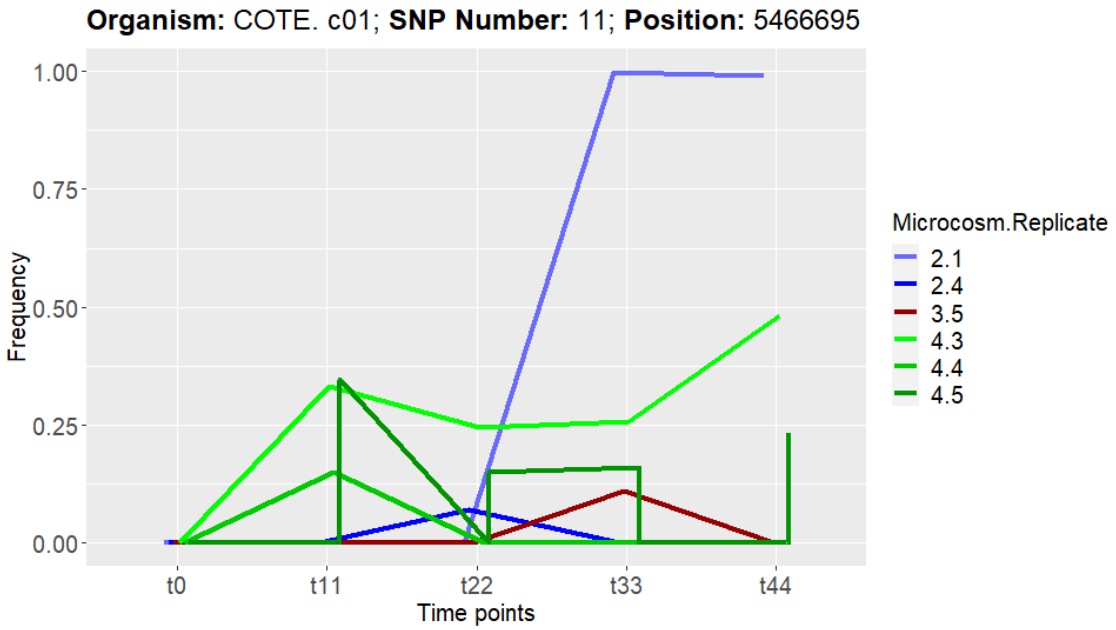

Figure 9: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. All SNPs are in position 5466695 of the first chromosome of *C. testosteroni*. The fixed SNP (SNP 11) occurs in the monoculture (microcosm.replicate 2.1). The additional six SNPs are in the mono- (2.4) and co-cultures (3.5, 4.3–4.5). All SNPs are non-synonymous except the second SNP in microcosm 4.5 and belong to the gene nasR.

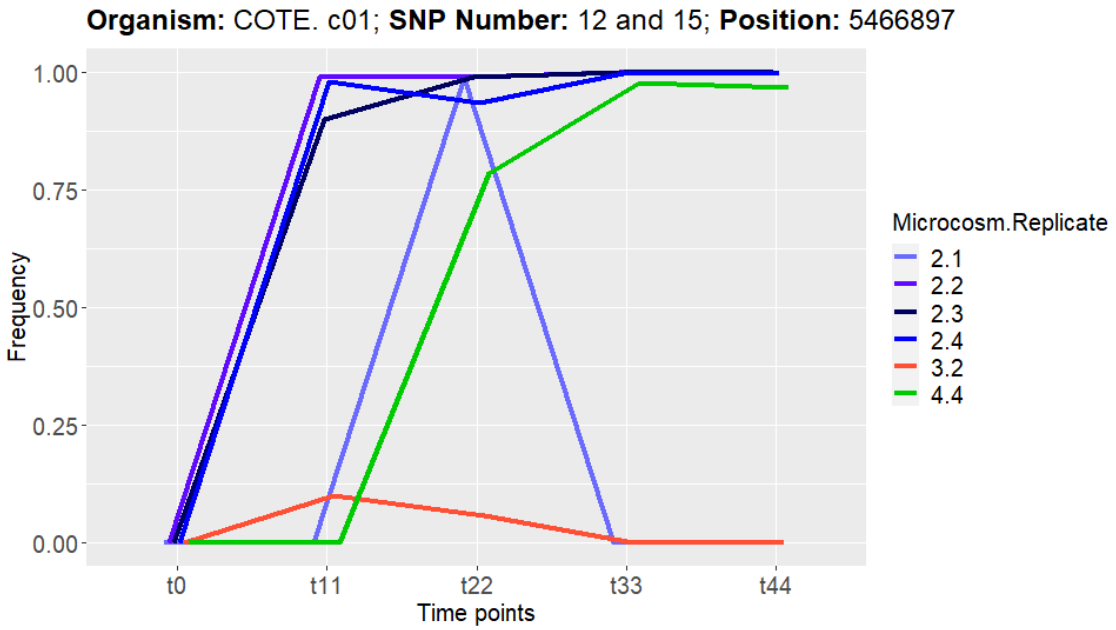

Figure 10: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. All SNPs are in position 5466897 of the first chromosome of *C. testosteroni*. Two SNPs are fixed (SNP 12 and 15) and occur in the monoculture (microcosm.replicate 2.2, 2.3). The other four SNPs are in the mono- (2.1, 2.4) and co-cultures (3.2, 4.4). All SNPs are synonymous and belong to the gene nasR.

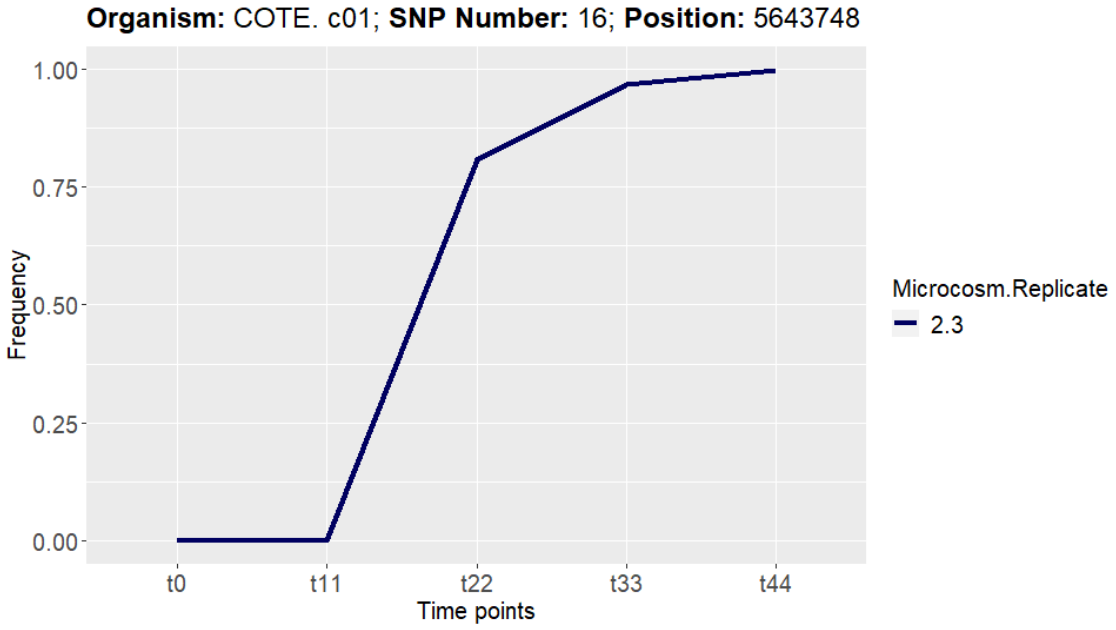

Figure 11: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 16 is in position 5643748 of the first chromosome of *C. testosteroni*. It occurs in the monoculture (microcosm.replicate 2.3), is non-synonymous and belongs to the gene trkH.

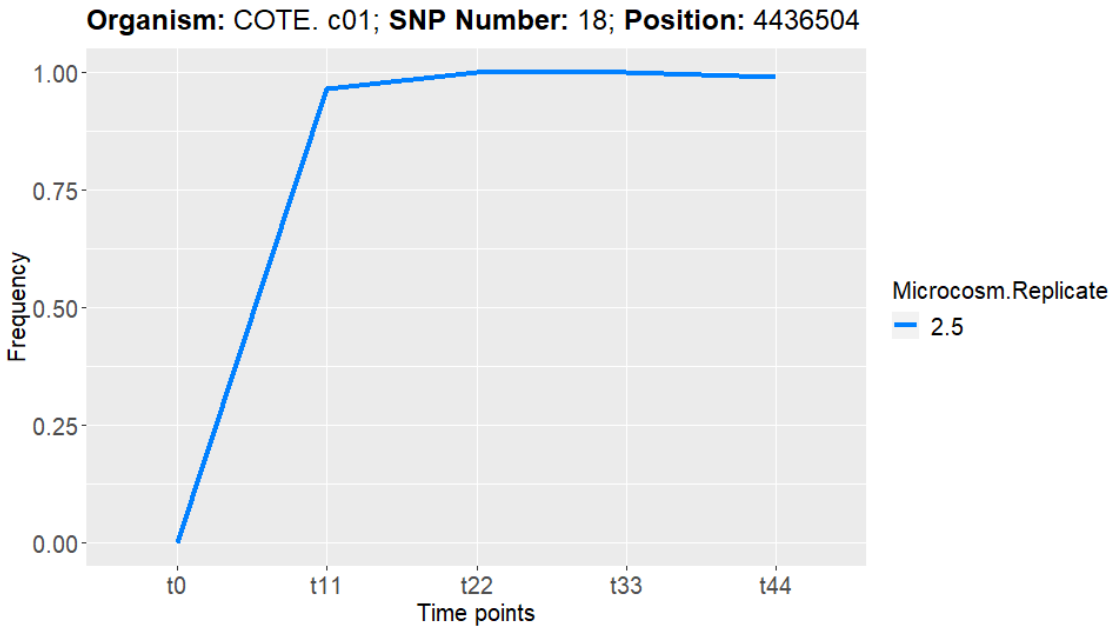

Figure 12: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. SNP 18 is in position 4436504 of the first chromosome of *C. testosteroni*. It occurs in the monoculture (microcosm.replicate 2.5), is non-synonymous and belongs to the gene COTE_04126.

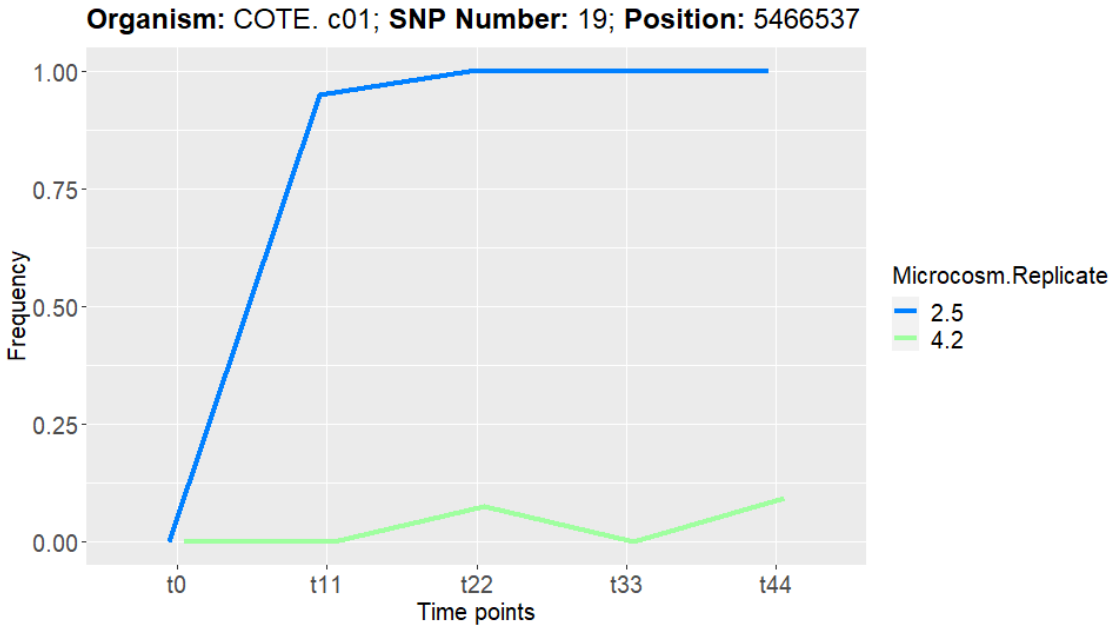

Figure 13: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. Both SNPs are in position 5466537 of the first chromosome of *C. testosteroni*. The fixed SNP (SNP 19) occurs in the monoculture (microcosm.replicate 2.5) and the additional SNP in the co-culture with *C. testosteroni*, *M. saperdae* and *O. anthropi* (4.2). Both SNPs are non-synonymous and belong to the gene nasR.

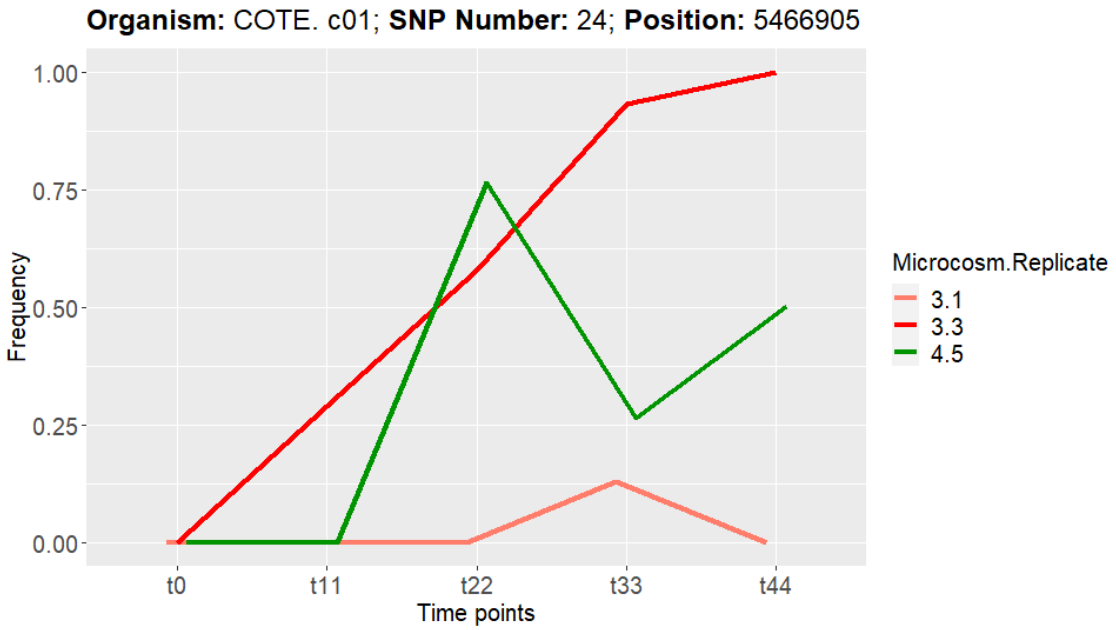

Figure 14: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. All SNPs are in position 5466905 of the first chromosome of *C. testosteroni*. The fixed SNP (SNP 24) occurs in the co-culture (microcosm.replicate 3.3) with *A. tumefaciens* and *M. saperdae*. The additional two SNPs are also in the co-cultures (3.3, 4.5). All SNPs are non-synonymous and belong to the gene nasR.

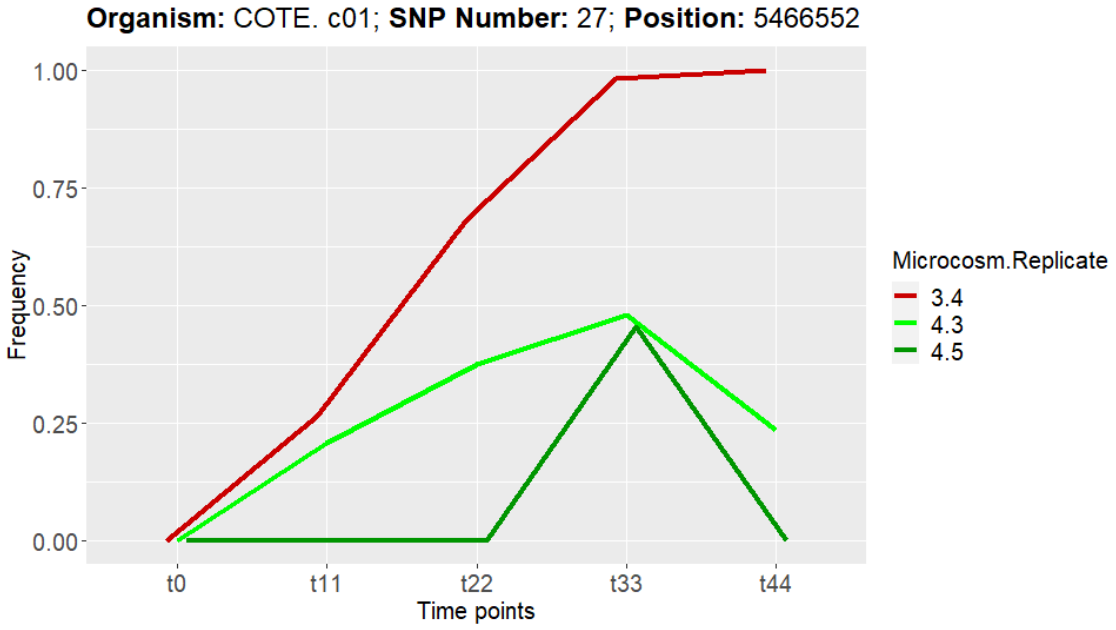

Figure 15: **SNP frequency over time.** X-axis: four sequencing time points; Y-axis: frequency of SNP. All SNPs are in position 5466552 of the first chromosome of *C. testosteroni*. The fixed SNP (SNP 27) occurs in the co-culture (microcosm.replicate 3.4) with *A. tumefaciens* and *M. saperdae*. The additional two SNPs are also in the co-cultures with *A. tumefaciens*, *M. saperdae* and *O. anthropi* (4.3, 4.5). All SNPs are non-synonymous and belong to the gene nasR.

Gene	Tag	SNP	Cosm.	count NS	count S	dN/dS
ropA_2	AGTU.c01_006770	1	1	1	0	*NA*
			3	1	0	*NA*
fadD	AGTU.c01_000790	3	1	2	0	*NA*
			3	5	0	*NA*
			4	2	0	*NA*
AGTU_00650	AGTU.c01_006500	4	3	1	0	*NA*
			4	2	0	*NA*
yjgN	AGTU.c01_027410	5	1	0	1	*NA*
			3	1	4	0.25
			4	0	5	*NA*
AGTU_01326	AGTU.c01_013260	6	4	5	2	2.5
secA	AGTU.c02_004900	7	4	1	0	*NA*
frmR	AGTU.p02_001750	8	1	0	3	*NA*
			3	0	2	*NA*
			4	4	2	2
nasR	COTE.c01_050720	11/12/15/19/24/27	2	4	4	1
			3	13	4	3.25
			4	17	4	4.25
trkH	COTE.c01_052100	16	2	1	1	1
			4	2	0	*NA*
COTE_04126	COTE.c01_041260	18	2	1	0	*NA*
			4	0	1	*NA*
MISA_01549	MISA.c01_015490	30	3	1	0	*NA*
puuP_3	MISA.c01_024740	31	3	10	8	1.25
			4	2	0	*NA*
ftsH_1	MISA.c01_006580	32	3	2	0	*NA*
			4	4	0	*NA*
dedA_1	MISA.c01_012280	33	3	0	2	*NA*
			4	3	0	*NA*
srlR	MISA.c01_035320	34	3	1	1	1
			4	2	0	*NA*
yjiA	OCAN.c02_020540	38/39/40/41	4	5	0	*NA*

Table 5: **dN/dS.** dN/dS is calculated for all fixed SNPs belonging to a gene. dN/dS is calculated for all microcosms over all replicates. SNP: number of fixed SNP belonging to the gene; Cosm.: microcosm; count NS/count S: number of non-synonymous/synonymous SNPs.