universität wien

MASTERARBEIT | MASTER'S THESIS

Titel | Title

predicting nitrite oxidation using evolutionary genomics

verfasst von | submitted by

Kilian Gandolf BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 875
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Masterstudium Bioinformatik
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Acknowledgements

Firstly, I want to thank my two amazing supervisors Daan and Ade for their hard work and guidance throughout the whole project.

Thank you also to members of CUBE and the Nitrification group, and my ever-changing officemates for creating a pleasant working atmosphere and providing help and feedback when needed.

Lastly, I want to thank my friends and family who supported me all these years.

Abstract

Nitrification is a multistep process that is a crucial component of the global nitrogen cycle. The last step of nitrification is nitrite oxidation. This project investigated the genomes of organisms capable of NarG-mediated nitrite oxidation. Phylogenetic analysis of NarG-sequences suggests that the nitrifying variant does not initially stem from the organisms that are now known to perform nitrite oxidation and is likely to be present in other organisms as well. Our goal was to identify properties common to nitrifying organisms that set them apart from their non-nitrifying relatives by the example of the well-known genera *Nitrobacter* and *Nitrococcus*. These properties should aid in the prediction of nitrite oxidation abilities in other organisms.

During the project, we conducted phylogenetic and pangenomic analyses to identify genes exclusive to these nitrifiers, with particular emphasis on colocalized gene clusters. Once the set of exclusive genes was known, we performed structure predictions to assess different types of interaction with the Nar complex.

One key discovery was the identification of the protein NarC, whose distribution across the NarG phylogeny suggests a conserved evolutionary role in the Nar complex. Structural predictions indicated a four-subunit complex (NarGHIC), and we investigated the possibility that the membrane-bound cytochrome PIP 736 might mediate electron transport between NarGHIC and COX, though this remains to be confirmed.

Zusammenfassung

Nitrifikation ist ein mehrstufiger Prozess der ein wesentlicher Bestandteil des globalen Stickstoffkreislaufs ist. Der letzte Schritt der Nitrifikation ist die Oxidation von Nitrit zu Nitrat. In diesem Projekt wurden die Genome von Organismen untersucht, die zur NarG-catalysierten Nitritoxidation fähig sind. Eine phylogenetische Analyse der NarG-Sequenzen deutet darauf hin, dass die nitrifizierende Variante nicht ursprünglich von den Organismen stammt, die heute als Nitritoxidierer bekannt sind, und daher auch in anderen Organismen vorhanden ist. Unser Ziel war es, Eigenschaften von nitrifizierenden Organismen zu identifizieren, die sie von ihren nicht-nitrifizierenden Verwandten unterscheiden, am Beispiel der bekannten Genera *Nitrobacter* und *Nitrococcus*. Diese Eigenschaften können helfen, die Nitritoxidationsfähigkeit in anderen Organismen vorherzusagen.

Im Rahmen des Projekts führten wir phylogenetische und pangenomische Analysen durch, um Gene zu identifizieren die exklusiv für diese Nitrifizierer sind. Schließlich führten wir Strukturvorhersagen durch, um verschiedene Interaktionsmöglichkeiten mit dem Nar-Komplex zu bewerten.

Eine wichtige Entdeckung war das Protein NarC, dessen Verbreitung in der NarG-Phylogenie auf eine konservierte evolutionäre Rolle im Nar-Komplex hindeutet. Strukturvorhersagen wiesen auf einen vierteiligen Komplex (NarGHIC) hin. Wir untersuchten auch die Möglichkeit, dass das membrangebundene Cytochrom PIP 736 den Elektronentransport zwischen NarGHIC und COX vermitteln könnte.

Abbreviations

anaerobic ammonia oxidation	-	anammox
complete ammonia oxidation	-	comammox
proton motive force	-	PMF
average nucleotide identity	-	ANI
single copy core	-	SCC
Markov chain clustering	-	MCL
hidden Markov model	-	HMM
cluster of exclusive genes	-	CEG
protein of unknown function	-	PUF
cytochrome c oxidase	-	COX
cytochrome c oxidase subunit 1/2/3	-	COX I/II/III
Nar-operon associated cytochrome	-	NarC
expected position error matrix	-	EPE
(interface) predicted template modelling	-	(i)pTM
potential interaction partner	-	PIP

1 - Introduction

1.1 - The Global Nitrogen Cycle

The global nitrogen cycle is a collection of biogeochemical processes in which nitrogen compounds are transformed across oxidative states ranging from -3 (ammonia) to +5 (nitrate) (Kuypers, Marchant, and Kartal 2018). Figure 1 shows the processes that make up the nitrogen cycle. Key steps are nitrogen fixation, where dinitrogen gas is transformed into ammonium (Beijerinck 1901), nitrification describes the oxidation of ammonium to nitrate (Winogradsky. 1890) and denitrification is the reduction of nitrate to N_2 (Kluyver and Donker 1926). In anaerobic ammonia oxidation (Anammox), ammonia and nitrite are transformed to N_2 in anaerobic conditions (van de Graaf et al. 1996).

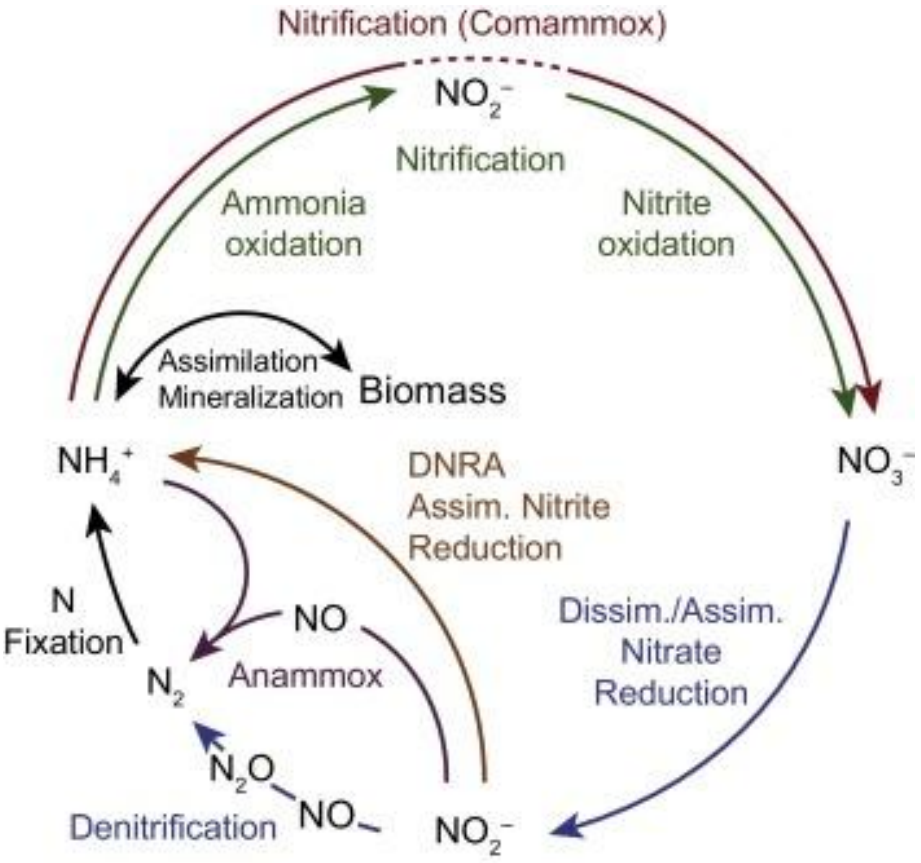


Figure 1. Schematic depiction of the global Nitrogen cycle

N_2 - Dinitrogen gas; **NH_4^+** - Ammonium; **NO_2^-** - Nitrite; **NO_3^-** - Nitrate; **NO** - Nitric oxide; **N_2O** - nitrous oxide *(Daims, Lücker, and Wagner 2016)*

Nitrogen is a nutrient essential for all life on earth as it is part of universal biomolecules such as DNA and proteins and can also be used as electron donor/acceptor in the

respiratory chain. Despite it being the most abundant element in our atmosphere, it is often a growth-limiting factor. This is because not all nitrogen compounds are equally reactive and accessible for use by organisms (Bernhard, Anne 2010). N_2 is only available to bacteria and archaea that are able to perform nitrogen fixation, while plants assimilate nitrogen in the form of ammonium or nitrate into biomass (Kuypers, Marchant, and Kartal 2018). This is why many plants enter a symbiotic relationship with nitrogen-fixing organisms, by forming nodules providing them with additional nitrogen (Robertson et al. 1985). Even still, more nitrogen is needed to fertilise the crops that sustain the human population. To alleviate this problem, humankind makes use of the Haber-Bosch process, which transforms the unreactive but very abundant N_2 in our atmosphere into bioavailable ammonium. This process was invented in 1908 by Fritz Haber and industrialised in 1911 by Carl Bosch (Haber 1920). It is estimated that 50% of the population is sustained through crops grown with Haber-Bosch synthesised fertiliser (Erisman et al. 2008). Human nitrogen consumption is on the rise, in 2008 ~100 Tg of nitrogen was fixed with the Haber-Bosch process and predictions say by 2150 we will increase the production to 200 Tg per year (Erisman et al. 2008). For comparison, all biological processes are estimated to make up 300 Tg of fixed nitrogen per year (Kuypers, Marchant, and Kartal 2018). However, a large percentage of the nitrogen put into the soil as fertiliser does not end up in crop biomass but is instead lost to the environment, leading to eutrophication events. Estimations of nitrogen use efficiency range between 30 and 50% (Anas et al. 2020; Erisman et al. 2008), which is partially due to the process of nitrification.

Nitrification is a series of oxidative reactions in which ammonia is transformed into nitrite and eventually nitrate. Ammonia oxidising bacteria (AOB) catalyse the oxidation of ammonia to nitrite, and then nitrite oxidising bacteria (NOB) transform the generated nitrite to nitrate (Winogradsky. 1890). In more recent years other mechanisms have been discovered as well - Comammox is the complete oxidation of ammonia to nitrate by a single organism (Daims et al. 2015). All of these reactions are catalysed by a few key enzymes that are linked to the electron transport chain to generate proton motive force. Quinones generally play an important role in these systems and act as an electron transporter between the different components of the ETC along the membrane (Simon and Klotz 2013). This thesis discusses the two-step nitrification by AOB and NOB, more specifically the last step which is catalysed by NOB.

The enzyme that NOB use to catalyse nitrite oxidation is called nitrite oxidoreductase (NXR). There are two types of NXR used by different groups of NOB. The majority of NOB use periplasmic NXR (Daims, Lücker, and Wagner 2016). Periplasmic NXR has

its catalytic subunit on the periplasmic side of the membrane, where protons derived from the oxidation reaction directly contribute to Proton motive force (PMF) across the membrane (Lücker et al. 2010). Cytoplasmic NXR's catalytic subunit sits on the cytoplasmic side of the membrane. Evolutionarily, cytoplasmic NXR is more closely related to Nitrate reductase (Nar) than to periplasmic NXR (Hemp et al. 2016). Because of this, cytoplasmic NXR is referred to as Nar from here on.

Only a few organisms are known that use Nar to perform nitrite oxidation (Daims, Lücker, and Wagner 2016), most prominent are ***Nitrobacter*** and ***Nitrococcus*** which were discovered and cultivated as early as 1892 and 1971 (Spieck and Bock 2015a; 2015b). These are the focus of the project. Other representatives of this category of NOB are *Thiocapsa KS1* - a phototrophic NOB (Hemp et al. 2016) and members of the *Chloroflexota* phylum (Spieck et al. 2020; Sorokin et al. 2014).

1.2 - Nar Protein Complex

The Nar complex is the key enzyme for the first step of denitrification. Usually, Nar enables organisms to use nitrate as an electron acceptor and reduce it to nitrite (Blasco et al. 1989). In this case, electrons are generated by oxidation of an often organic electron donor and transported to Nar via the quinone pool (Bertero et al. 2003). In contrast, when Nar runs in the oxidative direction, nitrite serves as the electron donor and is oxidised to nitrate. The electrons are transported from Nar to the terminal oxidase (COX). The exact transport mechanism is unknown but COX receives the electron from a cytochrome c, not from the quinone pool (Starkenburg et al. 2006). Organisms that use Nar to facilitate nitrite oxidation are chemolithoautotrophs, they create PMF across the membrane through chemical reactions without an organic electron donor and rely on CO_2 to assimilate biomass (Starkenburg et al. 2008).

Nar consists of three subunits, NarG is the largest, catalytic subunit and is located on the cytoplasmic side of the membrane along with NarH. NarI is a cytochrome b-type transmembrane subunit with two heme-binding domains for electron transfer (Bertero et al. 2003). Assembly of the complex is managed by the chaperone NarJ which is not part of the final complex (Lanciano et al. 2007). In *Thermos thermophilus* a fourth subunit has been described, a c-type cytochrome that has been proposed to serve as an alternate pathway for electron flow when nitrate is low (Cava, Zafra, and Berenguer 2008). It has also been proposed that this cytochrome may be part of the nitrifying *Nitrobacter* Nar complex (Starkenburg et al. 2008).

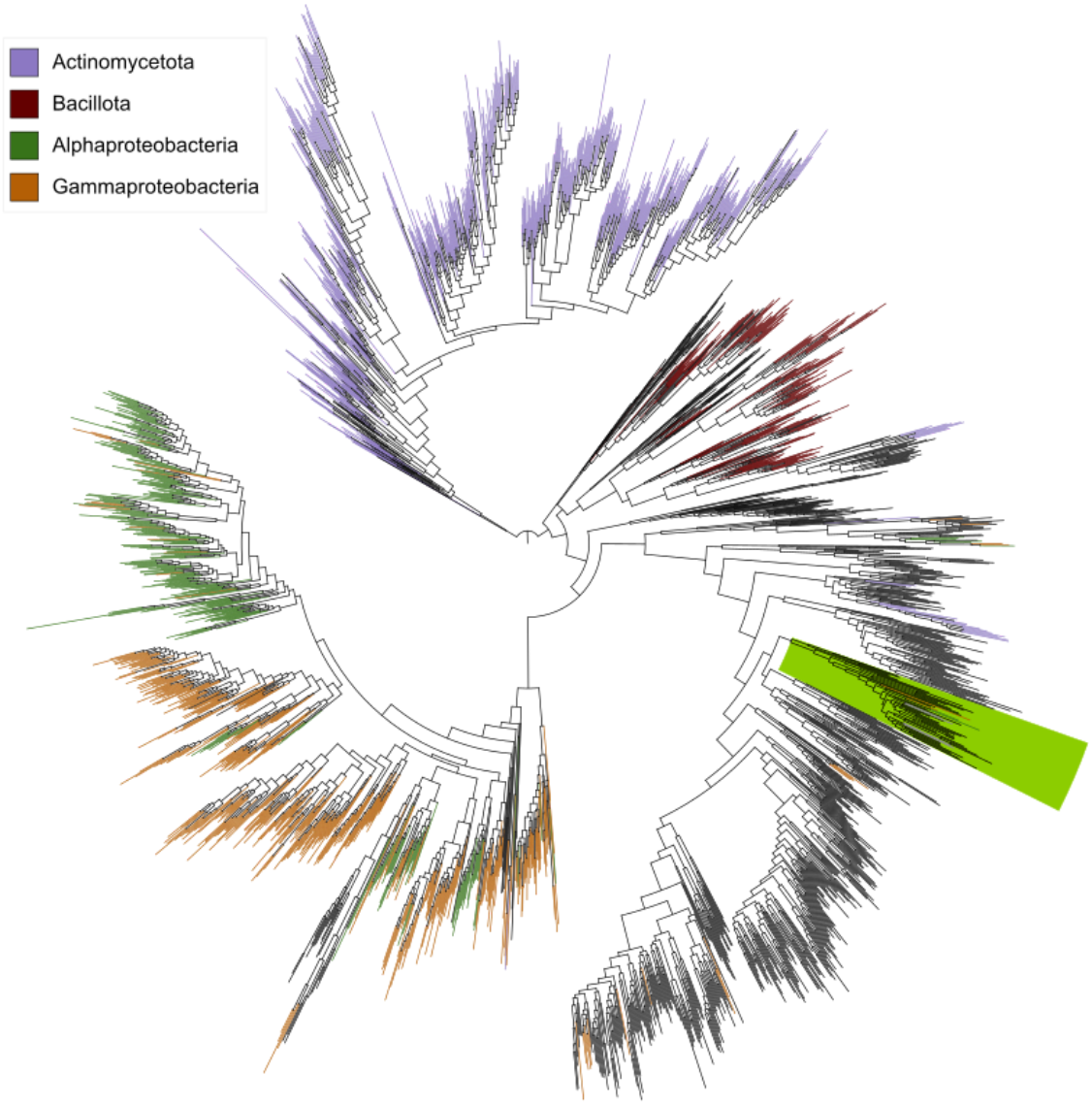


Figure 2. Phylogenetic tree of NarG-sequences.

10218 sequences were clustered at 85% identity which resulted in the 2523 centroid sequences that make up this tree. Branches are coloured according to the Phylum that the corresponding NarG sequence stems from. The clade highlighted in green contains all NarG sequences that stem from known nitrifiers, except the phototrophic *Thiocapsa KS1.*

Figure 2 shows a phylogenetic tree of NarG sequences. The depicted sequences are centroids resulting from clustering all NarG sequences in the project database consisting of 180,161 genomes from publicly available databases (see section 2.2) at 85 % identity. There are three prevalent phyla in the tree: Actinomycetota, Bacillota, and Proteobacteria, each with evolutionarily distinct NarG sequences that form separate clades, with little gene transfer between them. However, there is a fourth major clade in the tree that does not contain a majority of sequences from any phylum. The narG genes in this clade appear to have largely been horizontally acquired. This

major clade has a smaller subclade, which contains NarG sequences from the nitrifying *Nitrobacter*, *Nitrococcus* and different *Chloroflexota.* We hypothesise it is likely that this clade contains sequences from other nitrifying organisms that have not yet been identified. The rapid increase in availability of genomic data in recent years now gives us the opportunity to investigate this possibility using a bioinformatics approach.

1.3 - Project Aim

The goal of the project was to predict NarG-mediated nitrite oxidation ability from genomic data. To facilitate this we aimed to find properties common to nitrifying organisms that distinguish them from organisms that do not use Nar in the oxidative direction. The overall analysis was guided by the evolutionary history of the Nar-complex, as well as the evolutionary history of the genera *Nitrococcus* and *Nitrobacter*. A key assumption of the project is that genes that are present in nitrifying species, but not in their non-nitrifying relatives would be more informative than genes that are shared within the corresponding families. It was therefore tried to identify these **exclusive genes** through pangenomic analysis, investigate their distribution by the example of a representative genome and then infer their potential functions and interactions in the nitrification reaction through structure prediction.

2 - Methods

2.1 - Overview

The workflow is divided into three main parts. In the first and most broad scale part **whole genome comparisons** were performed. Through the creation of pangenomes a set of “exclusive genes” was identified that are present in *Nitrococcus* and *Nitrobacter,* but not in their closest relatives.
In the second part, these exclusive genes were further investigated. Here, genomic regions with clusters of exclusive genes were identified and additional functional annotations of genes of interest were done. **Gene cluster Comparison** was performed between these clusters of exclusive genes.
The third part focuses on the proteins themselves, **structural predictions** were made to infer about function of and interaction between proteins.
All of this was always done with the phylogenetic context in mind. The **Generation of phylogenetic trees** was therefore an essential part of the project that is described separately from the other methods.

2.2 - Data

Genomes used for this project stem from the genome taxonomy database (GTDB r214) (Parks et al. 2022), the genomic catalogue of earth's microbiomes (GEM) (Nayfach et al. 2021), the Searchable, Planetary-scale mIcrobiome REsource (SPIRE) (Schmidt et al. 2024) and the genomic catalogue of soil microbiomes (SMAG) (Ma et al. 2023). All genomes were taxonomically classified with *GTDB-Tk 2.3.2* (Chaumeil et al. 2022). In brief, GTDB-tk assigns taxonomy based on average nucleotide identity (ANI) and relative evolutionary divergence (RED). First, ANI screening is performed, to check if the genome belongs to a known species, a threshold of 95% ANI is used here. If no match is found, the genome is placed in a reference tree using pplacer (Matsen, Kodner, and Armbrust 2010) and the taxonomic level is assigned with the RED score (Chaumeil et al. 2020; 2022).
Not all genomes from each of the databases were used. Only the 95 % ANI dereplicated genomes (species representatives) were considered. The set of dereplicated genomes from each database was dereplicated again at 96% ANI and 50% aligned fragments. This was done in the order of CheckM/CheckM2 inferred quality scores (Parks et al. 2015; Chklovski et al. 2023) as listed in Table 1, meaning all representatives from the GTDB were used; from GEM only representatives that were

not in GTDB; from SPIRE only representatives that were neither in GTDB nor GEM; from SMAG only representatives that were not in any of the other databases.

	GTDB	GEM	SPIRE	SMAG	Sum
Total Genomes	402,709	524,046	1,160,000	40,039	2,126,794
Species representatives	85,205	45,599	92,134	16,530	239,468
Genomes included	85,205	12,022	70,892	12,042	**180,161**

Table 1. Databases and number of genomes used for the project.

GTDB-representatives : CheckM-inferred completeness - 5*contamination > 50%

GEM: CheckM inferred completeness ≥ 50% and contamination ≤ 5%

SPIRE: CheckM2-inferred completeness ≥ 50% and contamination ≤ 10%

SMAG: CheckM-inferred completeness ≥ 50% and contamination < 10%

2.3 - Tree Generation

Single copy core (SCC) genes are genes that are thought to be essential to - and therefore shared by - all members of a phylogenetic group and are only present once per genome. Selecting SCC genes for the analysis is meant to ensure that no sequences were subject to horizontal gene transfer and therefore track the phylogeny of the organisms they stem from. Comparing those sequences between different organisms thus provides a good estimation of the evolutionary distance between them. The list of SCC genes used in this project is based on those used in Laura Hug et al's Tree of Life (Hug et al. 2016) and can be found in the supplementary materials.

Concatenated SCC gene phylogenies were created following Anvio-8's workflow for phylogenomics (Eren et al. 2021), which works by extracting key genes from the target genomes using HMM-profiles based on the profiles used in GToTree (Lee 2019), performing MSA, concatenating the resulting alignments, and lastly inferring the phylogenetic tree based on the concatenated alignment. This involved generating anvio-readable "contigs databases" which store hits to HMM profiles for each genome and a file containing all paths to those databases and their names. This file could then

be used as input for *anvi-get-sequences-for-hmm-hits,* which in turn exports the desired HMM-hits in Fasta format.
The SCC HMM profiles were compared against the provided contig databases (genomes) to extract and concatenate SCC protein sequences in all target genomes. This was done with the –return-best-hit flag to obtain only one sequence per profile and genome. The concatenated sequences were aligned with muscle3 (Edgar 2022).
Next, the anvi-gen-phylogenomic-tree function was called to infer a phylogenetic tree based on the previously generated multiple sequence alignment using Fasttree 2.1.11 (Price et al. 2010). Fasttree constructs an initial tree with a nearest neighbour-joining approach and then refines the tree using a maximum likelihood (ML) method and subtree-pruning-regrafting (SPR). In the ML method, sequence evolution is simulated with a transition rate matrix, and the tree that best fits the simulated outcome is chosen. In SPR a subtree is pruned and put into a different position in the tree for evaluation. Both methods come with several heuristics to speed up the process.

Single protein phylogenies of NarG were done similarly. Sequences were obtained by filtering for the COG annotation in the gff-file. In the case of the Phylogeny of all NarG-sequences in the database clustering was performed with USEARCH v11.0.667_i86 (Edgar 2010) at 85% identity to reduce the initial set of 10218 sequences down to 2523. Instead of calling anvi-get-sequences-for-hmm-hits for MSA generation, *muscle 5.1.0*'s super5 algorithm was used directly (Edgar 2022). Muscle creates multiple sequence alignments in a divide-and-conquer approach by splitting the input sequences into smaller clusters and then building an MSA following a guide tree. This may introduce biases, which is why multiple variations with different guide trees are performed.

2.4 Whole genome comparison

2.4.1- Genome annotation

After the identification of suitable groups for pangenomic analysis, the chosen genomes were annotated with *Prokka-1.14.6* (Seemann 2014). Prokka is a well-established pipeline for quick annotation of bacterial genomes. It predicts a variety of genomic features by making use of other tools, specialising in different kinds of feature prediction:

Prodigal V2.6.3 is used for the prediction of coding sequences;
Barrnap 0.9 predicts rRNA genes (Hyatt et al. 2010);
ARAGORN v1.2.41 identifies tRNA genes(Laslett and Canback 2004) ;
and *Infernal* is used for the prediction of ncRNA(Eric P. Nawrocki, Sean R. Eddy 2013).
Since Prodigal only predicts coding sequences and not their function, hits are searched against the following databases for functional annotation: 16000 high-quality proteins from the UniProt (Apweiler et al. 2004) database are searched with Blast+, then, if a target genus is specified, corresponding complete genomes from the RefSeq database are searched. Lastly, Pfam (Punta et al. 2012) and TIGRFAM (Haft et al. 2013) are searched using hmmscan (Eddy 2011). Once one match (*e*-value ≤ 10^{-6}) for a given sequence is found, no further databases are searched (Seemann 2014).
Prokka requires a genome fasta file as input and produces annotations in the GFF3 and GBK file formats, alongside a few different fasta files describing identified protein and gene sequences. GFF files contain a header stating the file format and a description of all sequence regions - in this case contigs, as we are working with MAGs - present in the genome. After that, each feature predicted by Prokka is listed with nine tab-separated descriptors, the most important being: The contig where the feature is located, feature type, feature start and end, its orientation (+ or - strand), as well as a functional description/gene product. The file ends with the nucleotide sequences of all contigs.

2.4.2 - Pangenome creation

Pangenomes are the union of all genes in a set of genomes. They are commonly used to gain insight into the genomic evolution of a group of organisms. Typically, the genomes used for the creation of pangenomes stem from one species, but pangenomic analysis can also be applied to broader taxonomic units. Genes within a pangenome can be differentiated into two major groups - core genes and accessory genes. Core genes are genes that are shared by all genomes in the pangenome, accessory genes are not. The analysis of pangenomes yields insights into gene loss events and gene gain events through horizontal gene transfer and helps to infer when those events have happened in the evolution of the organism (Brockhurst et al. 2019).

Roary 3.13.0 (Page et al. 2015) was used to create the pangenomes in this project. Roary's basic workflow consists of retrieving all protein sequences from the input files, filtering sequences with 5% or more Ns, Iteratively clustering sequences with 98% or

more sequence identity using CD-hit and then performing all against all sequence comparison with BLASTp. Hits over a user-defined identity threshold are reported, and included in a network with nodes representing sequences and edges representing BLASTp hits over the user-defined identity threshold. The sequences are then clustered again with Markov chain clustering (MCL). MCL works by performing random walks through the network of sequences with edge weights based on the e-values reported by BLASTp. Clusters are then assigned based on the connectivity of the network, i.e. the distribution of probabilities of landing at each node in the network after starting at a node and taking a set number of steps. Because of the properties of Markov chains these probabilities are easily calculated. (Page et al. 2015) There are two main ways in which the network can be altered. One is the MCL inflation value, which scales the edge weights of the graph. The second is the identity threshold, which determines which edges are included in the network. We found that the inflation value had little effect on the roary output and therefore focussed on finding the appropriate identity threshold by comparing the outcomes at different values.

The GFF3 files produced by Prokka served as input to create a CSV file with all genes of the pangenome and their presence and absence across all genomes, as well as a text file containing summary statistics of genes classified in core and accessory groups. Additionally, dendrograms in newick format were created based on the binary presence or absence of genes. These trees are to be considered less accurate than trees based on sequence alignment and only serve as a rough estimation of the evolutionary trajectory. They can however be used, in conjunction with the CSV file, to obtain a visual representation of the results. This was done with the roary_plots.py function recommended in the roary documentation.

2.4.3 - Identification of exclusive genes

Genes that were part of the core in the nitrifying *Nitrobacter* or *Nitrococcus* and not present in any of the genomes of their closest relatives were considered **exclusive genes**. They were identified with a Python script iterating over all lines in the CSV file of gene presences and absences across all genomes. The script receives the path of the CSV file and the accession numbers of the genomes of interest as arguments and finds genes that are present in the genomes of interest and nowhere else. This minimum number of nitrifying genomes that a gene had to be present in to be considered exclusive was changed depending on the size of the group of interest, for the *Nitrobacter* group at least three of the 7 genomes needed to have a gene for it to

be considered, with 2 of the three being N. Winogradskyi and N. Hamburgensis, which are considered to be complete genomes. In the case of *Nitrococcus* 2 out of 3 members had to have it. These values were chosen to allow for genes missing from some of the MAGs while still not allowing too much noise. Since the majority of genes in the pangenomes were only present in one member and are not of interest, the threshold needed to include more than one genome. The exclusive genes were written into a CSV file in the same format as the input.

2.5 - Gene Cluster Comparison

2.5.1 - Functional annotation of exclusive genes

Many of the genes could not be functionally annotated in the initial round of functional annotations using prokka. Therefore a second round of annotation was done for the identified exclusive genes, this time using *eggNOG-mapper 2.1.1* (Cantalapiedra et al. 2021). eggNOG-mapper is another tool for annotation, it yields more precise annotations than Prokka but runs considerably slower because a substantially larger and more comprehensive search database is used. Similarly to Prokka, it predicts coding sequences with Prodigal, but hits are searched against the EggNOG 5.0.2 database which contains 4.4 million orthologous groups of proteins (Huerta-Cepas et al. 2019). EggNOG-mapper was run according to the two-step recipe recommended on Github, using *DIAMOND* in more-sensitive mode. DIAMOND is a sequence alignment tool that can achieve comparable accuracy to BLAST, while being much faster for searches against large databases (Buchfink, Reuter, and Drost 2021).

2.5.2 - Identification of HGT-rich genomic regions

I wrote a Python package to plot the distribution of exclusive genes along the genomes. Additionally, the annotations obtained from EggNOG were merged with the roary output. The genes were reorganised in the order that they appeared in a chosen **representative genome** (N. winogradskyi) that contained all exclusive genes for each group. In this way, a visual representation of the distribution of exclusive genes, which could easily be cross-referenced to see their functions, was achieved.

The package consists of 3 modules - the *main* module, the *gene_distribution* module and the *graphs* module - and a config file containing all relevant paths and accession numbers for the genomes of interest. The first step in the main module is to create a csv_reader object out of the CSV file with exclusive genes using Python's csv.reader

function. Then the *get_df* function of the *gene_distribution* module is called to obtain a pandas dataframe with further information on the exclusive genes.

Get_df receives the csv_reader object, a directory with gff files, a list of target genomes, and a path to EggNOG annotations. The header is read in the beginning to identify the positions of all genomes of interest. A for-loop runs over every identified position of interest for every line in the csv_eader. If the position in the line is not empty, it means the gene is present in that genome. In turn, information about the gene can be retrieved from the gff-file (such as contig, gene size, start and end positions, ...). The initial functional annotation by Prokka is modified such that each annotation is unique by adding a count to the end of it if another annotation with the same name already exists. Additionally, the EggNOG output file is read to obtain the EggNOG functional annotation as well as potential information about pathways the protein may be involved in. All of this information is stored and returned in the form of a pandas dataframe where each row represents one contig and all the exclusive genes and information about them are stored in lists.

In the main function, this “all-genomes-dataframe” (AGDF) is received and the rows with contigs belonging to the representative genome are put into a new dataframe with one row for each gene. The rows are sorted by their gene accession number and the “representative-genome-dataframe” is printed to a CSV file. Afterwards, the csv-reader is reset and passed to another function in the *genes_distribution* module.

This function only receives the csv_reader and the list of genomes of interest. A gene presence/absence data frame (GPA) is created by iterating over each line in the reader and storing the presence and absence of each gene in the form of a 1 or a 0. Again, the functional annotations are modified to be unique. This is done in the same way as in the *get_df* function so that the outputs can easily be cross-referenced. The GPA is returned to the main function.

All necessary inputs for visualisation are now assembled. The first visualisation is achieved by the *get_pa_matrix* function in the *graphs* module. It receives the GPA from *main,* turns it into an R DataFrame and then creates a gene presence/absence matrix using the *heatmaply* R-library. R can be run in Python by importing the *rpy2* module. (“Rpy2 Github” [2018] 2024; “Heatmaply Documentation,” n.d.)

The second visualisation is a plot of the exclusive genes along the length of each contig. For this purpose, I wrote the *plot_gene_distribution* function in the *graphs* module. It receives the GPA, AGDF and a path to the desired output directory from the main function. For each row (contig) in the AGDF, gene starts, lengths and products (Prokka annotations modified to be unique) are retrieved as lists. A for-loop starts, iterating over all items in the three lists.

In the loop the number of genomes sharing the current gene is calculated by getting the sum over the corresponding row in GPA. Then a marker with size corresponding to the gene length and colour based on the number of genomes sharing the gene is added to the current plot. The x-value corresponds to the gene start, the y-value corresponds to the gene count. The gene count was added, so that genes would not clump together on the plot. Once the loop over all genes is done, the plot is saved to the output directory and the next iteration starts with genes from another contig.
The third visualisation function in the *graphs* module is called *plot_gene_distribution*. It works similarly to the second visualisation. The only difference is, that genes are only plotted if another exclusive gene is within 5 or fewer positions (based on accession number, not on nucleotide position) on the same contig.

2.5.3 - Exclusive gene cluster alignment

Clusters of exclusive genes were compared with *clinker 0.0.28. Clinker* (Gilchrist and Chooi 2021) is a Python-based tool for the alignment and interactive visualisation of gene clusters. It takes as input either GenBank files or GFF files with specified regions that are to be aligned. The alignment is done using the Biopython package with scoring according to BLOSUM62. The interactive visualisation is achieved using an accompanying JavaScript library called *clustermap.js.* (Gilchrist and Chooi 2021)
To generate the necessary input for *clinker* to visualise all clusters of exclusive genes I wrote a Python script. The script loops over each gene in the previously generated representative-genome CSV file that contains all exclusive genes. If a cluster of exclusive genes is found, the counterparts for all genes in the cluster across all other genomes are gathered. 2 or more genes within a range of 5 accession numbers were considered as a cluster. The output is a file with a list of commands to run clinker with correct filenames, contig and region specifications that has to be run from the directory containing all necessary GFF files.

2.6 - Protein Structure Predictions

2.6.1 - Signal peptide trimming

SignalP 6.0 (Teufel et al 2022) was used to predict and trim signal peptides from protein sequences. SignalP predicts all known types of signal peptides through the use of protein language models (PLMs). PLMs are machine learning models trained on large amounts of protein sequences that can be used for functional predictions. SignalP 6.0 offers two models for prediction, a reduced and a full model. The reduced

model is smaller and therefore less accurate but much faster than the full model. The required input is a (multi-)fasta file, the outputs are text files noting the predicted signal peptide locations, a classification in the 5 types of signal peptides as well as a fasta file with the trimmed sequences (Teufel et al 2022).
The prediction time of signal peptides was not a bottleneck in our workflow, therefore the bigger model was chosen by running with the *-m slow-sequential* flag. Otherwise, default values were used.

2.6.2 - Structure predictions

When structure predictions were started the latest AlphaFold version was AlphaFold2 (v2.3.1) (Jumper et al. 2021)), over the course of the project AlphaFold3 was made available (Abramson et al. 2024). Because of this, predictions were first made with AlphaFold2 and ColabFold (Mirdita et al. 2022), and later during the project a switch to AlphaFold3 was made.

AlphaFold2 (Jumper et al. 2021) is a machine learning system that uses multiple sequence alignments (MSAs) to predict three-dimensional protein structures based on the assumption that interacting residues will co-vary - if one residue changes, the other one is expected to adapt as well. The first step of the AlphaFold workflow is MSA generation. For this purpose the input sequence is searched against the Big Fantastic Database (BFD) database with jackHMMER and HHBlits, two widely used hidden Markov model (HMM) based alignment tools known for their speed (Johnson, Eddy, and Portugaly 2010; Remmert et al. 2012). The BFD contains more than 2 Billion protein sequences represented as HMMs and MSAs and clustered into families. After the MSA is assembled, it is fed into AlphaFold2's neural network, which consists of two main components. The first component is the **Evoformer**. The Evoformer processes the MSA and a pair representation matrix through a series of operations that focus on capturing relationships both within the MSA (sequence relationships) and between residue pairs (spatial relationships).
The second element of the network then uses this information to infer a three-dimensional structure. It starts by placing all atoms at the same point and then iteratively refines the predicted structure to adhere. Lastly, the predicted structure is "relaxed" by putting it in a forcefield using the amber molecular dynamics simulation program. This is meant to ensure that all solutions are physically feasible (Jumper et al. 2021; Case et al. 2005).

AlphaFold2 was run using the container software singularity (v3.8.6). A container is a piece of software that packages up an application and its dependencies. This way the code can be run reliably in different environments. The singularity script receives the necessary filepaths and model specification (multimer for multimer prediction, monomer for monomers), loads a container image that can also be specified (AlphaFold:2.3.1 was used) and then runs AlphaFold. The original script was slightly modified by commenting out these lines in the environment variables definition:
"TF_FORCE_UNIFIED_MEMORY": "1",
"XLA_PYTHON_CLIENT_MEM_FRACTION": "4.0",
The inclusion of these lines led to a GPU memory access error when the input was too large.
The script produced the following outputs:
A directory containing generated MSAs; 25 structural predictions in pdb format - before and after relaxation; a ranking of the predictions, and additional information including confidence values about each prediction in pkl format.

ColabFold (Mirdita et al. 2022) is a structure prediction tool based on AlphaFold2. It can be used through a few different Jupyter notebooks hosted on Google servers. The implementation used for this project makes use of MMseqs2 (Steinegger and Söding 2017) for faster MSA generation. It has access to the PDB70 and UniRef100 databases. ColabFold also allows the user to upload a custom alignment to skip the MSA generation step. This is also possible in AlphaFold2, but the ColabFold implementation only requires one MSA file whereas AlphaFold2 requires five MSAs in different formats. The custom MSAs for ColabFold were generated by aligning the input sequences with *MAFFT v7.526* (Katoh et al. 2002) and then reformatting the output to a3m format using the reformat.pl script of HHsuite v3.0 (Steinegger et al. 2019). This way it was possible to compare results generated with regular MSAs created by AlphaFold to MSAs against the project database that we identified through our generated phylogenies as being evolutionarily related. Since AlphaFold relies on MSAs to capture patterns of co-evolving residues, we thought our custom MSA with related sequences may yield better results, despite the smaller size,

AlphaFold3 (Abramson et al. 2024) is the latest version of AlphaFold at the time of this writing. The main difference to AlphaFold2 is the change of the Evoformer module to a simpler and faster “Pairformer” module. Similar to Evoformer, this module needs MSAs as input but it does not focus on processing the MSA, instead, the emphasis lies on generating a pairwise representation. This is the only input for the second block, a

diffusion model that is trained on noisy structures and learns to predict the true underlying atomic coordinates. AlphaFold 3 predicts structures with higher accuracy than previous versions, especially when it comes to the prediction of complexes. Additionally, it is capable of incorporating some ligands and ions into its prediction.

Since the source code of AlphaFold3 was not made available, it can only be run on Google Deepmind's AlphaFold server (https://alphafoldserver.com) by uploading Protein sequences in Fasta format and manually adding Ligands, if needed. The best-ranked result can be viewed directly on the server, the full results can also be downloaded. Those include five ranked predictions in cif format, a summary of confidence values for all structures and a summary of the job request that can be used to rerun the job.

3 - Results

The results are organised into four main sections. First, we constructed phylogenetic trees for *Nitrococcus* and *Nitrobacter*, showing that NarG was horizontally acquired in both groups. Next, we investigated through pangenomic comparisons which other genes were horizontally acquired. We operate under the assumption that these exclusive genes could provide insights into additional adaptations, beyond acquiring the Nar complex, that enabled these organisms to carry out nitrite oxidation. The third section focuses on analyzing the distribution of exclusive genes by the example of a representative genome. Finally, we conducted structure predictions for the complexes and proteins identified in our previous analyses.

3.1 - Phylogenetic trees

To identify suitable outgroups for the pangenomic analysis of *Nitrococcus* and *Nitrobacter*, we made phylogenetic trees of their closest relatives and corresponding narG sequences.

The genus *Nitrococcus* belongs to the *Nitrococales* order which has 65 species representatives in our genome database. In these 65 genomes, 21 narG sequences stemming from 20 different genomes were found (Figure 3). The *Nitrococcales* are divided into four families: The *Halorhodospiracea*, *Aquisalimondaceae*, *AK92*, and *Nitrococcaceae* which also include the genus *Nitrococcus*.
Figure 3 shows the phylogeny of the *Nitrococcales* and their respective narG-sequences. narG is present in all families of the *Nitrococcaceae*, though it is most prevalent in the *Aquisalimondaceae* (11 out of 14 genomes, 79%). In the other families, narG appears to be mostly present in the deep-branching species, suggesting multiple gene loss events within the *Nitrococcales* order.

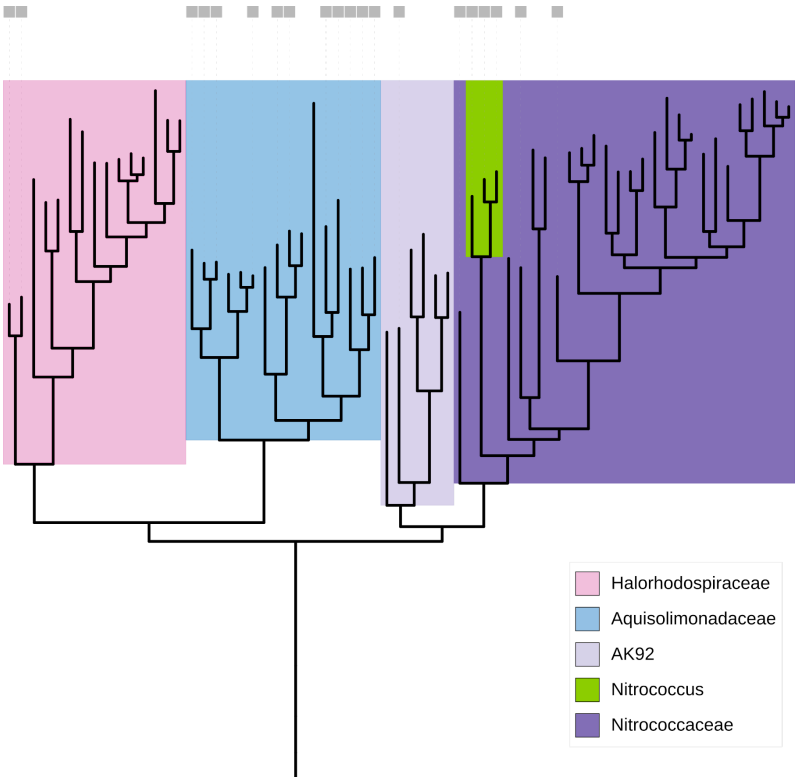


Figure 3. Phylogenetic tree of the *Nitrococcales*.
The tree is based on the alignment of concatenated core genes (see section 2.3). Clades are coloured according to the family they represent and the presence of narG is highlighted in grey on the outside of the tree.

The phylogeny of corresponding NarG sequences in Figure 4 is mostly congruent with the concatenated core gene tree (see section 2.3) as the branching order is largely the same in the NarG phylogeny and the species tree. One exception to this is the clade of 4 *Aquisalimonadaceae* NarG sequences which cluster with the two *Halorhodospiraceae* sequences in the left-most clade of the tree. Still, the NarG sequences of the *Aquisalimonadaceae, Holorhodospiraceae, Nitrococaceae* and *AK92* cluster closely together, indicating a shared evolutionary origin and therefore likely vertical inheritance. In contrast to this, the NarG sequences belonging to *Nitrococcus* cluster outside of the other groups on a comparatively long branch, indicating a horizontal gene-transfer event.
Overall, this makes the *Nitrococaceae* a suitable outgroup for pangenomic analysis. They are the closest relatives to *Nitrococcus* and none of them have acquired the same version of narG as *Nitrococcus*, meaning if another gene was acquired alongside narG to allow for nitrification, it will likely not be present in any of the *Nitrococaceae* and can thus be identified through comparative genomics.

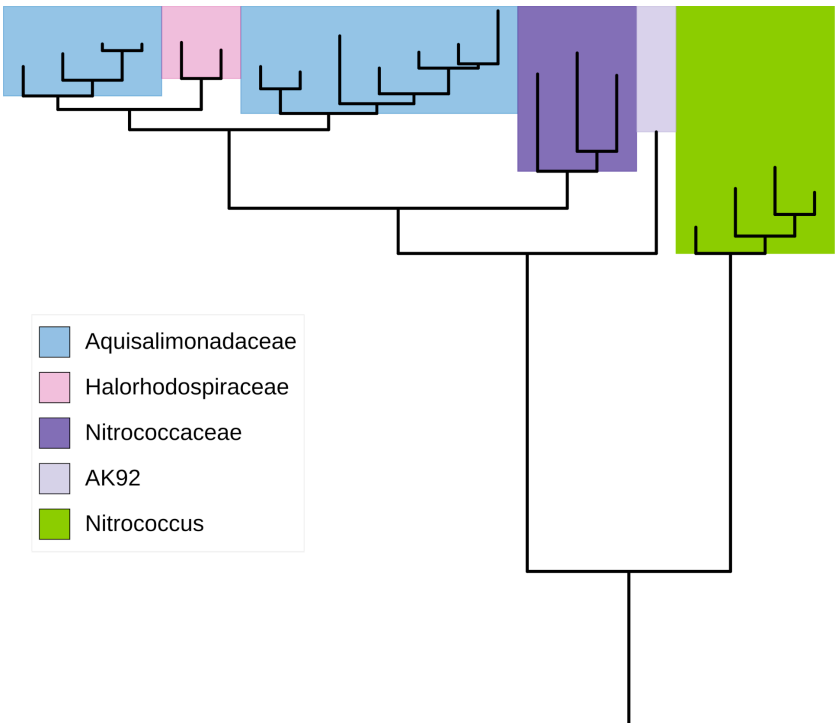


Figure 4. Phylogenetic tree of NarG-sequences of the *Nitrococcales*.
Clades are coloured according to the family that sequences stem from. The sequences stemming from Nitrooccucs (green) cluster outside of the other sequences, indicating they were acquired through horizontal gene transfer.

The genus *Nitrobacter* is a member of the *Xanthobacteraceae* family with 1370 representative genomes in the database. The full tree of all 1370 genomes can be found in the supplementary materials. Based on the phylogeny of the 1370 members of the *Xanthobacteraceae,* we selected a subset of 404 genomes from the 5 closest genera to *Nitrobacter* and compared concatenated core gene and NarG phylogenies. These 404 genomes contained 78 NarG sequences.

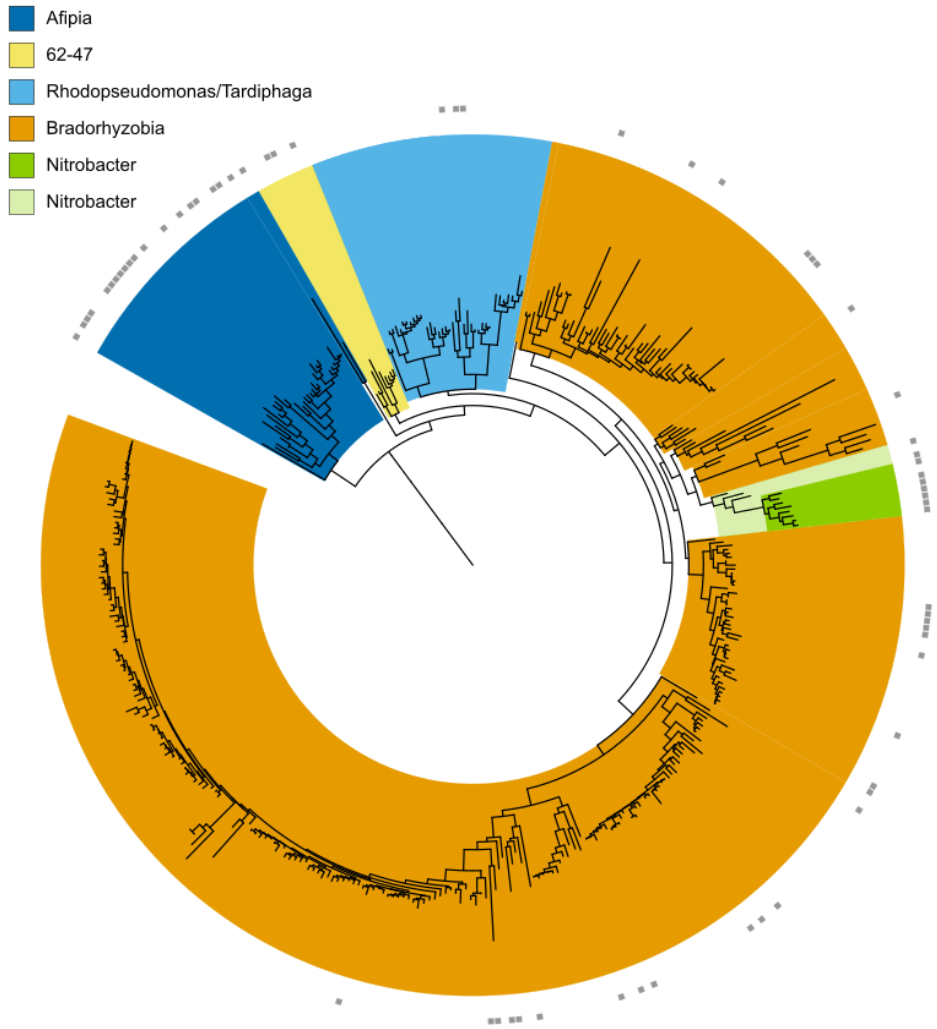


Figure 5. Phylogenetic tree of *Nitrobacter* and the 5 closest branching genera within the *Xanthobacteraceae*.

Clades are coloured according to the genera they represent. The *Rhodospseudomonas* and *Tardiphaga* are presented in one colour, as there was no clear separation between them in the tree. The *Nitrobacter* are separated into early (light green) and late-branching (dark green) *Nitrobacter*.

Figure 5 shows the phylogeny of *Nitrobacter* and the 5 closest related genera within the *Xanthobacteraceae*. The earliest branching clade comprises the *Afipia*, the majority of which (19 out of 33, 58%) possess a copy of narG, as indicated by the grey squares on the outside of the tree. In the next clade named *62-47* narG is already less abundant (3 out of 9, 33%) and in the clade after that, which comprises genomes from the *Rhodopseudomona* and *Tardiphaga*, narG is largely absent (3 out of 37, 8%). There was no clear separation between the *Rhodopseudomona* and *Tardiphaga* in the tree, which is why they are represented by the same colour. In the biggest and, apart

from the *Nitrobacter*, the latest branching clade - the Bradorhyzobia - narG is slightly more abundant (32 out of 319, 10%). The *Nitrobacter* clade was coloured with two shades of green, light for the early branching *Nitrobacter* and dark for the later branching *Nitrobacter*. We wanted to distinguish the two, because NarG sequences from the darker *Nitrobacter* clade branch differently from NarG sequences stemming from organisms in the lighter-coloured clade, as can be seen in Figure 6.

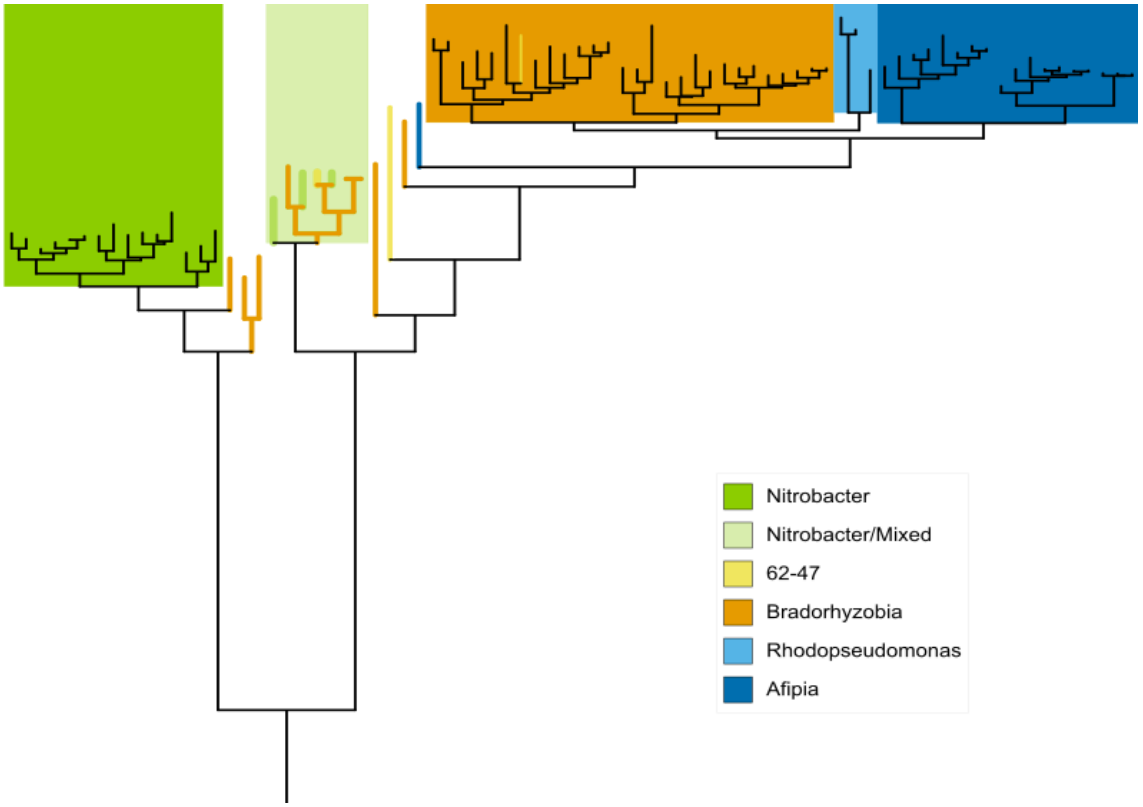


Figure 6. Phylogenetic tree of NarG sequences of *Nitrobacter* and their 5 closest related genera.

Clades are highlighted according to the genera that sequences stem from. The three NarG-sequences of the genus *62-47* are distributed in different parts of the tree, the corresponding branches are coloured yellow. The clade highlighted in light green contains sequences from *Nitrobacter*, *Bradorhyzobia* and *62-47*. Although most sequences in this clade stem from the *Bradorhyzobia* the clade is highlighted in light green, as it contains all three of the early branching Nitrobacter sequences, that are highlighted in Figure 5. Branches in this clade were coloured to reflect their origin (see legend). Some sequences did not branch to form a clade with other sequences, these branches are highlighted according to their origin as well.

The phylogeny of NarG in Figure 6 largely supports that it was vertically inherited in the *Xanthobacteraceae*. The majority of sequences seem to be closely related and are clustered in one part on the right side of the tree, which contains sequences from the

Afipia, *Rhodopseudomonas* and *Bradorhyzobia*. Interestingly, the sequences belonging to *Nitrobacter* are split into two parts. The sequences belonging to earlier branching *Nitrobacter* that were highlighted in light green in the species phylogeny (Figure 5) are contained in a clade of sequences with mixed origins. This clade is highlighted in light green and branches closer to the *Afipia*, *Rhodopseudomonas* and *Bradorhyzobia* than to the remaining *Nitrobacter*. The darker green clade of *Nitrobacter* sits on a long branch on the left side of the tree, indicating a horizontal gene transfer event. In contrast to the *Nitrococcales*, where the NarG sequences of the nitrifying Nitroccus were branching far away from all other sequences, some sequences of the *Xanthobacteracea* do branch closely to the NarG of *Nitrobacter*. The genomes that these sequences belong to were not considered for the pangenomic analysis. This was done to ensure that the outgroup only contained genomes of non-nitrifying organisms.
Otherwise, the *Xanthobacteraceae* depicted in Figure 5 are an excellent outgroup for the next steps, the nitrifying ability was likely acquired very recently in the evolutionary history of *Nitrobacter*.

3.2 - Whole genome comparison

We performed pangenomic analyses to identify genes that were horizontally acquired in and are exclusive to the nitrifying genera *Nitrococcus* and *Nitrobacter.* We hypothesise that these exclusive genes are more informative than genes that are shared with the non-nitrifying relatives.
The initial step was to assess the parameters for running the pangenome pipeline used in the project, Roary. As discussed in section 2.4.2, two key variables influence the resulting pangenomes: the MCL inflation value and the identity threshold. Since altering the inflation value had minimal impact on the results, we concentrated on the identity threshold.

3.2.1 - Threshold evaluation

The identity threshold is the percentage of identical positions that two protein sequences need to share to be connected in the network that determines which sequences belong to the same cluster (See section 2.4.2). If the threshold is too high, many proteins that were vertically inherited will not end up in the same cluster and the vast majority of genes will be assigned as being present in only one genome. If it is too low, the opposite will happen and genes that have different evolutionary origins and were horizontally acquired will end up in the same cluster.

Roary was run at different identity thresholds to determine an appropriate value for each group. The best threshold should capture the maximum amount of exclusive genes in the ingroup while stabilising the amount of core genes in the outgroup.

Identity	Core *Xanthobacteraceae*	Core *Nitrobacter*	Exclusive genes
60	1100	1450	103
70	940	1460	134
75	824	1458	167
80	571	1407	230
90	67	774	290

Table 2. Summary of roary results for the group of *Xanthobacteraceae* at different identity thresholds.

Genes were considered as belonging to the core set if they were present in at least 80 % of genomes. Exclusive genes are genes that are present in both closed genomes of *N. Winogradskyi* and *N. Hamburgensis*, at least one of the other 5 *Nitrobacter* genomes and not in any of the other *Xanthobacteraceae*.

Table 2 shows a summary of the roary results for the *Xanthobacteraceae* at different identity thresholds.

As expected, the number of core genes generally increases when lowering the identity threshold and the number of exclusive genes decreases. For the core genome of *Nitrobacter,* a plateau is hit at a threshold of 75% identity. Similarly, the core genome of the *Xanthobacteraceae* does not increase by as much after lowering beyond 75% identity as before. Therefore, the 167 exclusive genes that were found at the 75% identity threshold were chosen as the set to work with in further analysis.

Table 3 provides a summary of the Roary results for the *Nitrococcaceae*. The number of exclusive genes in the genus *Nitrococcus* is significantly higher compared to *Nitrobacter* (see Table 2). This difference is due to the smaller number of genomes in the outgroup. Additionally, the *Nitrococcaceae* appear to be less closely related than the *Xanthobacteraceae*, as they share fewer core genes at the same identity thresholds. This observation aligns with expectations since the *Nitrococcaceae* represent an entire family, while only the genera closely related to *Nitrobacter* were included from the *Xanthobacteraceae* family.

Identity	Core *Nitrococcaceae*	Core *Nitrococcus*	Exclusive genes
60	476	2320	1056
70	235	2261	1536
80	61	2152	1913
90	6	1530	1494

Table 3. Summary of roary results for the *Nitrococcaceae* at different identity thresholds. Core genes are present in 80% or more of the genomes. Exclusive genes are present in 2 out of 3 *Nitrobacter* genomes and not in any of the other genomes.

Interestingly, the number of exclusive genes in the genus *Nitrococcus* at 80% is higher than at 90%. This can be explained by the large number of genes that are only present in one genome if genes are clustered at 90% identity. As mentioned in section 2.4.2, genes that are only present in one genome were not included in the exclusive set. When lowering the identity threshold from 90 to 80%, some of the genes that were previously classified as being present in only one genome are now shared by multiple genomes, leading to the observed increase in exclusive genes.

After considering phylogenetic and pangenomic results from both *Nitrococcus* and *Nitrobacter*, we decided to focus the analysis on the group of *Nitrobacter*. The availability of two closed genomes within *Nitrobacter* allowed us to be more strict in the inclusion of genes into the exclusive set. Additionally, the greater set of genomes for the outgroup and the knowledge that the NarG associated with nitrifying abilities was acquired relatively recently in the evolution of *Nitrobacter* led to a more refined set of exclusive genes that was more accessible to interpretation in a biological context.

3.2.2 - Pangenome Matrix

Figure 7 shows the binary presence and absence matrix of all genes in the pangenome of the *Xanthobactereaceae* at identity level 75. The presence of a gene is marked in blue, absence is marked with white. A tree based on presence/absence distances between the genomes is shown to the left of the figure.
The matrix is very sparse and dominated by genes that are only present in one or two genomes. This is because genomes from different genera were considered for this pangenome, typically pangenomes are used to analyse species-level differences,

leading to many more shared genes. However, the seven *Nitrobacter* genomes belonging to potential nitrifiers all cluster together in one clade denoted in green. This indicates that there is a part of the genome that is shared between the nitrifying *Nitrobacter* that sets them apart from other *Xanthobacteraceae.*

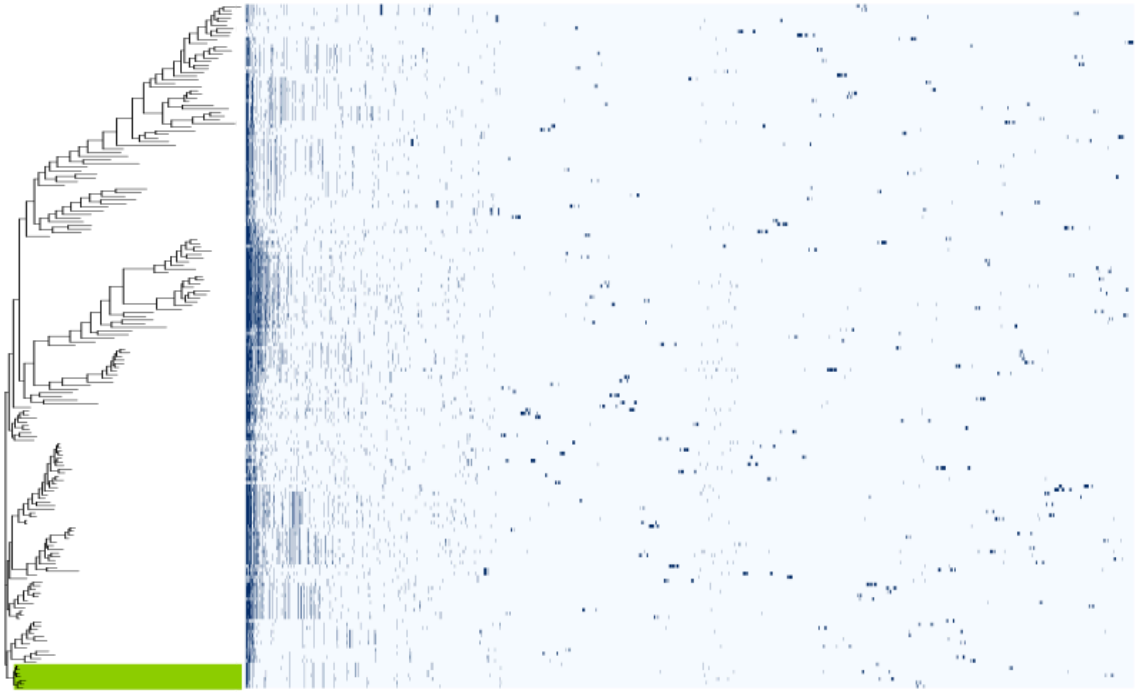


Figure 7. Pangenome matrix and gene presence/absence based tree of *Xanthobacteraceae.*

The clade highlighted in green contains only and all *Nitrobacter* genomes that are thought to be capable of nitrite oxidation.

3.2.3 - Nitrobacter exclusive genes.

After obtaining a set of exclusive genes for the nitrifying members of *Nitrobacter*, we wanted to get more insights into the funcion of these genes. Therefore we expanded on the initial annotations with Prokka, which were done as part of the roary pangenome pipeline, with another round of annotations using EggNOG. EggNOG has access to a much bigger database and is therefore more accurate but slower than Prokka (see sections 2.4.1 and 2.5.1). The amount of unannotated exclusive genes could be reduced from 120 with Prokka to 48 using Prokka and EggNOG annotations.

To gain a better understanding of the exclusive genes their distribution along the *Nitrobacter* genomes was plotted, reordering them in the order that they appeared in

each genome. This provided a framework to get a sense of the overall conservation and distribution. We chose the genome of *N. winogradskyi* to be the representative genome for plotting the distribution and the gene cluster comparisons that followed. It is one of the two available complete genomes of *Nitrobacter* and, in contrast to the genome of *N. hamburgensis*, it contains no plasmids.

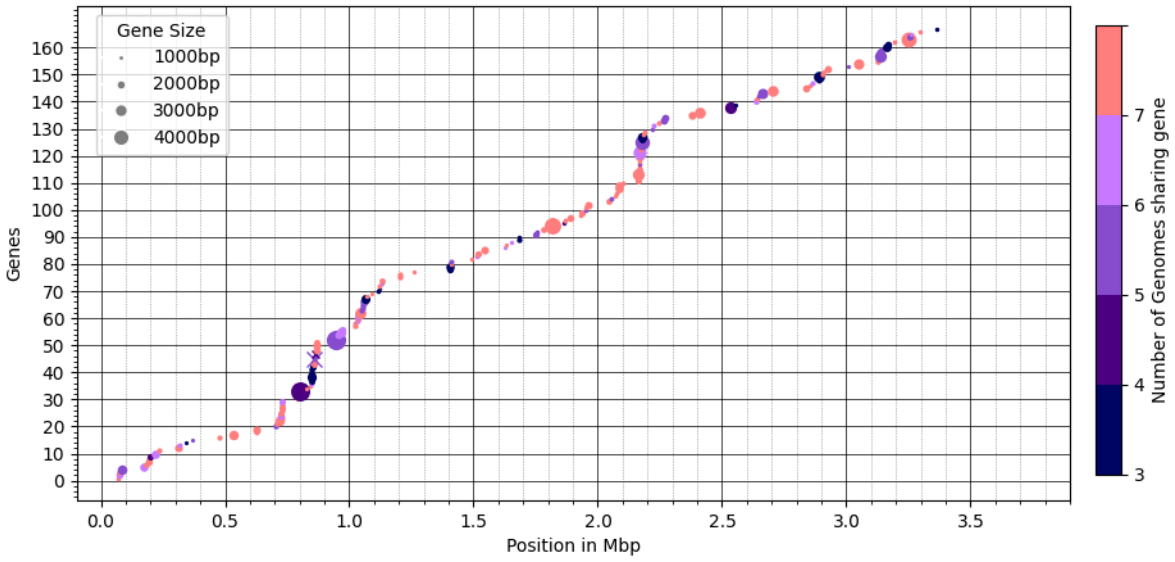


Figure 8. Distribution of exclusive genes in *N. Winogradskyi*

Exclusive genes are marked with a dot, sizing indicates the size of the gene and colouring indicates the number of total *Nitrobacter* genomes each gene is found in. The exclusive genes are counted on the y-axis and the x-axis shows the position within the genome.

Figure 8 shows the distribution of exclusive genes along the genome of *N. winogradskyi*, the representative genome. The y-axis represents the cumulative number of genes, the x-axis shows their positions within the genome. Throughout the genome, there is a steady increase in the number of exclusive genes, indicating an even distribution of exclusive genes in the majority of the genome. In some regions, the number of exclusive genes rises more sharply, these regions contain clusters of exclusive genes. Notably, between 0.7 and 1.1 Mbp, there are multiple such hotspots in close proximity, and another significant hotspot is located around 2.1 to 2.2 Mbp.

3.3 - Gene cluster comparison

3.3.1 - Cluster Alignment

Initial analysis of exclusive gene distribution, as shown in Figure 8, revealed - they are unevenly spread throughout the genome. We wanted to investigate hotspots in the genome, where exclusive tended to cluster more closely. For this purpose, a Python package (see section 2.5.2) was written, where clusters of exclusive genes (CEGs), could be identified and then aligned and visualised with Clinker. A CEG was defined as five or more exclusive genes where each of them was assigned within a range of five genes of the nearest exclusive gene in the representative genome. This resulted in five CEGs, three of which are shown in this section.

Figure 9 shows a CEG containing genes coding for all three subunits of cytochrome c oxidase (COX), protoheme IX farnesyltransferase, CtaG and several uncharacterized proteins. Cytochrome c oxidase is a protein complex that serves as a terminal oxidase in both bacteria and mitochondria. Protoheme IX farnesyltransferase is involved in the heme synthesis and CtaG plays a role in the maturation of COX subunit 3. With the exception of one gene of unknown function (second from the left), all genes of this cluster are in the exclusive set for *Nitrobacter*. It should also be noted that subunit 3 starts less than 7000 bp away from the start of the nar-operon in both of the complete *Nitrobacter* genomes, with 6 non-exclusive genes encoded between them.

BLAST search against *Nitrococcus* genomes revealed they also possess a copy of all three COX subunits, with sequence identities to *Nitrobacter* COX between 35 and 55%. The three subunits are in the exclusive set of *Nitrococcus* for identity levels above 60%, but not in the set at 60% identity. Overall, *Nitrococcus* and *Nitrobacter* don’t appear to have COX with the same evolutionary background. While the COX of *Nitrobacter* was likely horizontally acquired, it is unclear if *Nitrococcus* COX was vertically inherited or not.

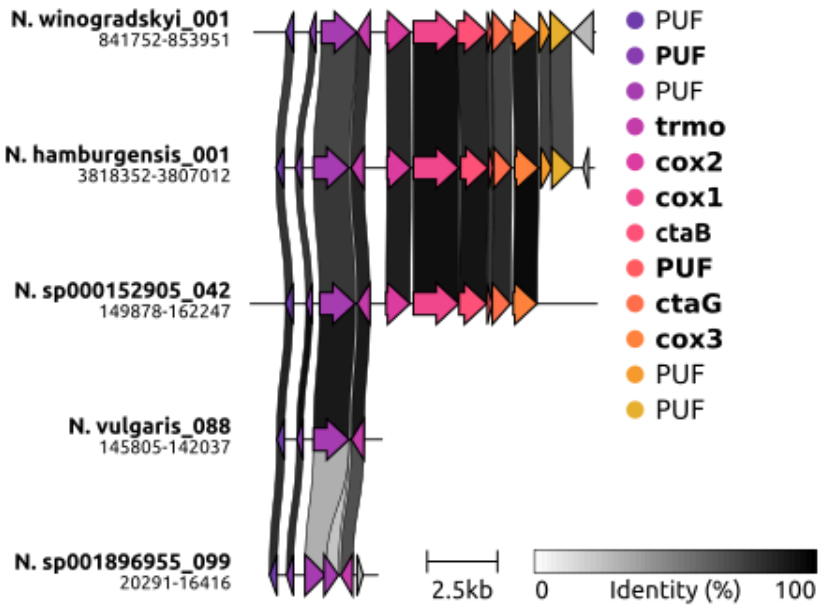


Figure 9. Visualisation of the alignment of CEG containing COX-operon.

Gene names are shown to the right, when no function could be identified genes were annotated as coding for Proteins of unknown function (PUFs). Exclusive genes are written in bold. This cluster contains genes coding for the three COX subunits and genes involved in the assembly of COX

The CEG in Figure 10 contains genes coding for proteins related to the assembly of a carboxysome and RubisCO. RubisCO is an extremely abundant protein complex that is responsible for carbon fixation. The carboxysome is a bacterial microcompartment, in which RubisCO is located. This operon is also present in the exclusive sets of *Nitrococcus*. BLAST results show that *Nitrococcus* and *Nitrobacter* have remarkably high levels of sequence identities around 90% in the RuBisCO operon.

RuBisCO is necessary to enable the autotrophic lifestyle of both *Nitrobacter* and *Nitrococcus*, but due to its prevalence in many organims, it is not a good marker for identifying nitrite-oxidizing bacteria. Nevertheless, the presence of RubisCO supports the validity of the approach, as its presence is expected in autotrophic *Nitrobacter* and *Nitrococcus*, but not in their heterotrophic relatives.

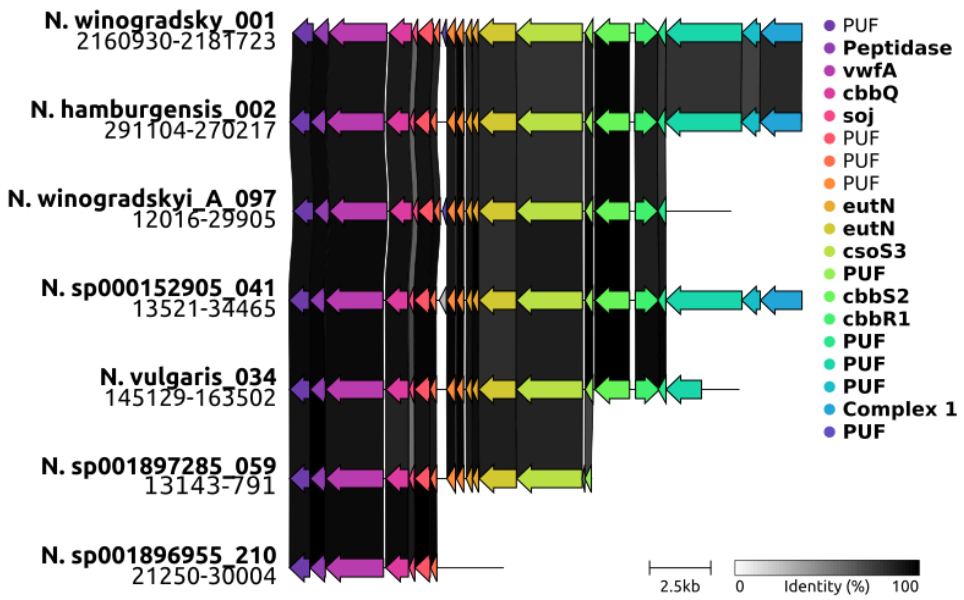


Figure 10. Visualisation of the alignment of CEG containing RuBisCO operon.
Gene names are shown to the right, when no function could be identified genes were annotated as coding for Proteins of unknown function (PUFs). Some genes were not annotated with names, but a function was proposed. In these cases, the proposed function of the protein is given. Exclusive genes are written in bold. This cluster contains genes coding for RuBisCO subunits, proteins responsible for RuBisCO assembly, and parts of the carboxysome.

The CEG in Figure 11 contains the nar-operon. The operon comprises the three subunits of the narGHI complex, as well as narJ which is needed for the assembly of the complex (Dubourdieu and DeMoss 1992; Vergnes et al. 2006). Other genes are annotated as coding for a nitrate transporter, an anion channel, an Isomerase and cytochrome c. The anion channel is not found in *Nitrococcus*, and the nitrate transporter is located outside of the operon. The other elements of the operon are conserved in the same operon with levels of sequence identity between the *Nitrococcus* and *Nitrobacter* proteins ranging between 40 and 70%.

The presence of an isomerase in the operon of both *Nitrococcus* and *Nitrobacter* is potentially significant. These isomerases are involved in protein folding, however, whether or not it is involved in the maturation of the narGHI complex is difficult to verify computationally. The role of cytochrome c in the operon may be more interesting for this project. Cytochromes are electron transporters, and the presence of a cytochrome in the nar-operon implies that it may be needed as an electron acceptor for the electron

generated by nitrite oxidation.

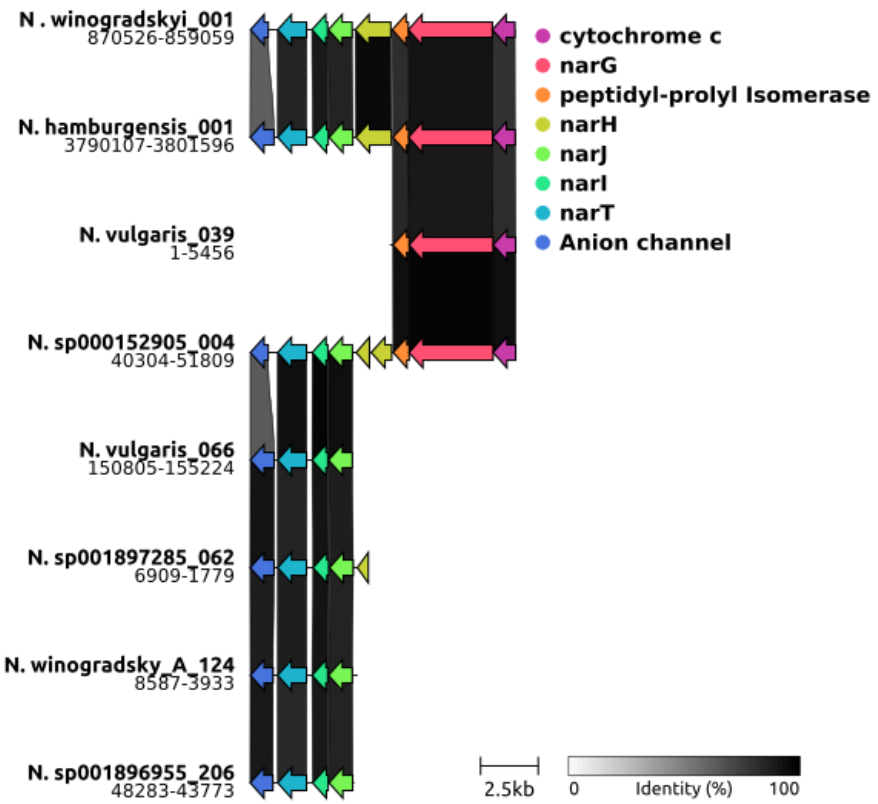


Figure 11. Visualisation of the alignment of Nar operon containing CEG.
Gene annotations are shown to the right. some genes were not annotated with names, but a function was proposed. In these cases, the proposed function of the protein is given.

3.3.2 - Distribution of nar-operon associated cytochrome

Because of their potential role in the electron transfer chain in narG-mediated nitrite oxidation, we decided to investigate the distribution of nar-operon associated cytochromes (NarC) in the context of the phylogeny of narG. For this purpose, we used BLAST to search for sequences similar to the narC of the *N. winogradskyi* in our database and filtered for hits within a range of 5 genes of narG. This led to the result in Figure 12, where the presence of all such cytochromes is marked with blue on the outside of the tree.

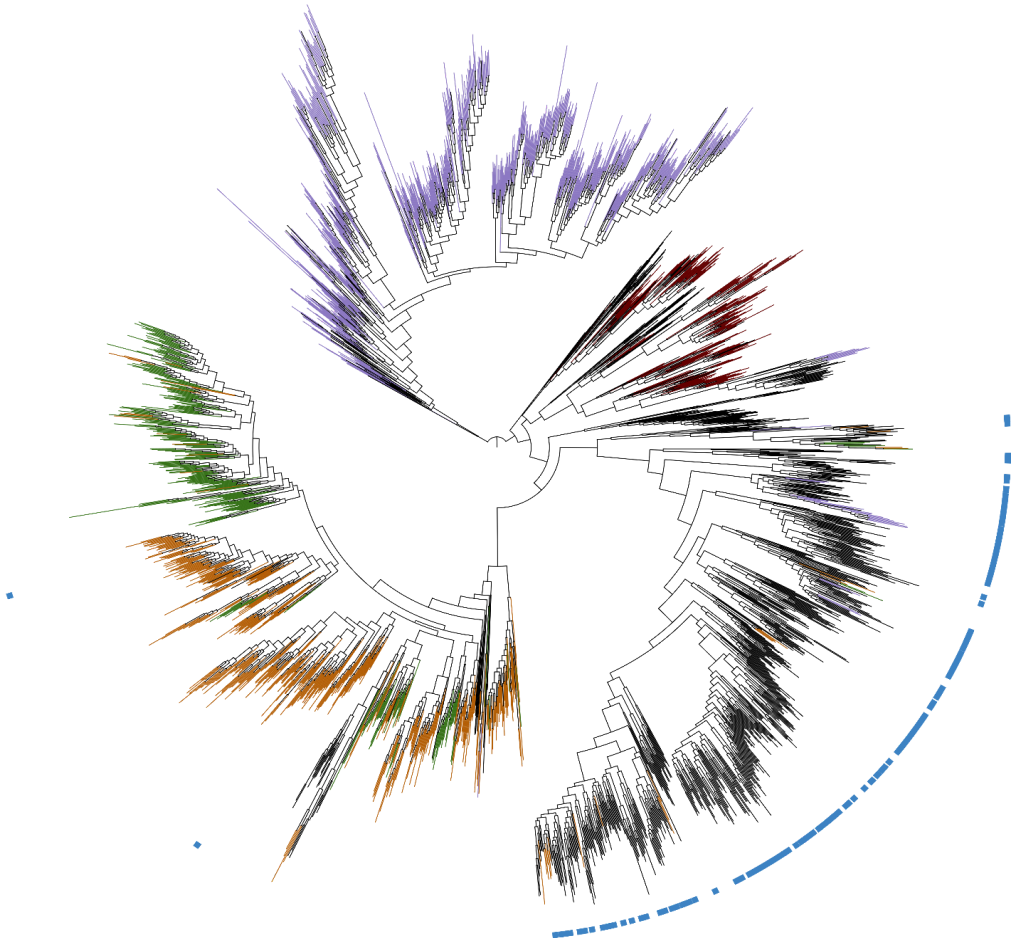


Figure 12. - Phylogenetic tree of NarG-sequences

Branches are colored according to the Phylum that the corresponding NarG sequence stems from: Actinomycetota (purple), Bacillota (red), Proteobacteria are divided in Alpha- (green) and Gammaproteobacteria (orange). The presence of narC within 5 genes of each narG sequence is indicated by blue markers on the outside of the tree.

Figure 12 shows the presence of NarC in the nar-operon is almost exclusive to the branch with an unclear evolutionary origin. The cytochrome appears to be ancestral to this clade that makes up 659 (or 26%) of the 2523 centroid sequences in this tree. Not all, but the vast majority of the sequences in this clade appear to be associated with the cytochrome. Such strong conservation points to a relevant function that the cytochrome must have in the operon. Therefore, we decided to search for potential interactions between cytochrome and NarGHI complex.

3.4 - Structure predictions

When we began working on structure predictions, we had limited knowledge of AlphaFold and were unsure of the time an resources our predictions would require and how accessible to interpretation they would be. To determine what could be achieved within the scope of the project, we adopted a progressive strategy, starting with monomer predictions, then moving on to multimer, and finally to whole complex predictions. Over the course of the project AlphaFold3 was made available and predictions became much faster and easier to achieve. This made it possible to not only predict whole protein complexes, but also interactions between them as well as the integration of selected ligands into the predictions.

3.4.1 - AlphaFold2 predictions

3.4.1.1 - cytochrome monomers

To evaluate whether making structure predictions with custom Multiple sequence alignments (MSAs) would influence results, we predicted the structure of the Nar-operon associated cytochrome (NarC) with ColabFold and AlphaFold2.3.1 (See section 2.6.2). ColabFold allows the user to provide a MSA in a3m format, whereas AlphaFold2 requires multiple inputs that are more complicated to generate. We performed one prediction with Colabfold and a custom MSA with input sequences from our database and one prediction with AlphaFold2, without a custom MSA, to compare the two methods, most notably to assess whether a curated sequence set would improve the structure prediction.

Figure 13 shows the top-ranked prediction of the *N. winogradskyi* NarC of both methods (a) and the custom-MSA coverage plot (b). The coverage plot reveals that the N-terminal globular domain is not present in about 50 of the 760 sequences. This can be seen by the uncoloured regions in positions ~50 to ~150 at the bottom of the graph. The sequences that lack the globular domain also have low sequence identity below 20% to the *N. winogradskyi* NarC for the rest of the protein. Otherwise, the identity levels range between 20 and 60 %

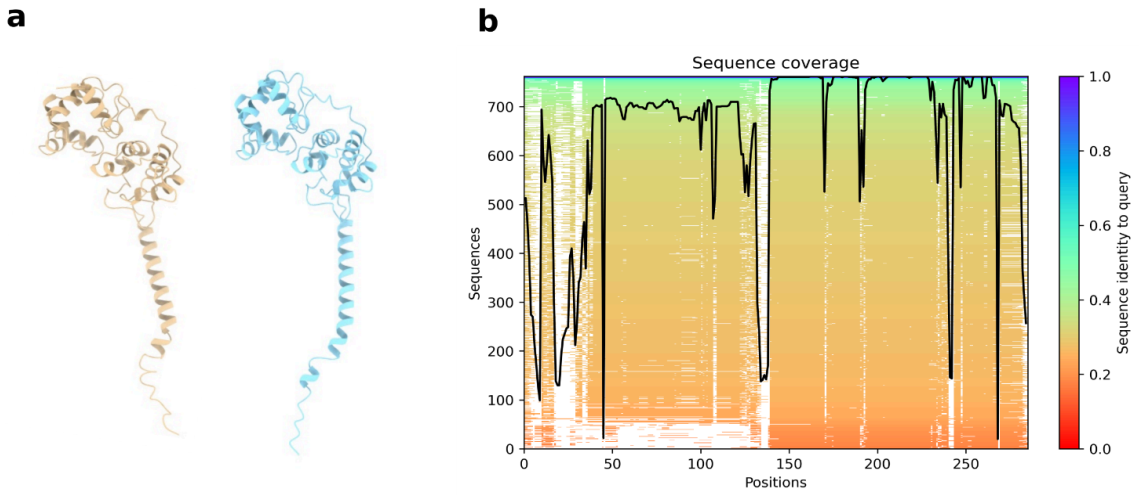


Figure 13. Structure prediction of cytochrome monomers with and without custom MSA. Panel a) shows the predicted structures, the prediction with custom MSA is coloured blue (right). Panel b) shows the sequence coverage of the cytochrome sequence in the custom MSA. Sequence identity levels are indicated by colour. There are approximately 100 sequences that share over 50% identity with the cytochrome, as indicated by the green colouring at the top of the plot.

The predictions in Panel a) show what appears to be a transmembrane helix at the c-terminus, preceded by two globular domains connected with a flexible linker. The structural predictions are oriented in such a way that the first globular domains of the two visualisations are aligned. There are two slight differences between the two predictions, firstly the angle of the transmembrane helix differs between the two predictions and secondly, there is a small helix that is predicted after the main transmembrane helix that is only present in the custom-MSA prediction. The two globular domains are almost identical, the same number of helices are predicted in the same positions for both. Overall we decided against using the custom MSAs for further complex structure predictions. The predictions without custom MSAs yielded similar results, and the workflow was faster and more reliable when using the automatically generated MSAs.

3.4.1.2 - NarIC multimer

After the successful prediction of monomer structures, we performed multimer predictions to check for interactions between the NarGHI complex and NarC. We started with the simplest case of NarI and NarC since these are two subunits that are expected to interact, with NarC on the periplasmic side of the membrane and NarI as

the membrane spanning subunit, whereas NarG and NarH are located on the cytoplasmic side (Simon and Klotz 2013).

Figure 14 shows the superposition of the top five ranked multimer predictions of NarIC by AlphaFold 2.3.1. The predictions show NarC sitting on top of NarI and the presumed transmembrane membrane running in parallel to NarI. One of the two globular domains of NarC interacts with NarI, the other one shows no interaction and is connected via a flexible linker. All five predictions show the helices in very similar positions, only the C-terminal unstructured tail of NarC varies in its position. The C-terminal Helix that was previously only predicted by Colabfold (Figure 13) is also predicted by AlphaFold2 when using the multimer model.

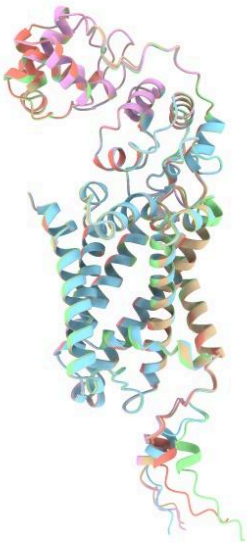


Figure 14. Superposition of top 5 ranked structural predictions of NarI and NarC.
Each prediction is coloured differently so that their relative positions can be assessed. Only the C-terminal tail of NarC (bottom right) varies significantly between the predictions.

The ipTM+pTM ([interface] predicted template modelling) scores range between .844 and .836 for the shown structures, indicating that AlphaFold has high confidence in these predictions. ipTM + pTM scores incorporate AlphaFold's confidence in the position of the two subunits relative to one another, as well as its confidence in each position of the complex as a whole.

3.4.2 - AlphaFold3 predictions

With the emergence of AlphaFold3 - specifically, the AlphaFold web-server - multimer predictions became considerably faster and easily acessible. This allowed for the

prediction of larger complexes, and the addition of a selected set of Ligands, such as hemes, to the structures. We proceeded with our approach of predicting more and more complex interactions, starting with the prediction of NarGHIC.

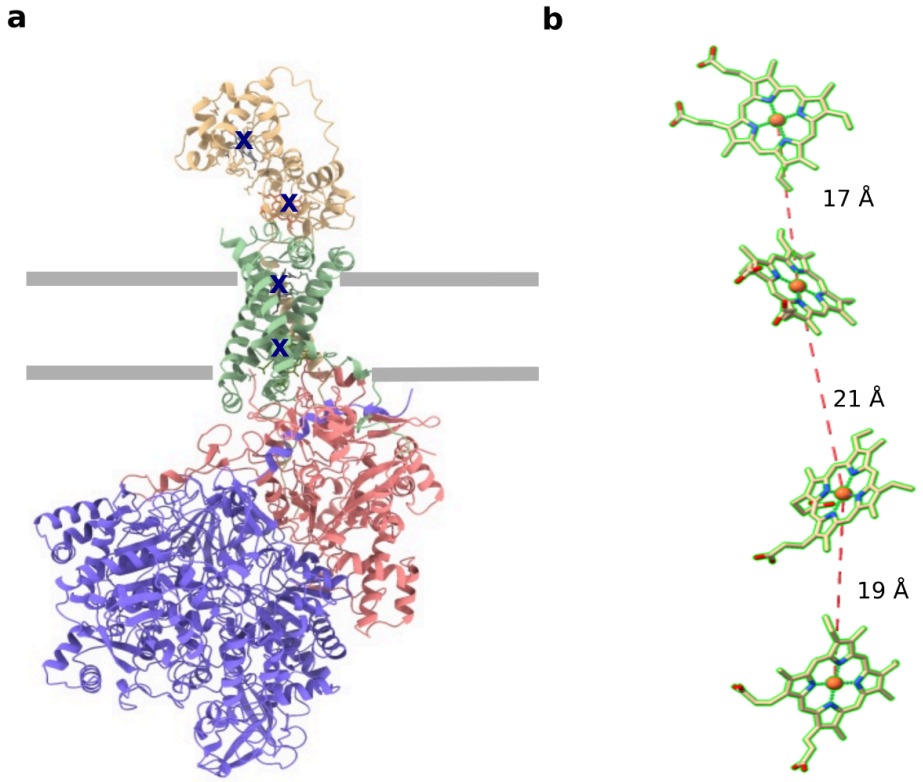


Figure 15. Structure predictions of NarGHIC with hemes.

Panel a) Shows the predicted structure. The blue X indicates the heme-positions. The membrane plane is indicated in grey with the periplasmic side on top and the cyotplasmic side below. Subunits can be distinguished by colour: NarG - purple; NarH - red; NarI - green; NarC - yellow

Panel b) shows the hemes without the surrounding structure. Centre to centre distances are drawn in red and the distance is given in Angstrom.

Figure 15 shows the top-ranked prediction by AlphaFold3 of the NarGHI + NarC complex, including four hemes - two heme c's and two regular hemes. The position of the hemes is shown to the right of the structural prediction. Other predictions show NarC and NarI in very similar positions with ipTM and pTM values between .81 and .84. The superposition of all structures can be found in the supplementary materials.

NarC is located on the periplasmic side of the membrane spanning subunit NarI, which is connected to NarGH on the cytoplasmic side of the membrane. The placement of the

hemes aligns with expectation; NarI is known to have two heme-binding domains, (Rothery et al. 1999) and the cytochrome contains two heme-c binding motifs, one in each globular domain. To see if electron transfer would be possible in the predicted complex, the centre-to-centre distances between all hemes were measured. The distances between the two NarC-heme c's was measured at 17 Å the distance between the NarI hemes is at 19 Å and the distance between the NarI and NarC heme is at 21 Å. Literature indicates that electron transfer is possible below a distance of 25 Å, meaning the predicted heme arrangement would allow for electron transfer within the complex (Cahen, Pecht, and Sheves 2021; Gray and Winkler 2005).

Figure 16 shows the expected position error (EPE) of each residue in relation to all other residues. Positions 1 through 1188 belong to narG, then follows the narH subunit which ends at 1782, NarI ranges until 2079 and the cytochrome follows after that. The ligands/hemes can be seen as four squares at the end of the residues at the bottom right.

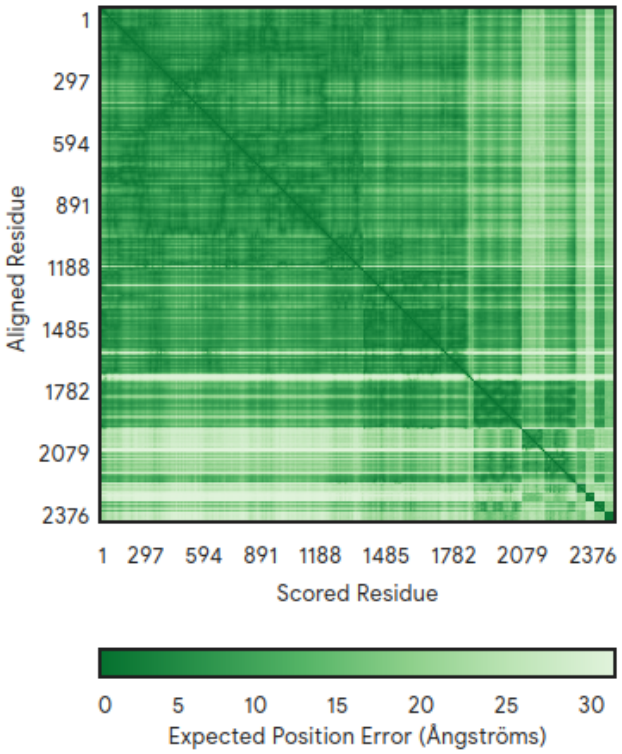


Figure 16. Expected position Error Matrix of the top-ranked NarGHIC prediction.
Darker colour indicates higher confidence in the relative positions. NarG ranges from 0 to 1214; NarH from 1215 to 1728, NarI from 1729 to 1954; NarC from 1955 to 2208; the four Hemes are at the end.

The position of NarG and NarH is very confidently predicted as indicated by the dark

green coloring of the corresponding area. The position of NarI in relation to narGH is slightly less confidently predicted. This may be related to the assembly process of the complex, where NarG and NarH are assembled first and later attached to the NarI subunit (Lanciano et al. 2007). Their interaction may be stronger, whereas the NarI subunit is integrated in the membrane and does not interact as much with the other subunits, therefore making the signal harder to detect. The position of NarC is less confidently predicted than other parts of the complex, with the position of NarI and NarC being slightly more confidently predicted than the position of NarC in relation to NarGH. This is to be expected, as NarC is only interacting with NarI directly. Additionally,

Overall the AlphaFold3 predictions show a strong indication of the formation of a 4 subunit complex for nitrite oxidation.

Potential interaction partners

We searched for potential interaction partners to the NarGHIC complex in the genome of *N. winogradskyi.* Identifying the interaction partner that accepts the electron from NarGHIC would help explain why the complex physiologically performs nitrite oxidation rather than nitrate reduction. Therefore we searched the genome for all proteins containing a heme-binding motif (CxxCH), which yielded 21 potential interaction partners (PIPs). Four PIPs were part of the exclusive set and after searching the literature, we found that one of the four exclusive PIPs was previously described as leading to an increase in nitrite consumption when added to an artificial system containing the nitrate reductase complex and cytochrome c oxidase (COX) extracted from *N. winogradskyi* (Nomoto, Fukumori, and Yamanaka 1993). It was hypothesised that this cytochrome acts as an electron shuttle between the Nar-complex and COX. From here on this cytochrome is referred to as **PIP 736**, after its accession number in *N. winogradskyi.* Additionally, we considered the possibility of direct transfer of the electron from NarGHIC to COX.

To assess the possibility of direct electron transfer from NarC to COX we performed structure predictions of NarGHIC + COX and NarC + COX. The reason we performed both predictions was that the presumed permanent interaction of NarGHI + NarC is likely stronger than the transient interaction of NarC + COX would be and the latter interaction might not be seen when predicting both complexes.

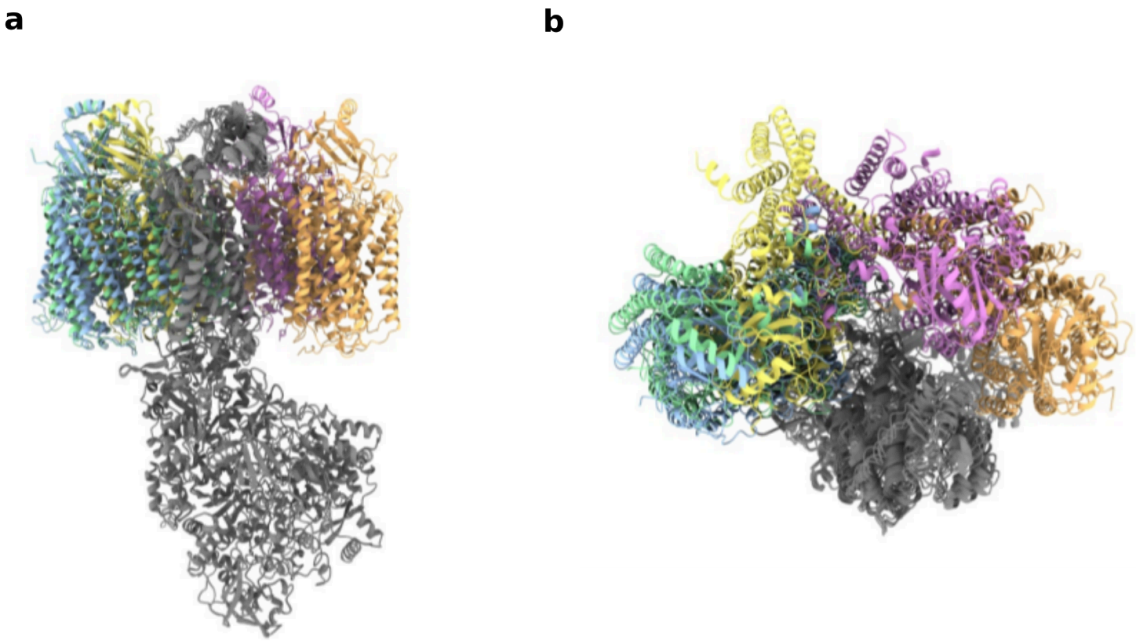


Figure 17. The superposition of the top 5 ranked predictions of NarGHIC and COX
NarGHIC is coloured grey, the 5 predictions of COX are coloured differently, so their relative positions can be seen. Panel a) shows the side view. NarGHIC is predicted stably in the middle, with different COX predictions to the left and right. Panel b) shows a view from above. NarGHIC is stably predicted at the bottom, the different COX predictions form a half-circle above.

Figure 17 shows the superposition of the top five predictions of NarGHIC +COX from two perspectives. NarGHIC is shown in grey, its position is stable throughout all five predictions. Panel a) shows the side view of the complex with the cytoplasmic side below and the periplasmic side above and Panel b) shows the view from the top. The position of COX varies between the different predictions and the COX predictions form a ring around the NarI subunit. AlphaFold3 does not predict a stable interaction between the two complexes.

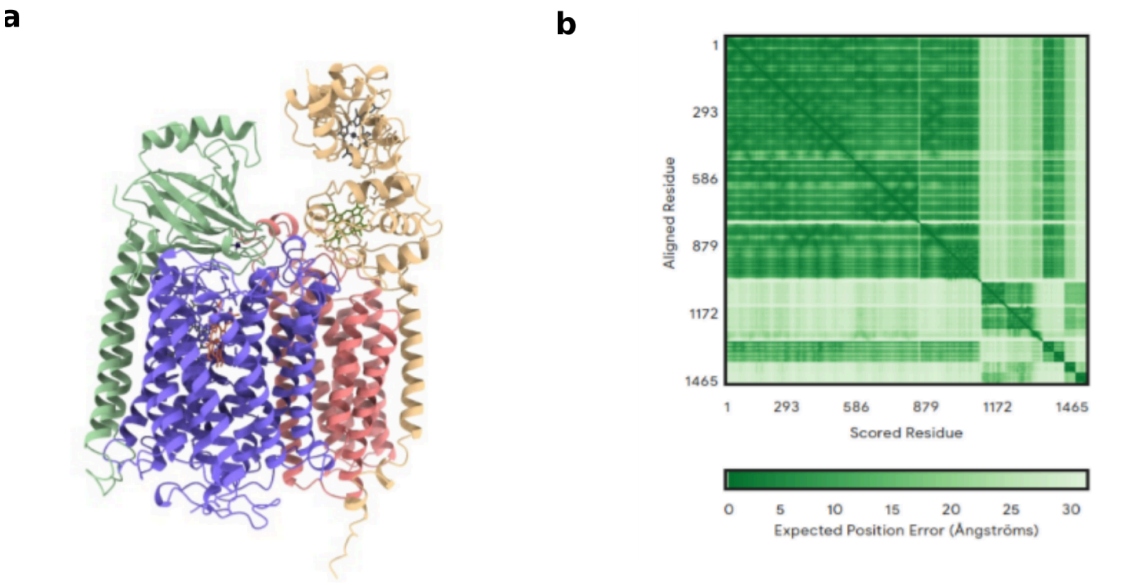


Figure 18. Structure prediction of COX and NarC

Panel a) shows the top-ranked predicted structure. NarC is shown in yellow, COX I in purple, COX II in green and COX III in red. Panel b) shows the corresponding EPE matrix. NarC ranges from 1040 to 1291. The COX subunits range from 0 to 1039 and Hemes are at the end.

Figure 18 shows the top-ranked prediction of NarC + COX with the corresponding EPE matrix. The cytochrome is shown in yellow, the three COX subunits are shown in purple green and red. COX II contains the electron acceptor site. This means that, if direct electron transfer from NarC to COX were to happen, the second globular domain of NarC and COX II are expected to interact. However, the structure in panel a) shows no signs of the two subunits interacting. This is also evident in the expected position error, where AlphaFold shows very low confidence in the position of NarC relative to COX. Only the transmembrane domain shows signs of interaction, this may be due to shared hydrophobic properties between the transmembrane domains of COX and NarC.

Next, we tested for interactions of the 21 PIPs in the genome of *N. winogradskyi* with NarGHIC and COX. Here, we focus on the results from PIP 736, it was the only PIP that showed promising AlphaFold results and it was previously demonstrated to have an impact on nitrite oxidation rates in an *in vitro* experiment (Nomoto, Fukumori, and Yamanaka 1993).

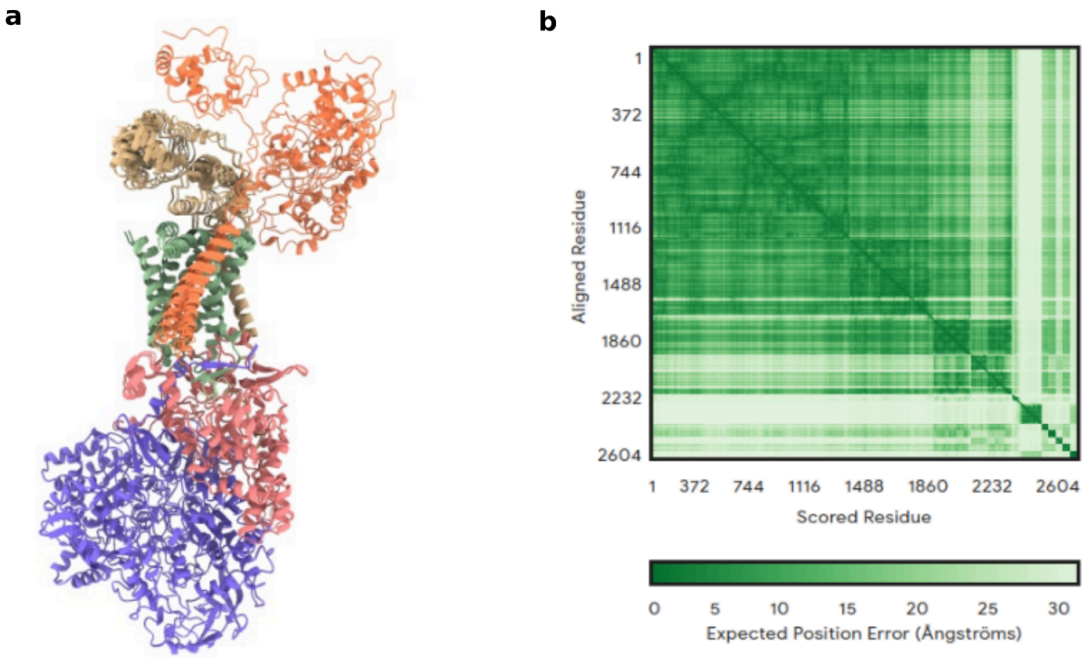


Figure 19.Structure predictions of NarGHIC and PIP 736.

Panel a) shows the superposition of the 5 predicted structures of NarGHIC and PIP 736. PIP 736 is shown in orange, its position varies between the predictions. NarGHIC are stably predicted. They are coloured purple, red, green and yellow respectively. Panel b) shows the EPE matrix of the top-ranked prediction. NarGHIC ranges until position 2208, PIP 736 follows after and ranges to 2333, the hemes are at the end.

Figure 19 shows the superposition of the top 5 ranked structure predictions of NarGHIC + PIP 736. The position of NarGHIC is stable throughout all 5 predictions, but the position of PIP 736 varies, showing no interaction with the second globular domain of NarC, where the electron transfer is expected to happen. PIP 736 has a long transmembrane domain which associates with NarI due to its hydrophobic properties. This domain is followed by a flexible linker that is much longer than the linker of NarC and one globular heme-binding domain. The position of the globular domain varies between the different predictions.

In Panel b) the EPE matrix is shown. NarGHIC ranges from positions 1-2208 and is predicted confidently as a complex. The position of PIP 736 relative to Nar is prone to error, as can be seen by the white markings in the corresponding area of the matrix

Figure 20 shows the superposition of the top 5 ranked structures of NarGHIC + PIP 736 + COX from two different perspectives. In contrast to previous predictions with NarGHIC + COX or NarGHIC + PIP 736, when predicting all three, the positions of all proteins stay stable throughout the predictions. In figure b) NarGHIC is shown right. COX is depicted on the right side with the COX II subunit in green on top, where

electron transfer is expected to happen. The transmembrane domain of PIP 736 is in between the two complexes and its globular domain can not be seen because it is hidden behind COX II. In figure a) the perspective from the opposite side is shown. NarGHIC is now to the right and COX is on the left. PIP 736 can now be seen as sitting in front of COX II, suggesting an interaction between the two proteins.

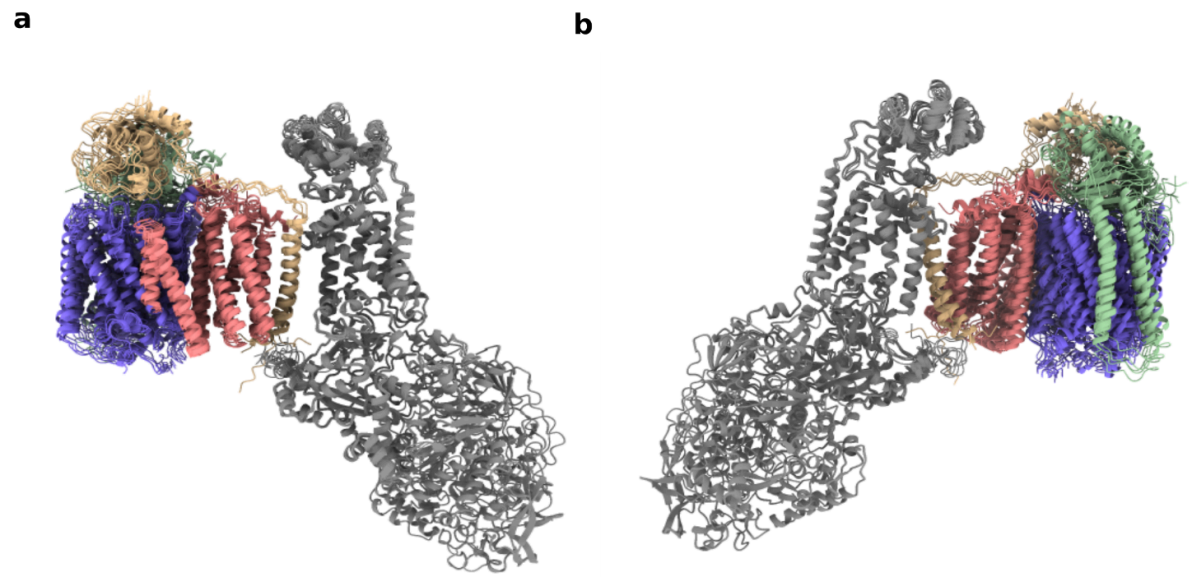


Figure 20. Superposition of the top five structure predictions of NarGHIC and COX and PIP 736

NarGHIC is shown in grey, PIP 736 in yellow, COX I in purple, COX II in green and COX III in red. Panel a) shows the front view of the predicted structure. NarGHIC is shown on the right, COX to the left. PIP 736 has a transmembrane domain between the two complexes and its heme-binding domain sits on top of COX. The view on COX II is obstructed by PIP 736. Panel b) shows the view from the other side. The heme binding domain of PIP 736 is now hidden behind COX II. All subunits are stably predicted in both figures.

Figure 21 shows a side view of COX and PIP 736 with visible ligands. This prediction is part of the supercomplex prediction in Figure 20, but NarGHIC was removed to showcase the interaction of PIP 736 and COX. In literature COX II is reported as having a dinuclear copper centre, from which electrons are transported a heme in COX I and then a heme-cu centre (Pitcher, Cheesman, and Watmough 2002). This is consistent with the predictions shown here. PIP 736 is shown in yellow, with one bound heme. It is not directly attached to COX II but is consistently predicted closely to it. Panel b) shows the position of the ligands without the surrounding proteins. Centre to centre distance measurements yield distances below 25 Å, where electron transfer is possible (Gray and Winkler 2005).

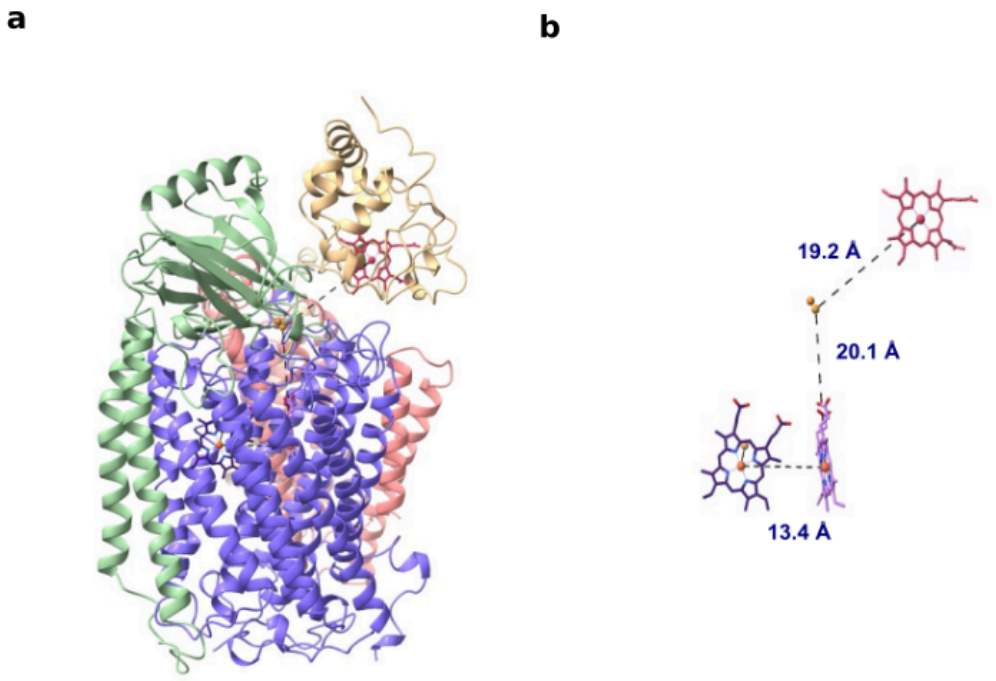


Figure 21. View on the interaction of COX and 736

The top-ranked prediction of NarHIC, COX and PIP 736 is shown. NargHIC is not shown in the figure. PIP 736 is yellow, COX I is purple, COX II is green and COX II is red. Panel a) shows the structure with hemes. Figure B shows only the hemes with centre-to-centre distances.

Overall, the structure predictions show strong indication of a 4 subunit NarGHIC complex in *Nitrobacter*. Additionally, the supercomplex prediction suggests that PIP 736 acts as an electron transporter between NarGHIC and COX, stabilising their interaction. Testing these hypotheses in a laboratory setup is an exciting direction for future work.

4 - Discussion

Throughout this project, we were successful in identifying properties common to organisms capable of NarG-mediated nitrite oxidation by the example of *Nitrococcus* and *Nitrobacter*. We made use of current bioinformatic tools to investigate the data at hand at different levels of resolution. Based on the constructed phylogenies we performed pangenomic analyses and found genes exclusive to the nitrifying *Nitrobacter* and *Nitrococcus*. To guide further analysis, we considered the distribution of these exclusive genes and further investigated those that were colocalized in clusters along the genome. When we then contextualised one of the identified exclusive genes - NarC - by overlaying its distribution on the NarG phylogeny, we found a strongly conserved signal in a major branch in the evolutionary history of NarG and therefore most likely the Nar-complex. We proceeded to perform structure predictions, our results strongly suggest the formation of a four-subunit complex. We then looked for potential interaction partners to transport electrons from NarGHIC to the terminal electron acceptor COX. We propose that the membrane-bound cytochrome we termed PIP 736 may stabilise an interaction between COX and NargHIC in *N. winogradskyi* and act as an electron shuttle between the two complexes, although further confirmation is needed.

One potential limitation of the analysis is that it may miss genes essential for nitrification if those genes are also found in some non-nitrifying relatives. These genes wouldn't be detected in the analysis because they weren't uniquely acquired by *Nitrobacter* or *Nitrococcus*, even if they were introduced through horizontal gene transfer before these species diverged evolutionarily. Still, we found the RuBisCO and Carboxysome operon to be exclusive to both nitrifiers. These genes were previously described as being horizontally acquired in *Nitrobacter* (Starkenburg et al. 2008) and therefore are expected to be part of the exclusive set. Although they are not suitable markers for nitrification, they are essential in supporting the autotrophic lifestyle of *Nitrococcus* and *Nitrobacter* (Starkenburg et al. 2008) and finding them as part of the exclusive set of genes supports the approach. Another drawback is the uncertain quality of genomes. Although each MAG had to meet quality standards to be included in the database, MAGs are inherently approximations of the true genome (Setubal 2021). In some instances, we detected signs of hybrid assembly, leading to the exclusion of those MAGs. However, if other cases of hybrid assembly were overlooked, the results could be affected. That said, only hybrid assemblies where genes exclusive

to *Nitrobacter* or *Nitrococcus* were incorrectly assigned to a related species would influence the outcomes. Overall there is a limited number of genomes of organisms performing NarG-mediated nitrite oxidation. Especially in the case of *Nitrococcus,* the limited availability of genomes led to a high proportion of genes being assigned as exclusive, including a lot of uninformative genes.

When it comes to structure prediction, it is unclear how well transient interactions and formation of supercomplexes can be predicted by the new AlphaFold3 software (Wee and Wei 2024; Abramson et al. 2024). For instance, we were unsuccessful in showing interaction between NarC and any of the PIPs or COX. This is likely due to the limitations of AlphaFold in predicting transient interaction, since there must be a way of electron transport from NarGHIC to COX and COX can only receive electrons from a cytochrome c. Recent work has also shown that such respiratory supercomplexes may have multiple stable states, which makes the interpretation of our results more difficult (Zheng et al. 2024).

The existence of a four-subunit complex in *Nitrobacter* has been proposed before (Simon and Klotz 2013), and our findings align with this idea. Tanaka et al. also show results that we think are consistent with a four subunit complex, though their interpretation differed at the time (Tanaka, Fukumori, and Yamanaka 1983). Evidence now strongly suggests that this four-subunit complex is far more widespread than previously thought. Although this discovery does not directly help in predicting Nar-mediated nitrite oxidation, it is a novel and intriguing finding. In organisms using NarGHIC to facilitate nitrate reduction, the NarC subunit may serve as an alternative pathway for electron transport, as previously suggested in *T. thermophilus*, where electron transport can shift to the nitrite reductase Nir under low nitrate conditions by forming a respiratory supercomplex (Cava, Zafra, and Berenguer 2008). In *T. thermophilus*, the Nar operon has been described as “integrated within highly variable regions of easily transferable extrachromosomal elements” (Alvarez et al. 2014: p19). This could explain the limited vertical inheritance of NarGHIC. In *N. winogradskyi*, the Nar operon was transferred into the chromosome because it is essential for survival, whereas in *T. thermophilus*, it remains an accessory gene that may not always be necessary.

Previous research identified a membrane-bound cytochrome as the electron transporter, which we believe is PIP 736 based on sequence similarity (Nomoto, Fukumori, and Yamanaka 1993). Structure predictions suggest that PIP 736 could facilitate the formation of a Nar/COX supercomplex. Interestingly, the COX enzymes in *Nitrococcus* and *Nitrobacter* share only 60% identity, indicating some variability in their

interaction. This suggests that NarGHIC is highly versatile in how it interacts within these complexes.

To build on the existing results, the nitrifying *Chloroflexota* should be investigated as well. They are not as closely related as the nitrifying *Nitrococcus* or *Nitrobacter*. This could point to the fact that the *Chlorofleoxota* already have some of the genomic requirements for nitrification and the chemolithoautotrophic lifestyle, which led to such gain of function events multiple times in the evolutionary history of the *chloroflexota*. This would make them a harder group to investigate, nevertheless it should be tried. Another indicator may be the distribution of PIP 736, although as stated previously, NarGHIC seems very flexible in its interactions and other organisms may use a different cytochrome for electron transport.

Overall, we did not achieve our initial goal of predicting NarG-mediated nitrite oxidation and it is unclear whether there is enough data to confidently be able to do so. Even still, I consider the project to be a success. The widespread distribution of NarC is a novel finding and it appears to modify the Nar complex to be highly versatile in its interactions. In previous work, there were some misunderstandings about NarC and the membrane-bound cytochrome that acts as electron transfer (Nomoto, Fukumori, and Yamanaka 1993; Starkenburg et al. 2006; 2008; Simon and Klotz 2013). We found that these are in fact two distinct proteins. We used tools from different fields of bioinformatic research and I personally gained knowledge in Biology, but also regarding phylogenies and pangenomes as well as structural bioinformatics.

References

Abramson, Josh, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, et al. 2024. “Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3.” *Nature*, May, 1–8. https://doi.org/10.1038/s41586-024-07487-w.

Alvarez, Laura, Carlos Bricio, Alba Blesa, Aurelio Hidalgo, and José Berenguer. 2014. “Transferable Denitrification Capability of Thermus Thermophilus.” *Applied and Environmental Microbiology* 80 (1): 19–28. https://doi.org/10.1128/AEM.02594-13.

Anas, Muhammad, Fen Liao, Krishan K. Verma, Muhammad Aqeel Sarwar, Aamir Mahmood, Zhong-Liang Chen, Qiang Li, Xu-Peng Zeng, Yang Liu, and Yang-Rui Li. 2020. “Fate of Nitrogen in Agriculture and Environment: Agronomic, Eco-Physiological and Molecular Approaches to Improve Nitrogen Use Efficiency.” *Biological Research* 53 (1): 47. https://doi.org/10.1186/s40659-020-00312-4.

Apweiler, Rolf, Amos Bairoch, Cathy H. Wu, Winona C. Barker, Brigitte Boeckmann, Serenella Ferro, Elisabeth Gasteiger, et al. 2004. “UniProt: The Universal Protein Knowledgebase.” *Nucleic Acids Research* 32 (Database issue): D115–19. https://doi.org/10.1093/nar/gkh131.

Beijerinck. 1901. “Über oligonitrophile Mikroben,” 1901.

Bernhard, Anne. 2010. “The Nitrogen Cycle: Processes, Players, and Human Impact | Learn Science at Scitable.” *Nature Education Knowledge 3(10):25*. https://www.nature.com/scitable/knowledge/library/the-nitrogen-cycle-processes-players-and-human-15644632/.

Bertero, Michela G., Richard A. Rothery, Monica Palak, Cynthia Hou, Daniel Lim, Francis Blasco, Joel H. Weiner, and Natalie C. J. Strynadka. 2003. “Insights into the Respiratory Electron Transfer Pathway from the Structure of Nitrate Reductase A.” *Nature Structural & Molecular Biology* 10 (9): 681–87. https://doi.org/10.1038/nsb969.

Blasco, Francis, Chantal Iobbi, Gerard Giordano, Marc Chippaux, and Violaine Bonnefoy. 1989. “Nitrate Reductase of Escherichia Coli: Completion of the Nucleotide Sequence of the Nar Operon and Reassessment of the Role of the α and β Subunits in Iron Binding and Electron Transfer.” *Molecular and General Genetics MGG* 218 (2): 249–56. https://doi.org/10.1007/BF00331275.

Brockhurst, Michael A., Ellie Harrison, James P. J. Hall, Thomas Richards, Alan McNally, and Craig MacLean. 2019. “The Ecology and Evolution of Pangenomes.” *Current Biology* 29 (20): R1094–1103. https://doi.org/10.1016/j.cub.2019.08.012.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND.” *Nature Methods* 18 (4): 366–68. https://doi.org/10.1038/s41592-021-01101-x.

Cahen, David, Israel Pecht, and Mordechai Sheves. 2021. “What Can We Learn from Protein-Based Electron Transport Junctions?” *The Journal of Physical Chemistry Letters* 12 (47): 11598–603. https://doi.org/10.1021/acs.jpclett.1c02446.

Cantalapiedra, Carlos P, Ana Hernández-Plaza, Ivica Letunic, Peer Bork, and Jaime Huerta-Cepas. 2021. “eggNOG-Mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale.” *Molecular Biology and Evolution* 38 (12): 5825–29. https://doi.org/10.1093/molbev/msab293.

Case, DAVID A., THOMAS E. CHEATHAM, TOM DARDEN, HOLGER GOHLKE, RAY LUO, KENNETH M. MERZ, ALEXEY ONUFRIEV, CARLOS SIMMERLING,

BING WANG, and ROBERT J. WOODS. 2005. “The Amber Biomolecular Simulation Programs.” *Journal of Computational Chemistry* 26 (16): 1668–88. https://doi.org/10.1002/jcc.20290.
Cava, Felipe, Olga Zafra, and José Berenguer. 2008. “A Cytochrome c Containing Nitrate Reductase Plays a Role in Electron Transport for Denitrification in Thermus Thermophilus without Involvement of the Bc Respiratory Complex.” *Molecular Microbiology* 70 (2): 507–18. https://doi.org/10.1111/j.1365-2958.2008.06429.x.
Chaumeil, Pierre-Alain, Aaron J Mussig, Philip Hugenholtz, and Donovan H Parks. 2020. “GTDB-Tk: A Toolkit to Classify Genomes with the Genome Taxonomy Database.” *Bioinformatics* 36 (6): 1925–27. https://doi.org/10.1093/bioinformatics/btz848.
———. 2022. “GTDB-Tk v2: Memory Friendly Classification with the Genome Taxonomy Database.” *Bioinformatics* 38 (23): 5315–16. https://doi.org/10.1093/bioinformatics/btac672.
Chklovski, Alex, Donovan H. Parks, Ben J. Woodcroft, and Gene W. Tyson. 2023. “CheckM2: A Rapid, Scalable and Accurate Tool for Assessing Microbial Genome Quality Using Machine Learning.” *Nature Methods* 20 (8): 1203–12. https://doi.org/10.1038/s41592-023-01940-w.
Daims, Holger, Elena V. Lebedeva, Petra Pjevac, Ping Han, Craig Herbold, Mads Albertsen, Nico Jehmlich, et al. 2015. “Complete Nitrification by Nitrospira Bacteria.” *Nature* 528 (7583): 504–9. https://doi.org/10.1038/nature16461.
Daims, Holger, Sebastian Lücker, and Michael Wagner. 2016. “A New Perspective on Microbes Formerly Known as Nitrite-Oxidizing Bacteria.” *Trends in Microbiology* 24 (9): 699–712. https://doi.org/10.1016/j.tim.2016.05.004.
Dubourdieu, M, and J A DeMoss. 1992. “The narJ Gene Product Is Required for Biogenesis of Respiratory Nitrate Reductase in Escherichia Coli.” *Journal of Bacteriology* 174 (3): 867–72. https://doi.org/10.1128/jb.174.3.867-872.1992.
Eddy, Sean R. 2011. “Accelerated Profile HMM Searches.” *PLOS Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.
Edgar, Robert C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics* 26 (19): 2460–61. https://doi.org/10.1093/bioinformatics/btq461.
———. 2022. “High-Accuracy Alignment Ensembles Enable Unbiased Assessments of Sequence Homology and Phylogeny.” bioRxiv. https://doi.org/10.1101/2021.06.20.449169.
Eren, A. Murat, Evan Kiefl, Alon Shaiber, Iva Veseli, Samuel E. Miller, Matthew S. Schechter, Isaac Fink, et al. 2021. “Community-Led, Integrated, Reproducible Multi-Omics with Anvi’o.” *Nature Microbiology* 6 (1): 3–6. https://doi.org/10.1038/s41564-020-00834-3.
Eric P. Nawrocki, Sean R. Eddy. 2013. “Infernal 1.1: 100-Fold Faster RNA Homology Searches,” November. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3810854/.
Erisman, Jan Willem, Mark A. Sutton, James Galloway, Zbigniew Klimont, and Wilfried Winiwarter. 2008. “How a Century of Ammonia Synthesis Changed the World.” *Nature Geoscience* 1 (10): 636–39. https://doi.org/10.1038/ngeo325.
Gilchrist, Cameron, and Yit-Heng Chooi. 2021. “Clinker & Clustermap.Js: Automatic Generation of Gene Cluster Comparison Figures.” Python. https://doi.org/10.1093/bioinformatics/btab007.
Graaf, Astrid A. van de, Peter de Bruijn, Lesley A. Robertson, Mike S. M. Jetten, and J. Gijs Kuenen. 1996. “Autotrophic Growth of Anaerobic Ammonium-Oxidizing Micro-Organisms in a Fluidized Bed Reactor.” *Microbiology* 142 (8): 2187–96. https://doi.org/10.1099/13500872-142-8-2187.
Gray, Harry B., and Jay R. Winkler. 2005. “Long-Range Electron Transfer.” *Proceedings of the National Academy of Sciences* 102 (10): 3534–39. https://doi.org/10.1073/pnas.0408029102.

Haber, Fritz. 1920. “The Synthesis of Ammonia from Its Elements Nobel Lecture.” *Resonance* 7 (9): 86–94. https://doi.org/10.1007/BF02836189.
Haft, Daniel H., Jeremy D. Selengut, Roland A. Richter, Derek Harkins, Malay K. Basu, and Erin Beck. 2013. “TIGRFAMs and Genome Properties in 2013.” *Nucleic Acids Research* 41 (Database issue): D387–95. https://doi.org/10.1093/nar/gks1234.
“Heatmaply Documentation.” n.d. Accessed June 1, 2024. https://cran.r-project.org/web/packages/heatmaply/vignettes/heatmaply.html.
Hemp, James, Sebastian Lücker, Joachim Schott, Laura A. Pace, Jena E. Johnson, Bernhard Schink, Holger Daims, and Woodward W. Fischer. 2016. “Genomics of a Phototrophic Nitrite Oxidizer: Insights into the Evolution of Photosynthesis and Nitrification.” *The ISME Journal* 10 (11): 2669–78. https://doi.org/10.1038/ismej.2016.56.
Huerta-Cepas, Jaime, Damian Szklarczyk, Davide Heller, Ana Hernández-Plaza, Sofia K Forslund, Helen Cook, Daniel R Mende, et al. 2019. “eggNOG 5.0: A Hierarchical, Functionally and Phylogenetically Annotated Orthology Resource Based on 5090 Organisms and 2502 Viruses.” *Nucleic Acids Research* 47 (D1): D309–14. https://doi.org/10.1093/nar/gky1085.
Hug, Laura A., Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, et al. 2016. “A New View of the Tree of Life.” *Nature Microbiology* 1 (5): 1–6. https://doi.org/10.1038/nmicrobiol.2016.48.
Hyatt, Doug, Gwo-Liang Chen, Philip F. LoCascio, Miriam L. Land, Frank W. Larimer, and Loren J. Hauser. 2010. “Prodigal: Prokaryotic Gene Recognition and Translation Initiation Site Identification.” *BMC Bioinformatics* 11 (1): 119. https://doi.org/10.1186/1471-2105-11-119.
Johnson, L. Steven, Sean R. Eddy, and Elon Portugaly. 2010. “Hidden Markov Model Speed Heuristic and Iterative HMM Search Procedure.” *BMC Bioinformatics* 11 (1): 1–8. https://doi.org/10.1186/1471-2105-11-431.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89. https://doi.org/10.1038/s41586-021-03819-2.
Katoh, Kazutaka, Kazuharu Misawa, Kei-ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66. https://doi.org/10.1093/nar/gkf436.
Kluyver, and Donker. 1926. “Unity in Biochemistry.”
Kuypers, Marcel M. M., Hannah K. Marchant, and Boran Kartal. 2018. “The Microbial Nitrogen-Cycling Network.” *Nature Reviews Microbiology* 16 (5): 263–76. https://doi.org/10.1038/nrmicro.2018.9.
Lanciano, Pascal, Alexandra Vergnes, Steéphane Grimaldi, Bruno Guigliarelli, and Axel Magalon. 2007. “Biogenesis of a Respiratory Complex Is Orchestrated by a Single Accessory Protein *.” *Journal of Biological Chemistry* 282 (24): 17468–74. https://doi.org/10.1074/jbc.M700994200.
Laslett, Dean, and Bjorn Canback. 2004. “ARAGORN, a Program to Detect tRNA Genes and tmRNA Genes in Nucleotide Sequences.” *Nucleic Acids Research* 32 (1): 11–16. https://doi.org/10.1093/nar/gkh152.
Lee, Michael D. 2019. “GToTree: A User-Friendly Workflow for Phylogenomics.” *Bioinformatics* 35 (20): 4162–64. https://doi.org/10.1093/bioinformatics/btz188.
Lücker, Sebastian, Michael Wagner, Frank Maixner, Eric Pelletier, Hanna Koch, Benoit Vacherie, Thomas Rattei, et al. 2010. “A Nitrospira Metagenome Illuminates the Physiology and Evolution of Globally Important Nitrite-Oxidizing Bacteria.” *Proceedings of the National Academy of Sciences* 107 (30): 13479–84. https://doi.org/10.1073/pnas.1003860107.

Ma, Bin, Caiyu Lu, Yiling Wang, Jingwen Yu, Kankan Zhao, Ran Xue, Hao Ren, et al. 2023. “A Genomic Catalogue of Soil Microbiomes Boosts Mining of Biodiversity and Genetic Resources.” *Nature Communications* 14 (1): 7318. https://doi.org/10.1038/s41467-023-43000-z.

Matsen, Frederick A., Robin B. Kodner, and E Virginia Armbrust. 2010. “Pplacer: Linear Time Maximum-Likelihood and Bayesian Phylogenetic Placement of Sequences onto a Fixed Reference Tree.” *BMC Bioinformatics* 11 (1): 538. https://doi.org/10.1186/1471-2105-11-538.

Mirdita, Milot, Konstantin Schütze, Yoshitaka Moriwaki, Lim Heo, Sergey Ovchinnikov, and Martin Steinegger. 2022. “ColabFold: Making Protein Folding Accessible to All.” *Nature Methods* 19 (6): 679–82. https://doi.org/10.1038/s41592-022-01488-1.

Nayfach, Stephen, Simon Roux, Rekha Seshadri, Daniel Udwary, Neha Varghese, Frederik Schulz, Dongying Wu, et al. 2021. “A Genomic Catalog of Earth’s Microbiomes.” *Nature Biotechnology* 39 (4): 499–509. https://doi.org/10.1038/s41587-020-0718-6.

Nomoto, T, Y Fukumori, and T Yamanaka. 1993. “Membrane-Bound Cytochrome c Is an Alternative Electron Donor for Cytochrome Aa3 in Nitrobacter Winogradskyi.” 1993. https://doi.org/10.1128/jb.175.14.4400-4404.1993.

on, An anvi’o workflow for phylogenomics was published, and Last Modified on. 2017. “An Anvi’o Workflow for Phylogenomics.” Meren Lab. June 7, 2017. https://merenlab.org/2017/06/07/phylogenomics/.

Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T.G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” *Bioinformatics* 31 (22): 3691–93. https://doi.org/10.1093/bioinformatics/btv421.

Parks, Donovan H, Maria Chuvochina, Christian Rinke, Aaron J Mussig, Pierre-Alain Chaumeil, and Philip Hugenholtz. 2022. “GTDB: An Ongoing Census of Bacterial and Archaeal Diversity through a Phylogenetically Consistent, Rank Normalized and Complete Genome-Based Taxonomy.” *Nucleic Acids Research* 50 (D1): D785–94. https://doi.org/10.1093/nar/gkab776.

Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25 (7): 1043–55. https://doi.org/10.1101/gr.186072.114.

Pitcher, Robert S., Myles R. Cheesman, and Nicholas J. Watmough. 2002. “Molecular and Spectroscopic Analysis of the Cytochrome*Cbb*3 Oxidase from *Pseudomonas Stutzeri**.” *Journal of Biological Chemistry* 277 (35): 31474–83. https://doi.org/10.1074/jbc.M204103200.

Price et al. 2010. “FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments | PLOS ONE.” 2010. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0009490.

Punta, Marco, Penny C. Coggill, Ruth Y. Eberhardt, Jaina Mistry, John Tate, Chris Boursnell, Ningze Pang, et al. 2012. “The Pfam Protein Families Database.” *Nucleic Acids Research* 40 (D1): D290–301. https://doi.org/10.1093/nar/gkr1065.

Remmert, Michael, Andreas Biegert, Andreas Hauser, and Johannes Söding. 2012. “HHblits: Lightning-Fast Iterative Protein Sequence Searching by HMM-HMM Alignment.” *Nature Methods* 9 (2): 173–75. https://doi.org/10.1038/nmeth.1818.

Robertson, J. G., B. Wells, N. J. Brewin, E. Wood, C. D. Knight, and J. A. Downie. 1985. “The Legvme-Rhizobium Symbiosis: A Cell Surface Interaction.” *Journal of Cell Science* 1985 (Supplement_2): 317–31. https://doi.org/10.1242/jcs.1985.Supplement_2.17.

Rothery, R. A., F. Blasco, A. Magalon, M. Asso, and J. H. Weiner. 1999. “The Hemes of

Escherichia Coli Nitrate Reductase A (NarGHI): Potentiometric Effects of Inhibitor Binding to narI." *Biochemistry* 38 (39): 12747–57. https://doi.org/10.1021/bi990533o.
"Rpy2 Github." (2018) 2024. Python. rpy2. https://github.com/rpy2/rpy2.
Schmidt, Thomas S. B., Anthony Fullam, Pamela Ferretti, Askarbek Orakov, Oleksandr M. Maistrenko, Hans-Joachim Ruscheweyh, Ivica Letunic, et al. 2024. "SPIRE: A Searchable, Planetary-Scale mIcrobiome REsource." *Nucleic Acids Research* 52 (D1): D777–83. https://doi.org/10.1093/nar/gkad943.
Seemann, Torsten. 2014. "Prokka: Rapid Prokaryotic Genome Annotation." *Bioinformatics* 30 (14): 2068–69. https://doi.org/10.1093/bioinformatics/btu153.
Simon, Jörg, and Martin G. Klotz. 2013. "Diversity and Evolution of Bioenergetic Systems Involved in Microbial Nitrogen Compound Transformations." *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, The evolutionary aspects of bioenergetic systems, 1827 (2): 114–35. https://doi.org/10.1016/j.bbabio.2012.07.005.
Sorokin, Dimitry Y., Dana Vejmelkova, Sebastian Lücker, Galina M. Streshinskaya, W. Irene C. Rijpstra, Jaap S. Sinninghe Damsté, Robbert Kleerbezem, Mark van Loosdrecht, Gerard Muyzer, and Holger Daims. 2014. "Nitrolancea Hollandica Gen. Nov., Sp. Nov., a Chemolithoautotrophic Nitrite-Oxidizing Bacterium Isolated from a Bioreactor Belonging to the Phylum Chloroflexi." *International Journal of Systematic and Evolutionary Microbiology* 64 (Pt_6): 1859–65. https://doi.org/10.1099/ijs.0.062232-0.
Spieck, Eva, and Eberhard Bock. 2015a. "Nitrobacter." In *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–11. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm00803.
———. 2015b. "Nitrococcus." In *Bergey's Manual of Systematics of Archaea and Bacteria*, 1–7. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm01130.
Spieck, Eva, Michael Spohn, Katja Wendt, Eberhard Bock, Jessup Shively, Jeroen Frank, Daniela Indenbirken, Malik Alawi, Sebastian Lücker, and Jennifer Hüpeden. 2020. "Extremophilic Nitrite-Oxidizing Chloroflexi from Yellowstone Hot Springs." *The ISME Journal* 14 (2): 364–79. https://doi.org/10.1038/s41396-019-0530-9.
Starkenburg, Shawn R., Patrick S. G. Chain, Luis A. Sayavedra-Soto, Loren Hauser, Miriam L. Land, Frank W. Larimer, Stephanie A. Malfatti, et al. 2006. "Genome Sequence of the Chemolithoautotrophic Nitrite-Oxidizing Bacterium Nitrobacter Winogradskyi Nb-255." *Applied and Environmental Microbiology* 72 (3): 2050–63. https://doi.org/10.1128/AEM.72.3.2050-2063.2006.
Starkenburg, Shawn R., Frank W. Larimer, Lisa Y. Stein, Martin G. Klotz, Patrick S. G. Chain, Luis A. Sayavedra-Soto, Amisha T. Poret-Peterson, et al. 2008. "Complete Genome Sequence of Nitrobacter Hamburgensis X14 and Comparative Genomic Analysis of Species within the Genus Nitrobacter." *Applied and Environmental Microbiology* 74 (9): 2852–63. https://doi.org/10.1128/AEM.02311-07.
Steinegger, Martin, Markus Meier, Milot Mirdita, Harald Vöhringer, Stephan J. Haunsberger, and Johannes Söding. 2019. "HH-Suite3 for Fast Remote Homology Detection and Deep Protein Annotation." *BMC Bioinformatics* 20 (1): 473. https://doi.org/10.1186/s12859-019-3019-7.
Steinegger, Martin, and Johannes Söding. 2017. "MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets." *Nature Biotechnology* 35 (11): 1026–28. https://doi.org/10.1038/nbt.3988.
Tanaka, Yoshikazu, Yoshihiro Fukumori, and Tateo Yamanaka. 1983. "Purification of Cytochrome A1c1 from Nitrobacter Agilis and Characterization of Nitrite Oxidation System of the Bacterium." *Archives of Microbiology* 135 (4): 265–71. https://doi.org/10.1007/BF00413479.

Teufel et al. 2022. “SignalP 6.0 Predicts All Five Types of Signal Peptides Using Protein Language Models - PMC.” 2022. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9287161/.

Vergnes, Alexandra, Janine Pommier, René Toci, Francis Blasco, Gérard Giordano, and Axel Magalon. 2006. “NarJ Chaperone Binds on Two Distinct Sites of the Aponitrate Reductase of Escherichia Coli to Coordinate Molybdenum Cofactor Insertion and Assembly.” *The Journal of Biological Chemistry* 281 (4): 2170–76. https://doi.org/10.1074/jbc.M505902200.

Wee, JunJie, and Guo-Wei Wei. 2024. “Evaluation of AlphaFold 3’s Protein–Protein Complexes for Predicting Binding Free Energy Changes upon Mutation.” *Journal of Chemical Information and Modeling* 64 (16): 6676–83. https://doi.org/10.1021/acs.jcim.4c00976.

Winogradsky. 1890. “Sur Les Organisms de La Nitrification,” 4, 215–31, 257, – 275, 760–71.

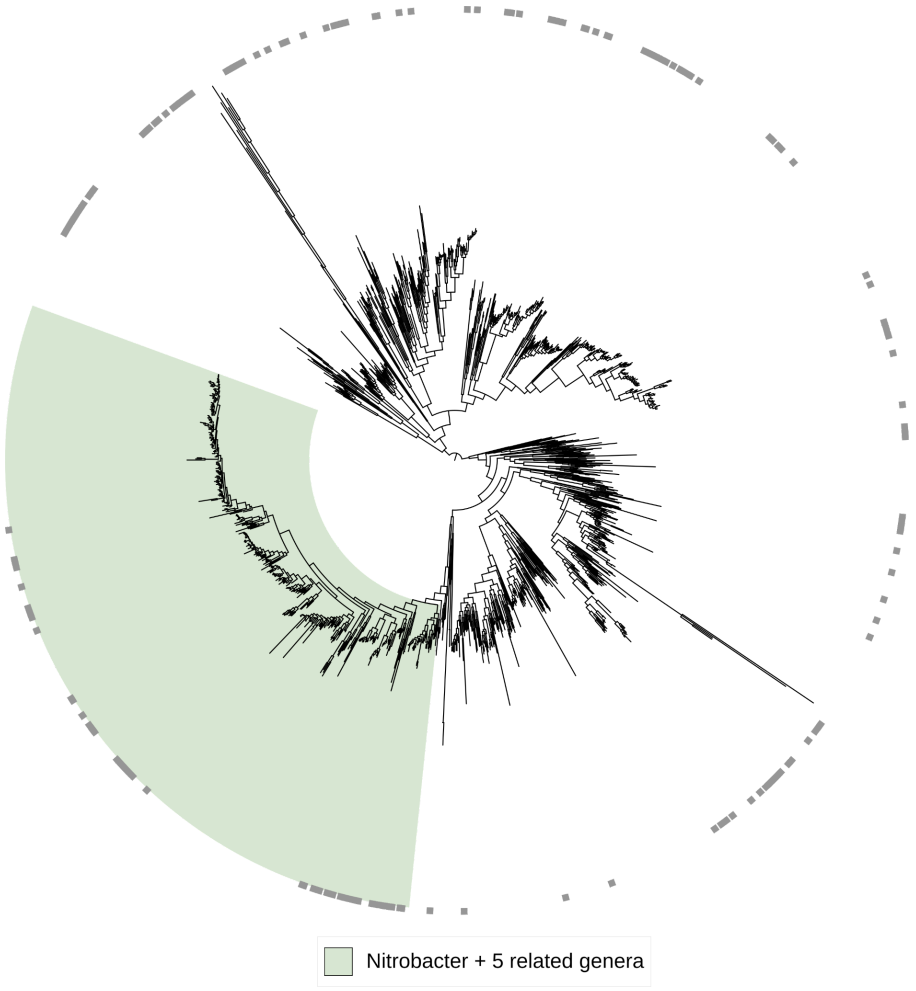
Zheng, Wan, Pengxin Chai, Jiapeng Zhu, and Kai Zhang. 2024. “High-Resolution in Situ Structures of Mammalian Respiratory Supercomplexes.” *Nature* 631 (8019): 232–39. https://doi.org/10.1038/s41586-024-07488-9.

Supplementary material

1. SCC gene profiles used for concatenated core gene phlogenies

Ribosomal_L2, Ribosomal_L3, Ribosomal_L4, Ribosomal_L5, Ribosomal_L6, Ribosomal_L14, Ribosomal_L16, Ribosomal_L18p, Ribosomal_L22, ribosomal_L24, Ribosomal_S3_C, Ribosomal_S8, Ribosomal_S10, Ribosomal_S17, Ribosomal_S19

2. Full Tree of the Xanthobacteraceae



3. Code

Code is available at https://github.com/KilianG98/EGDist