Evolutionary ecology of natural comammox *Nitrospira* populations

Alejandro Palomo[1], Arnaud Dechesne[1], Otto X. Cordero[2] and Barth F. Smets[1]

[1]Department of Environmental Engineering, Technical University of Denmark, Kgs. Lyngby, Denmark

[2]Ralph M. Parsons Laboratory for Environmental Science and Engineering, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

Microbial life on Earth commonly occurs in diverse and complex communities where species interact, and their genomic repertoires evolve over time. Our understanding of species interaction and evolution has increased during last decades, but most studies of evolutionary dynamics are based on single species in isolation or experimental systems composed of few interacting species. Here, we use the microbial ecosystem found in groundwater-fed sand filters as a model to avoid this limitation. In these systems, diverse microbial communities experience relatively stable conditions, and the coupling between chemical and biological processes is generally well defined. Metagenomic analysis of 12 sand filters revealed systematic co-occurrence of at least five comammox *Nitrospira* species, favoured by low ammonium concentrations. *Nitrospira* species showed intra-population sequence diversity, although possible clonal expansion was detected in few abundant local comammox populations. *Nitrospira* populations were separated by gene flow boundaries, suggesting natural and cohesive populations. They showed low homologous recombination and strong purifying selection, the latest process being especially strong in genes essential in energy metabolism. Positive selection was detected on genes related to resistance to foreign DNA and phages. Additionally, we analysed evolutionary processes in populations from different habitats. Interestingly, our results suggest that in comammox *Nitrospira* these processes are not an intrinsic feature but greatly vary depending on the habitat they inhabit. Compared to other habitats, groundwater fed sand filters impose strong purifying selection and low recombination. Together, this study improves understanding of interactions and evolution of species in the wild, and sheds light on the environmental dependency of evolutionary processes.

Microorganisms dominate the tree of life based on species number and diversity, and they play an essential role in Earth's global biogeochemical cycles. Microbial species interact with each other and with the environment (ecological processes), and also undergo changes in their genomic repertoire over time (evolutionary processes). Yet, the interaction between ecological and evolutionary processes is largely unknown, especially for complex open communities. For many years, most studies of microbial communities in open, complex environments have focused on ecological aspects as it was believed that evolutionary changes happen at much larger timescale[1]. However, in recent years, with the growth of population-genomics analysis, researchers have started to investigate both ecological and evolutionary processes in microbial communities. Yet, most studies of evolutionary dynamics are based on single species in isolation[2] or experimental systems composed of only a few interacting species[3]. Although these analyses have helped to better understand some aspects of evolutionary processes patterns, they have limitations because they lack many characteristics of actual natural populations (spatial structure, existence of microdiversity, predation, immigration, etc.). On the other hand, observing populations in the wild also has limitations because the conditions vary with little control (hence uncontrolled variation in population size, selection regime) and the typically unknown ecophysiology of retrieved genomes makes it difficult to interpret the observed patterns. Therefore, studying well-defined model microbial ecosystems can help to understand ecological and evolutionary processes in microbial communities[4].

Rapid sand filters (RSF), widely used to produce drinking water from surface- or groundwater, can be a useful model system as they are characterized by stable conditions and active growth primarily driven by the oxidation of ammonia, methane, and other inorganic compounds present at low concentration in the influent water, large populations (~ 10^9 cells/g), significant mixing, continuous but limited immigration from prokaryotes in the influent water, no dispersal between separate sand filters (resulting in allopatric populations), and relatively well defined coupling between chemical and biological processes[5–7]. In addition, microbial communities inhabiting these systems have been described and show the dominance of complete ammonia oxidizers (comammox)[8,9], which are expected to have a relatively simple basic ecology (due to their chemolithoautotrophic metabolism)[10], yet are poorly studied in terms of what drives their diversity, distribution and evolution. Furthermore, as comammox bacteria occur in RSF as coexisting populations[9,11], RSF offer an appropriate opportunity for resolving fine-scale genomic heterogeneity within closely related strains, and investigate if they show similar patterns in evolutionary processes (selection, recombination, etc.).

With the RSF microbial ecosystem model, different eco-evolutionary questions can be addressed. Of particular interest is to what extent the evolutionary processes that drive the diversification of *Nitrospira* species in RSF are dependent on their environment, as opposed to intrinsic properties of the species. Environmental dependency of microbial evolution has been investigated from different perspectives. Several studies have focused on genome signatures variations (GC, tetranucleotide signatures, codon usage, purine-pyrimidine ratio) associated with different environments (reviewed in Dutta and Paul (2012)[12]). Others, have studied bacterial adaptation to shifting environments[13], or have targeted a specific evolutionary process across several lifestyles (homologous recombination[14], selection[15], etc.). Most of these studies, however, considered different species living in different environments, or closely related species with a different lifestyle (i.e. free-living organisms vs pathogens). Yet, little is known about ongoing evolutionary processes of species belonging to the same lineage inhabiting different open environments. In this study, taking advantage of the multiple *Nitrospira* species present in several groundwater-fed RSF, we thoroughly investigated evolutionary processes on this local environment, and compared these observations with those in *Nitrospira* species inhabiting other open environments.

Results and Discussion

In this study, we examined ecological and evolutionary patterns within comammox dominated-bacterial communities inhabiting groundwater-fed rapid sand filters. To that end, we retrieved *Nitrospira* metagenome-assembled genomes (MAGs) from 12 similarly operated waterworks in Denmark using a combination of automatic and manual binning (Supplementary Table 1). These MAGs spanned 16 putative species (further on simply referred to as 'species') using a threshold average nucleotide identity (ANI) of ≥ 95%[16–18]. The phylogenomic analysis placed one *Nitrospira* species into lineage I, 14 into lineage II, and one into other lineages (Fig. 1). Of the 16 *Nitrospira* species, 12 were classified as comammox *Nitrospira* (5 clade A and 7 clade B) (Fig. 1).

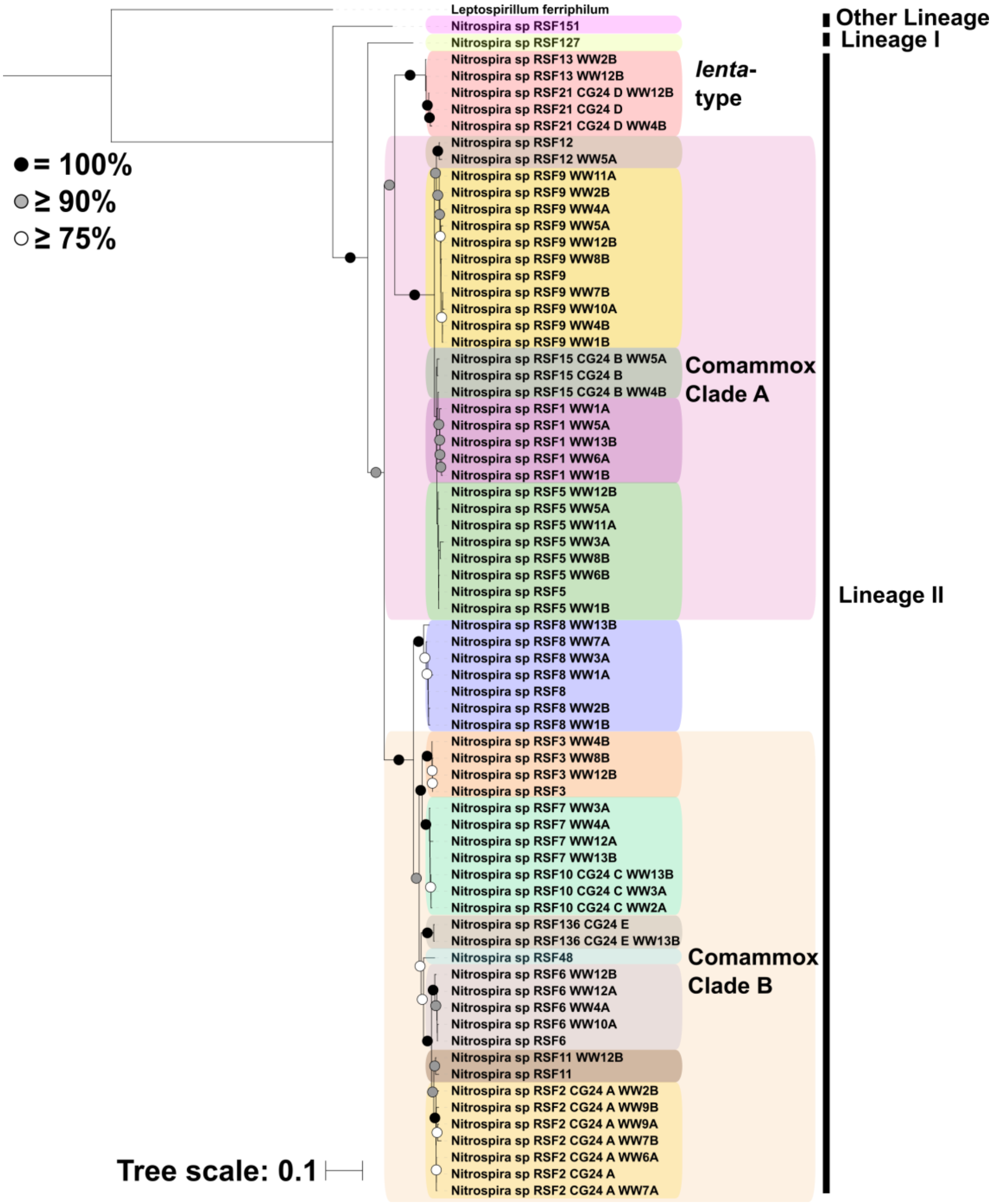

Fig. 1. Phylogenomic affiliation of *Nitrospira* MAGs retrieved from 12 waterworks. A phylogenetic tree was built based on the concatenation of 120 proteins. MAGs affiliated to same *Nitrospira* species (MAGS with ANI ≥ 95% are considered members of the same species) are shown with same colours. *Leptospirillum* was used to root the tree. The strength of support for internal nodes as assessed by bootstrap replicates is indicated as coloured circles (top left legend).

Nitrospira species comprised a large proportion of the microbial communities of the waterworks (27 - 70%), and comammox represented a large fraction of *Nitrospira* spp. (76 - 98%) (Fig. 2 and Supplementary Table 2). Multiple *Nitrospira* species (at least 5, 10 on average) co-occurred in all the waterworks (Fig. 2). In some cases, a single species constituted the majority of the community (WW4 and WW5), while, in other, two (WW9 and WW12) or more species dominated (Fig. 2). Distinct species were dominant in the different waterworks (Fig. 2). Generally, there was no significant correlation among the abundances of the comammox species, with the exception of few strong positive correlations (RSF5 and RSF15_CG24_B (ρ=0.84), RSF6 and RSF11 (ρ=0.84)), and a negative correlation between two species (RSF3 and RSF7 (ρ=-0.47)) (Supplementary Fig. 1). Thus, co-exclusion between comammox *Nitrospira* species seems to be rare in the studied waterworks. One reason to explain this phenomenon could be the existence of large amount of unique genes in each comammox *Nitrospira* species, which has been previously reported[11]. As in other coexisting microorganisms, these unique genes (although the function of most of them is still unknown) might promote differences in chemotactic strategies, attachment to particles strategies, secondary metabolism, defence against predation, etc. (Reviewed in Stilianos *et al.*, (2018)[19]). Furthermore, the sand grains of the investigated waterworks are very porous with bacterial cells embedded in different parts of the grain[20], which could enable fine-scale spatial separation among the comammox species. Nevertheless, more research is needed to properly characterise the reason enabling the observed co-occurrence of comammox *Nitrospira* species.

Although correlations between *Nitrospira* species were rare, positive correlations between the abundance of *Nitrospira* spp. and that of other microbes inhabiting the waterworks were more frequent (Supplementary Fig. 1). In particular, the abundance of most of the ammonia oxidizing bacterial (AOB) species positively correlated with canonical *Nitrospira* species, and interestingly, with one comammox *Nitrospira* species (RSF3) (Supplementary Fig. 1). We also observed that distinct *Nitrospira* species positively correlated with different phages retrieved from the waterworks, suggesting a specificity between phages and hosts (Supplementary Fig. 2).

The chemical characteristics of the water explained 31% of the variance in *Nitrospira* composition (permutation test: $p < 0.001$; Supplementary Fig. 3), suggesting that water chemistry is a strong filter for the assembly of these nitrifying communities. Among the measured water constituents, the influent ammonium concentration was the variable that best explained the *Nitrospira* distribution (explained 18%; permutation test: $p = 0.002$). Higher comammox species richness was detected in waterworks treating lower ammonium concentration (Supplementary Fig. 4, $R^2 = 0.54$, $p < 0.01$). In contrast, most of canonical *Nitrospira* and canonical ammonia oxidizers were related with waterworks containing higher ammonium concentration in their influent (Supplementary Fig. 3). These observations are in line whit a previous study which predicted that higher ammonium concentration favours emergence of division of labour (canonical ammonia and nitrite oxidiser)[21].

Nevertheless, we observed that a few comammox species seems to also cope with slightly higher ammonium concentrations (Supplementary Fig. 3). Moreover, as previously reported[22], no clear distinction in comammox species distribution was observed based on their clade affiliation (which depends on the phylogeny of ammonia monooxygenase subunit A) (Supplementary Fig. 3). Different from the water chemistry composition, the distribution patterns of *Nitrospira* species across the waterworks were not related to their geographic distance (Mantel test: r statistics = 0.08 and significance > 0.05 vs. r statistics = 0.36 and significance < 0.001 for water chemistry).

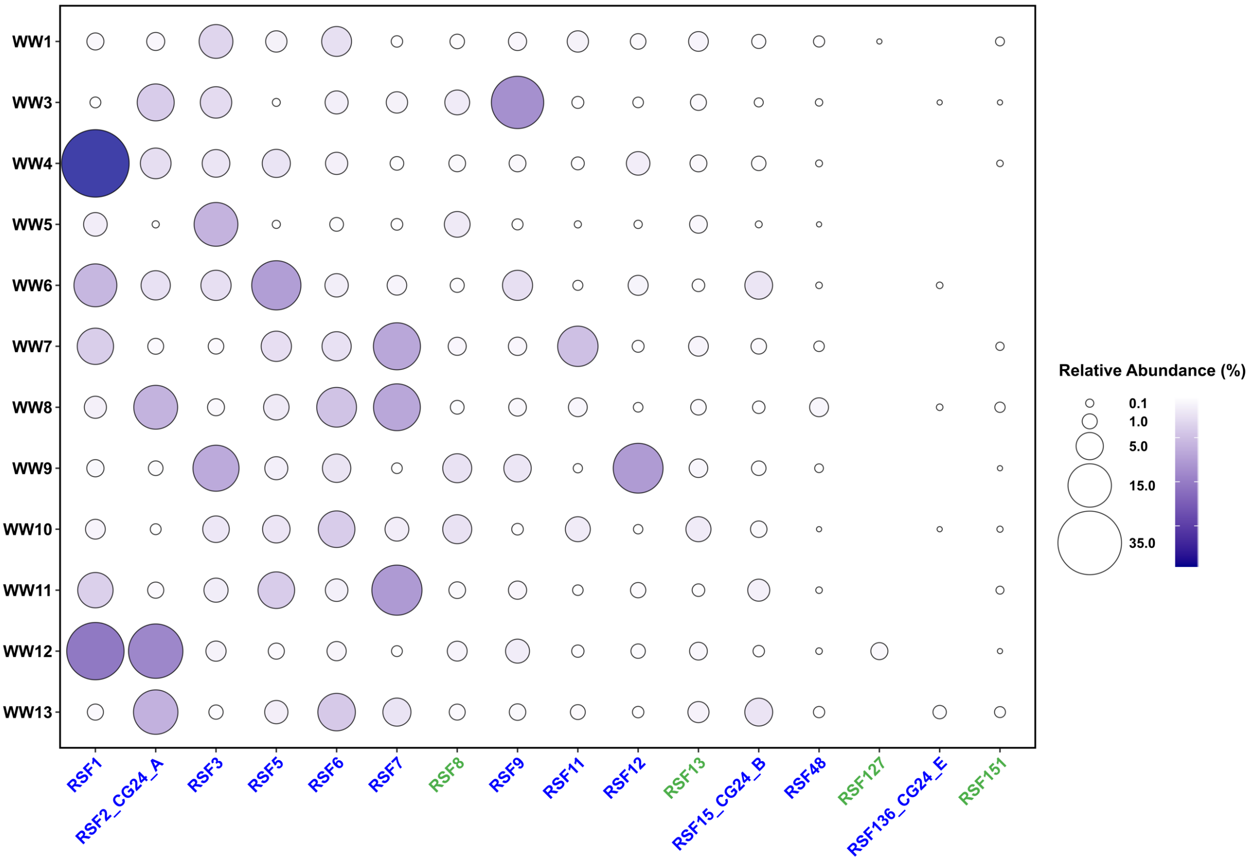

Fig. 2. Abundance of *Nitrospira* species across 12 waterworks.
Relative abundance of 16 *Nitrospira* species in 12 waterworks. Comammox and canonical *Nitrospira* species are denoted in blue and green, respectively.

Microdiversity within *Nitrospira* species

Strain-level analysis across the waterworks revealed that the *Nitrospira* populations contained intra-population sequence diversity. We exploited the shotgun metagenomic data from the different waterworks to perform strain-level analyses using single nucleotide polymorphisms (SNPs). The number of SNPs/Mbp in the populations across the waterworks ranged from 14,437 to 45,664 (Supplementary Table 3). Looking into the populations at local scale (species within waterworks), the number of SNPs/Mbp ranged from 249 to 37,663 (Supplementary Table 4). We observed a wide range of microdiversity (measured as nucleotide diversity (π)) among populations (Fig. 3A): canonical *Nitrospira* RSF8 was the most diverse species, with three times more nucleotide diversity than the less diverse *Nitrospira* species of our study (RSF1 and RSF12) (Fig. 3A). Both a homogeneous degree of microdiversity across the waterworks (e.g., RSF5 and RSF8), as well as a high microdiversity variation depending on the waterworks (e.g., RSF1, RSF9 and RSF11) was detected among the *Nitrospira* populations (Supplementary Fig. 5). Based on the observations done at species level, we hypothesised that microdiversity would be higher at low ammonium concentrations, where we observed higher comammox species diversity. However, this was not the case, as for each species, the correlations of microdiversity with ammonium concentration or comammox species richness were not significant ($p > 0.05$).

In contrast to the observation across all the waterworks (i.e.: all species showed significant microdiversity across waterworks), we did detect a few highly abundant comammox populations with almost no microdiversity at local scale (e.g., comammox *Nitrospira* RSF1 in WW4, and comammox *Nitrospira* RSF12 in WW9) (Supplementary Fig. 6), which suggests local clonal expansions of these comammox populations in specific waterworks. In the same line, the analysis of major allele frequencies of common SNPs (for each species, SNPs present in all the strains present in each waterworks) revealed that only a fraction of the subspecies diversity is found locally (for a specific species, we detected different subspecies among waterworks, some of them being genetically homogenous. E.g.: comammox RSF3 (WW10B vs WW5) (Supplementary Fig. 7)). These results contrast with what we observed at species level, where all the diversity was represented in each waterworks (as they all contain most of the *Nitrospira* spp.; Fig. 2).

Similar to what we observed at species level, there was no significant correlation between similarity in subspecies composition and the geographic distance of the waterworks, with exception of *Nitrospira* sp. RSF2 ($p < 0.01$) and *Nitrospira* sp. RSF8 ($p < 0.05$) (Supplementary Fig. 8). However, we observed a geographic organisation of the genetic structure at most loci across the studied genomes, indicating that the *Nitrospira* populations were more similar within than between waterworks. We calculated pairwise fixation indexes F_{ST} (which measures differences in allele frequencies between populations of the same species found in two distinct waterworks) for each gene between allele frequencies from the twelve waterworks. The mean gene F_{ST} values were ≥ 15% for all *Nitrospira* populations (Supplementary Fig. 9), the most extreme case being the RSF5 population ($F_{ST} > 40\%$) (Supplementary Fig. 9). In few populations (RSF2, RSF7 and RSF8), a higher dispersal ($F_{ST} < 20\%$) of most alleles between waterworks was observed (Supplementary Fig. 9). These observations differ from observations from soil bacterial populations across a meadow, where most of the populations had mean gene F_{ST} values < 5%[23]. These contrasting results support the notion that populations in the waterworks are much more allopatric than the ones from the mentioned meadow.

We also investigated local regions of the *Nitrospira* genomes with significantly higher F_{ST} values, as this is characteristic of local population-specific (here in each waterworks) selective pressures acting on specific loci[23]. Few loci were found with unusually high F_{ST} in some of the *Nitrospira* populations (Supplementary Fig. 10 and Supplementary Table 5), one of them being related with genes involved in nitrogen assimilation (*Nitrospira* sp. RSF2) (Supplementary Table 5). However, only two of these loci with unusually high site-specific differentiation of alleles (high F_{ST}) also had fewer recombinant events, and lower nucleotide diversity (Supplementary Table 5), which can be considered as a signal of recent selective sweep[23]. These results suggest that contrary to what it has been observed in several natural populations[23–26], gene-specific sweeps seems to play a minor role in the evolution of *Nitrospira* spp. inhabiting the waterworks. A possible explanation could be the low recombination rate that characterised the waterworks *Nitrospira* populations (Fig. 3B, discussed below), as opposed to genome-wide sweeps that are associated to low recombination rates[27], gene-specific sweeps are expected to occur with high recombination rates[27].

Overall, across the 12 waterworks, all species present significant genomic microdiversity but this diversity was not always represented locally, with a few occurrences of patterns consistent with clonal expansion. The reason for the difference of within-species diversity across waterworks is unknown but the allopatric nature of the communities likely contributes to their persistence.

Evolutionary processes at whole-genome level

The *Nitrospira* populations were characterised by a low degree of homologous recombination. We investigated homologous recombination in the *Nitrospira* populations based on linkage disequilibrium (D’ is only < 1 if all possible combinations of a pair of biallelic sites are observed[28]. Lower D’ values indicates higher levels of homologous recombination; Fig. 3B) and linkage decay (Supplementary Fig. 11). Similar results were observed for other abundant populations inhabiting the waterworks (Fig. 3B). In general, recombination was lower in the waterworks populations than in populations inhabiting a grassland meadow[23], where a similar study was conducted (Fig. 3B). To further examine the relative effect of homologous recombination on the genetic diversification of populations, we measured the rates at which nucleotides become substituted as a result of recombination versus mutation using the *r/m* ratio. Most of the *Nitrospira* populations had a relatively low *r/m* ($r/m < 2$) compared to recombinogenic species reported in literature ($r/m > 4$)[29] (Supplementary Fig. 12), although in one case (RSF15_CG24_B) the rate was similar to the value reported for a *S. flavogriseus* population ($r/m = 28$) considered to be approaching panmixia[30]. Overall, these results suggest a low effect of recombination in the *Nitrospira* population inhabiting the studied waterworks. Increasing recombination rate has been associated with fluctuating environments as a source of variation which can accelerate adaptation favouring survival in this type of environments[31,32]. On the other hand, constant environments - as the waterworks studied here - tend to reduce recombination rates of inhabiting microbes[31].

A phylogenetic analysis to assess the impact of recombination across the different *Nitrospira* species showed that these ones were separated by gene flow boundaries, consistent with the notion that these species represent cohesive populations (i.e. gene flow within species is higher than between species). We performed a quartet-based phylogenetic analysis for each *Nitrospira* species by building tree quartets of whole-genome orthologous genes (“gene trees”) and core genes (“species trees”) with four strains of the same species or with two pairs of strains from two different species. In most of the pairwise species comparisons, the percentage of gene trees supporting the species phylogeny was > 98% (Supplementary Fig. 13). In few cases, this percentage was lower (down to 67% in the comparison between RSF1 and RSF15_CG24_B), but these numbers were always above the within-species analysis (Supplementary Fig. 13).

The *Nitrospira* populations were characterised by strong purifying selection. We used the relation between non-synonymous and synonymous polymorphisms (pN/pS) to investigate this evolutionary process. We detected $pN/pS < 1$, indicating purifying selection, for all *Nitrospira* species (Fig. 3C). Similar results were observed for other abundant populations inhabiting the waterworks (Fig. 3C). Purifying selection has frequently been observed in wild populations, and it was the case for populations inhabiting a grassland meadow[23] ($pN/pS = 0.56 \pm 0.17$, $n = 19$), but this process seems to be especially strong in the waterworks populations ($pN/pS = 0.34 \pm 0.05$, $n = 24$) (Two-Sample t-test, $p < 0.0001$) (Fig. 3C), which suggests prior evolution to optimal adaptation to this stable environment, with purging of non-synonymous mutations.

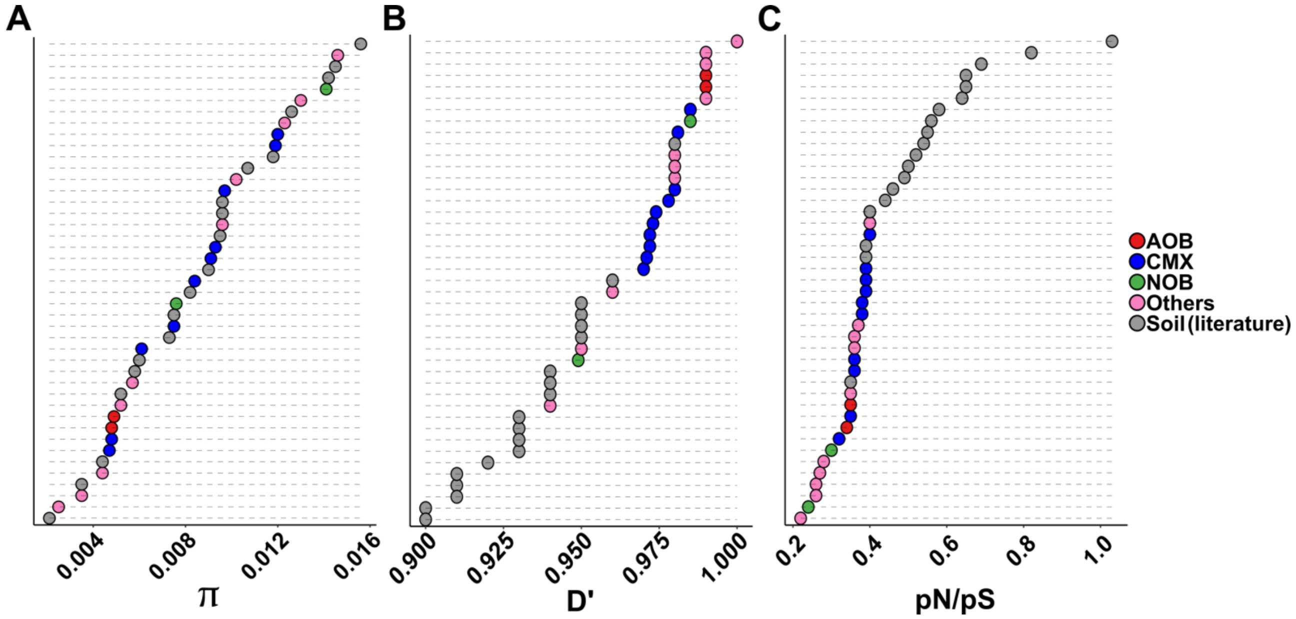

Fig. 3. Evolutionary metrics of *Nitrospira* populations across 12 waterworks.
A Right) Genetic diversity of most abundant bacterial populations across 12 waterworks. A left) It also includes most abundant bacterial populations across grassland meadow[23]. Microdiversity is measured as nucleotide diversity (π). B Right) Homologous recombination (D’) of most abundant bacterial populations across 12 waterworks. B Left) It includes most abundant bacterial populations across grassland meadow. C left) Selection (pN/pS ratio) of most abundant bacterial populations across 12 waterworks. C right) It includes most abundant bacterial populations across grassland meadow. Colour legends are displayed on the right of each figure.

Interestingly, the degree of recombination and diversity across different *Nitrospira* populations varied substantially with habitat (Fig. 4). High variability of recombination in closely related bacterial species has occasionally been reported[33], and lifestyle appear as one of the most relevant factors to explain this variability[14,33]. Our analysis of evolutionary processes in populations from different habitats (*Nitrospira* from drinking water treatment plants (DWTP), freshwaters and soils) suggests that the environment also influences ongoing evolutionary processes: different bacterial types inhabiting the same environment tended to share similar features (Fig. 4), while the evolutionary characteristics of comammox *Nitrospira* populations differed depending on the environment where they were retrieved (Fig. 4). Comammox species in the studied waterworks and in other DWTP were characterised by low recombination, strong purifying selection and moderate microdiversity (Fig. 4). On the other hand, comammox present in freshwater or, especially, in soils had higher microdiversity and recombination rate, and weaker purifying selection (Fig. 4). Intriguingly, we consistently observed that in drinking water treatment systems canonical *Nitrospira* species showed features similar to those of comammox *Nitrospira* but with even stronger purifying selection (Fig. 4). This feature, together with the much lower richness observed in canonical *Nitrospira* compared to comammox bacteria (Fig. 2), suggests that competition can play a more intense role in canonical *Nitrospira*, which might select for few species optimally adapted to this type of stable environment. However, a broader analysis is required to confirm this hypothesis.

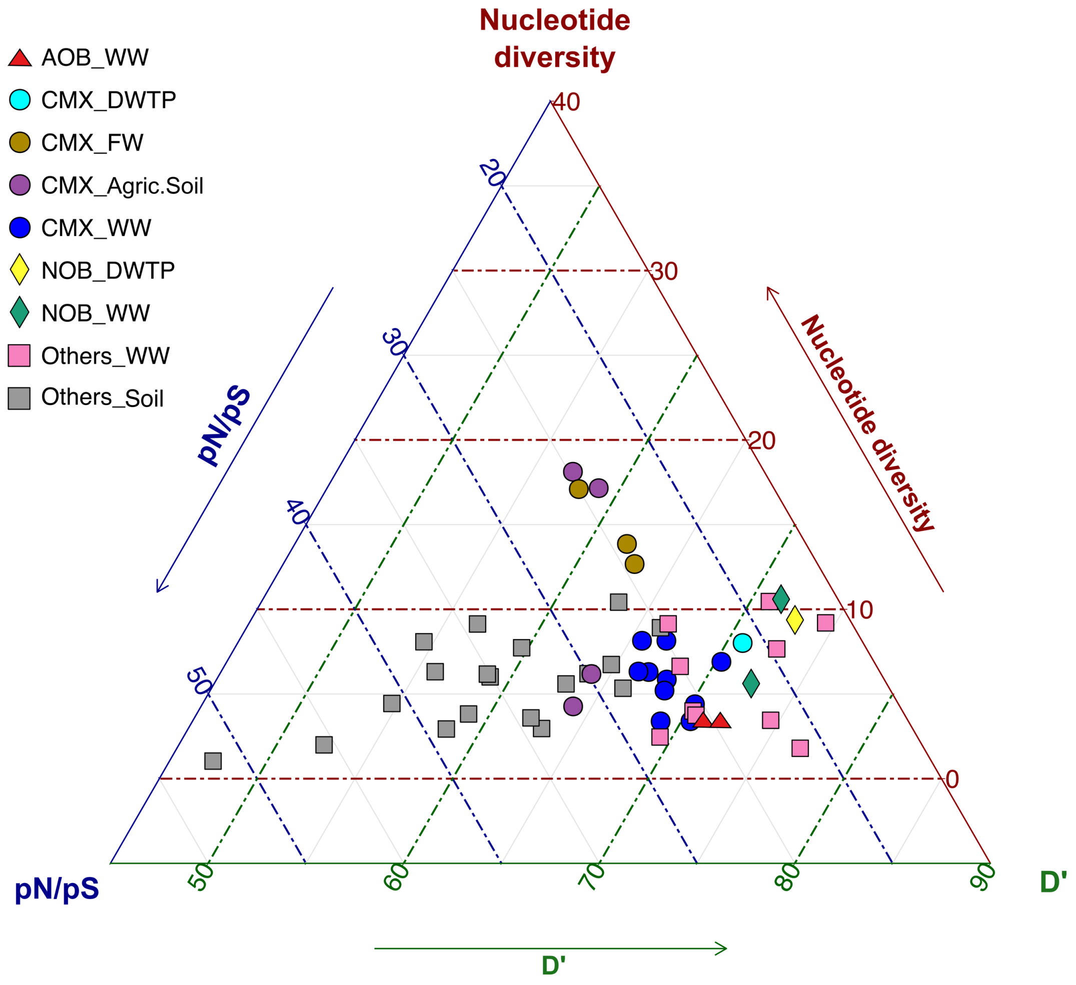

Fig. 4. Impact of environment and microbial type in evolutionary metrics

Triplot composed of the nucleotide diversity, pN/pS ratio and D' values for the bacterial populations of this study (WW) as well as most abundant bacterial populations across grassland meadow[23] (Soil), and other *Nitrospira* populations abundant in other systems (Supplementary Table 6; CMX_FW, CMX_DWTP, NOB_DWTP, and CMX_Agric.Soil). Colour legends are displayed on the left of the figure.

Evolutionary processes at the gene level

In addition to a genome-wide analysis, we investigated the evolutionary processes at the gene level. Genes involved in nitrification (ammonia monooxygenase: *amoA* and *amoB*; hydroxylamine dehydrogenase: *haoA* and *haoB*; nitrite oxidoreductase: *nxrA* and *nxrB*) in the studied *Nitrospira* populations generally had a similar nucleotide diversity (π) (Fig. 5A) and homologous recombination rate (D') (Fig. 5B) compared to the rest of the genome, but with higher levels of purifying selection (pN/pS) (Fig. 5C). The nucleotide diversities of genes related to nitrification were very similar with the exception of *amoB*, which had a significantly lower nucleotide diversity than *nxrB* ($p < 0.05$) (Fig. 5A). A similar pattern was detected for the recombination, but in this case *amoA*, as well as *amoB*, had significantly lower recombination than *nxrB* ($p < 0.05$) (Fig. 5B). We observed a very strong purifying selection for most of the nitrifying genes, especially for *amoA*, *nxrA*, and *nxrB* ($p < 0.01$) (Fig. 5C). In the case of *nxrB*, not a single non-synonymous mutation was found in most of the *Nitrospira* species (0-1 non-synonymous sites vs 17-66 synonymous sites), even though this gene had a higher nucleotide diversity and homologous recombination (Fig. 5A and Fig. 5B). Our observations on selection are in line with previous studies, as generally, bacterial essential genes and enzymes catalysing reactions that are difficult to by-pass through alternative pathways are subject to higher purifying selection compared to nonessential ones[34–37].

Even though the average pN/pS values were below 1 in all *Nitrospira* species (Fig. 3), indicating purifying selection, genes with pN/pS values above 1 and significantly higher than the genomic average were detected in each species (Supplementary Fig. 14). Many of those genes were related to defence mechanism against phages (e.g. genes putatively involved in phage entry into cells, ribonucleases, genes coding proteins associated to restriction-modification systems, genes related to toxin-antitoxin systems, etc (Supplementary Table 7)). Comparable findings were made in other abundant species from the waterworks (Supplementary Table 8), in the additional *Nitrospira* species retrieved from other environments (Supplementary Table 8), as well as by Petersen *et al.* (2007) and by Rabbi *et al*. (2015) in *E. coli* and *Vibrio* sp. strains, respectively[38,39]. These observations suggest that positive selection in phage-related genes is widespread across bacteria, and highlights the evolutionary arms race occurring between phages and bacteria as an important driver in bacterial ecology and evolution[27,40,41]. Additionally, we found nondefense, mobile genetic elements, such as transposons and integrases, with significantly higher pN/pS values than the genome average in the *Nitrospira* spp. (Supplementary Table 9).

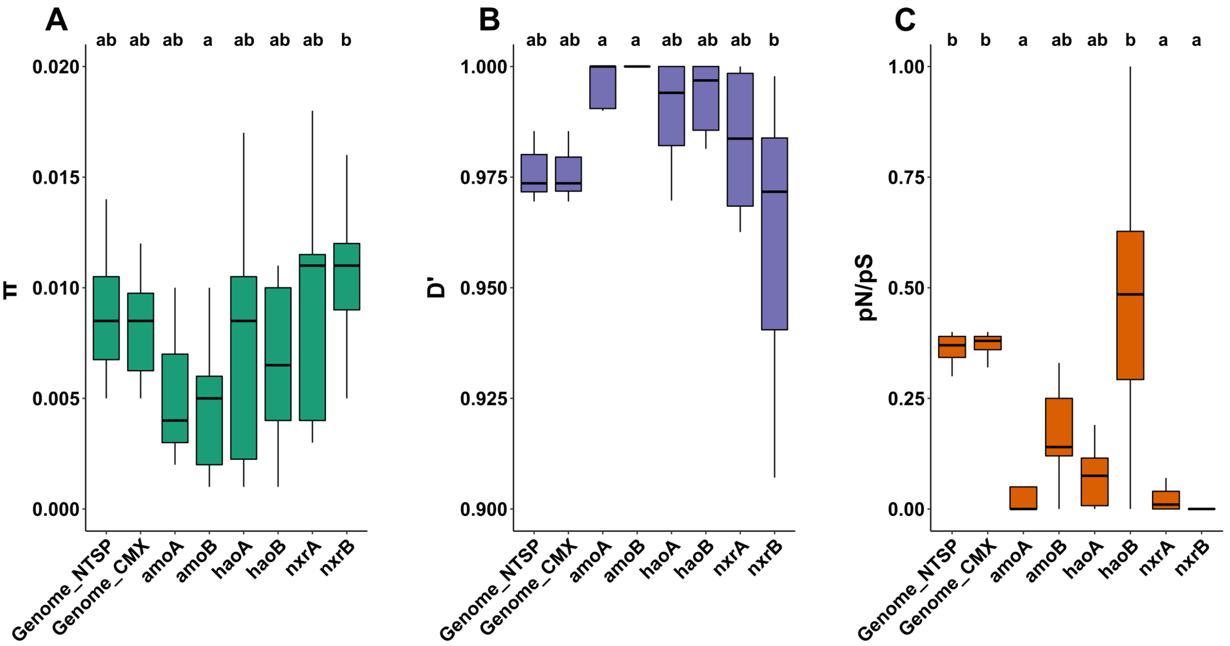

Fig. 5. Evolutionary metrics of nitrification genes in *Nitrospira* populations across 12 waterworks. Left) Boxplot of nucleotide diversity of *Nitrospira* bacterial populations for whole genome (all *Nitrospira* and comammox *Nitrospira*) and nitrification genes. Differences between the mean nucleotide diversities were assessed by a Dunn's test; same letter have means not significantly different from each other ($p < 0.05$). Middle) Boxplot of linkage disequilibrium of *Nitrospira* bacterial populations for whole genome (all *Nitrospira* and comammox *Nitrospira*) and nitrification genes. Differences between the mean linkage disequilibriums were assessed by a Dunn's test; same letter have means not significantly different from each other ($p < 0.05$). Right) Boxplot of pN/pS ratios of *Nitrospira* bacterial populations for whole genome (all *Nitrospira* and comammox *Nitrospira*) and nitrification genes. Differences between the mean pN/pS ratios were assessed by a Dunn's test; same letter have means not significantly different from each other ($p < 0.01$).

Conclusions

A major unresolved question is how the relationship between ecology and evolution shapes complex communities in wild environments. Here we use a model microbial system to examine this question. Strain-level analyses enabled us to decipher the degree of intra-population diversity in wild dominant comammox *Nitrospira* inhabiting rapid sand filters and estimate important evolutionary processes such as recombination and selection in these populations. We showed that compared to other environments the *Nitrospira* populations in rapid sand filters are characterized by low recombination and strong purifying selection therefore, we conclude that the evolutionary processes that drive the diversification of *Nitrospira* are dependent on the local environment, as opposed to intrinsic properties of the species.

Methods

Sampling, sequencing, and metagenomic assembled genomes recovery

The sampling description, DNA extraction and sequencing have been previously described[22]. Briefly, filter material was collected from two locations at the top of the filters of 12 Danish waterworks. DNA was extracted from 0.5 g of sand material using the MP FastDNA Spin Kit (MP Biomedicals LLC, Solon, USA). DNA libraries were generated using the 24 extracted DNA with the Nextera XT DNA library preparation kit (Illumina Inc.) according to the manufacturer's instructions. The samples were sequenced in one lane, with 2×150 paired read sequencing on the Illumina HiSeq4000 at BGIs facility in Copenhagen. As previously described[22], high-quality reads were used for metagenomic assembled genomes (MAGs) recovery using a combination of automatic and manual binning followed by filtering and refinement steps to improve the quality of the MAGs. The resulted MAGs were dereplicated using dRep[42] with the secondary clustering threshold -sa 0.99. Dereplicated MAGs completeness and contamination was evaluated using CheckM[43]. MAGs with completeness > 50% and contamination < 10% were kept for downstream analyses.

Species abundance estimation

A 95% average nucleotide identity (ANI) cut-off was used to define species as proposed by Klappenbach *et al.* (2007)[16]. The retrieved MAGs were dereplicated using dRep with the secondary clustering threshold set at 95% gANI. Among the genomes classified as belonging to the same species, the one with higher quality was chosen as representative genome for that species. The species abundance and coverage of each representative genome across the studied metagenomes was assessed using MIDAS [44]. Briefly, MIDAS uses reads mapped to 15 universal single-copy gene families (with ability to accurately recruit metagenomic reads to the correct species [44]) to estimate the abundance and coverage of bacterial species from a shotgun metagenome. We used the species retrieved in this study to build the database of universal-single-copy genes.

Genome classification and annotation

MAGs were classified (Supplementary Table 10) using the classify workflow of the GTDB-Tk v.0.1.3 tool[45]. Open reading frames were predicted using Prodigal v. 2.63[46], and annotated using blastp[47] against NCBI nr[48], UniProt[49], KEGG[50], PFAM[51] and eggNOG[52]. Genes were assigned to antiphage defense systems using the strategy described in Doron *et al.* (2018)[53].

Phylogenetic analysis

Phylogenetic analyses of *Nitrospira* genomes were conducted with the GTDB-Tk v.0.1.3 tool[45] using the *de novo* workflow with a set of 120 single copy marker proteins and the genome taxonomy database (GTDB)[54]. Concatenated alignments were used to construct a maximum likelihood tree using RAxML v. 8.2.11[55] with 400 rapid bootstraps (determined using the autoMRE option) and the LG likelihood model of amino acid substitution with gamma distributed rates and fraction of invariant sites (-m PROTGAMMAILGF; best model determined using ProtTest v. 3.4.2[56]). The tree was rooted using two *Leptospirillum* species as outgroup. The rooted tree was visualized using the online web tool from the Interactive Tree of Life (iTol)[57].

Quartet analysis

To retrieve waterworks-specific MAGs for each *Nitrospira* species, the reassembly module of metaWRAP[58] was used with individual reads from each sample and the representative *Nitrospira* species MAG. Resulted MAGs were kept for the quartet analysis if the completeness did not vary in more than 10% compared to the representative *Nitrospira* species MAG and the contamination remained < 5%. A species phylogenetic tree of the resulting *Nitrospira* MAGs was constructed as described above. For each quartet analysis we selected four *Nitrospira* MAGs. Four from the same *Nitrospira* species for within species analysis (the most phylogenetically distant MAGs), and two from one *Nitrospira* species and two from another one for between species analysis (the most phylogenetically distant MAGs). Orthofinder v. 2.3.3[59] was used to identify orthologous genes among each set of four genomes, retaining for subsequent analyses only single-copy orthologous genes. Orthofinder v. 2.3.3 with the options -M msa -T raxml was also used to produce phylogenetic trees for each orthologous gene. For each within or between species analysis, topological differences between each orthologous gene and the species tree were assessed by calculating the Robinson-Foulds (RF) distance[60] with the R function RF.dist of the phangorn package[61]. This analysis allowed to obtain, for each quartet, the percentage of single-copy orthologous genes phylogenetic trees which did not support the species phylogenetic tree topology.

Read mapping, SNP calling, and population genomic analysis

The population genomic analysis was done following the approach described in Crits-Christoph et al (2020)[23]. High-quality reads were mapped to an indexed database of the 176 species MAGs recovered from the waterworks using BWA-MEM[62]. Resulted alignments were filtered using samtools[63] view -q30 to remove reads with mapping quality less than 30, and also with the script filter_reads.py[23] (with the options: -m 96 to retain reads with a percent identity of at least 96% to the reference; and -q 2 to assure uniquely best mapping read pairs in the index). Downstream analysis was performed for 24 species genomes (all the 12 *Nitrospira* ones, and 12 other abundant species genomes). For each of these species genomes, we analysed its data in samples that passed a cutoff of at least 50% of the genome being covered with at least 5× coverage. 149 out of 576 sample genome comparisons (24 genomes × 24 samples) passed this minimum requirement. Sample read mappings were pooled by each waterworks and by all samples across the waterworks. Nucleotide diversity (π), linkage disequilibrium (D') and pN/pS ratio were calculated for each sample, each waterworks and across all the waterworks as described elsewhere[23] using the provided scripts. F_{ST} was calculated following the same procedure but on sites segregating across two waterworks being compared (for all the possible waterworks comparisons). As Crits-Christoph *et al.* (2020)[23] recommended, only sites with a coverage of at least 20× in each waterworks was used to calculate F_{ST}. In addition, genes with coverages in a waterworks outside of the range of two standard deviations were excluded from the analysis. As previously suggested[23], a two-sample Wilcoxon test was conducted to find out if average linkage of highly differentiated loci differed from the genomic average for each species. Similarly, a two-sample *t*-tests was used to conclude if average nucleotide diversity of highly differentiated loci differed from the genomic average. Both sets of tests were corrected for multiple hypotheses using the Benjamini–Hochberg method.

Same strain-level analysis as the one described above was conducted in *Nitrospira* MAGs previously recovered[22] that passed a cutoff of at least 50% of the genome being covered with at least 5× coverage in any of the metagenomes were *Nitrospira* MAGs were found to be present[22].

Statistical Analyses.

All statistical tests were performed using R v3.5.2[64]. For all statistical analyses, species abundances data was treated as followed: zeros were replaced with an estimate value using the Count Zero Multiplicative approach with the zCompositions R package[65], and data were further centred log-ratio transformed. *Nitrospira* community dissimilarities were calculated using the Jaccard index. The correlation between the *Nitrospira* community dissimilarities and geographic distances was calculated using the Mantel test (significance obtained after 100,000 permutations). Same analysis was used to assess the correlation between the *Nitrospira* community dissimilarities and the water composition dissimilarity, as well as the correlation between major allele dissimilarities and geographic distances.

Proportionality between abundances of the species across the 24 metagenomes were calculated using the propr R package[66] (with the options metric = "rho", ivar = "clr") and visualised using the corrplot R package[67]. For the network analysis, the function getNetwork from propr R package was used to retain proportionalities > 0.56 (FDR < 5%). The network was visualised using the igraph R package[68]. Same approach was used to build the network including phages but, in this case, proportionalities > 0.51 (FDR < 5%) were retained.

Redundancy analysis (RDA) was performed in R package vegan[69]. RDA was conducted using centred log-ratio transformed *Nitrospira* species abundances and chemical data of influent water. The constrained ordination model and the variable significance were determined by permutation tests (1000 permutations) with anova.cca in vegan. Principal components analysis (PCA) was performed in R package factoextra[70] using the nucleotide diversity, pN/pS ratio and D' values for the bacterial populations retrieved from the waterworks, as well as most abundant bacterial populations across grassland meadow[23], and other *Nitrospira* populations abundant in other systems.

Differences between the mean nucleotide diversities of the nitrifying genes, whole *Nitrospira* genomes, and whole comammox *Nitrospira* genomes were assessed using Kruskal–Wallis ANOVA followed by Dunn's test with the Holm-Bonferroni correction. Same analysis was performed for linkage disequilibrium and pN/pS ratios.

Recovery of draft phage genomes and abundance estimation

MARVEL[71] was used to recover draft phage genomes from the co-assembly generated from the waterworks (describe above). As recommended by MARVEL, Metabat[72] was run using the parameters -m 1500 -s 10000 to produce bins with contigs of at least 1500 bp and with a minimum total size of 10 kbp. Then, we executed MARVEL to identify phage bins from the 1026 bins generated with Metabat. The abundance in each sample of the 43 identified draft phage genomes was estimated using the quant module from metaWRAP[73] (Supplementary Table 11) .

Chemical analysis of influent water

Ammonium was measured using a standard colorimetric salicylate and hypochlorite method[74]. Iron and manganese content was determined by ICP-MS (7700x, Agilent Technologies). NVOC analysis was performed using a wet chemical TOC-analyser TOC-V WP (Shimadzu, Kyoto, Japan).

Data availability

All raw sequence data and *Nitrospira* genomes retrieved from the Danish rapid sand filters have been deposited at NCBI under the project PRJNA384587. The rest of the retrieved draft genomes from the Danish rapid sand filters are available on figshare (https://doi.org/10.6084/m9.figshare.12962075).

Acknowledgements

We thank Gabriel E. Leventhal for helpful discussions on data analysis and theory. This research was supported by a research grant (13391, Expa-N) from VILLUM FONDEN.

Authors contributions

A.P conceived the study and performed the bioinformatic analyses. A.P, O.X.C and A.D led interpretation of the results supported by B.F.S. A.P drafted the manuscript with input from A.D, O.X.C and B.F.S. All authors contributed to manuscript revision, and approved the final version of the manuscript.

References

1. Hairston, N. G., Ellner, S. P., Geber, M. A., Yoshida, T. & Fox, J. A. Rapid evolution and the convergence of ecological and evolutionary time. *Ecol. Lett.* **8**, 1114–1127 (2005).
2. Barroso-Batista, J. *et al.* The First Steps of Adaptation of Escherichia coli to the Gut Are Dominated by Soft Sweeps. *PLoS Genet.* **10**, e1004182 (2014).
3. Lawrence, D. *et al.* Species Interactions Alter Evolutionary Responses to a Novel Environment. *PLoS Biol.* **10**, e1001330 (2012).
4. Denef, V. J., Mueller, R. S. & Banfield, J. F. AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J.* **4**, 599–610 (2010).
5. Palomo, A. *et al.* Metagenomic analysis of rapid gravity sand filter microbial communities suggests novel physiology of Nitrospira spp. *ISME J.* **10**, 2569–2581 (2016).
6. Gülay, A. *et al.* DNA- and RNA-SIP Reveal Nitrospira spp. as Key Drivers of Nitrification in Groundwater-Fed Biofilters. *MBio* **10**, (2019).
7. Hu, W. *et al.* Metagenomics Unravels Differential Microbiome Composition and Metabolic Potential in Rapid Sand Filters Purifying Surface Water Versus Groundwater. *Environ. Sci. Technol.* **54**, 5197–5206 (2020).
8. Tatari, K. *et al.* Density and distribution of nitrifying guilds in rapid sand filters for drinking water production: Dominance of Nitrospira spp. *Water Res.* **127**, 239–248 (2017).
9. Fowler, S. J., Palomo, A., Dechesne, A., Mines, P. D. & Smets, B. F. Comammox Nitrospira are abundant ammonia oxidizers in diverse groundwater-fed rapid sand filter communities.

Environ. Microbiol. **20**, 1002–1015 (2018).

10. Koch, H., Kessel, M. A. H. J. van & Lücker, S. Complete nitrification: insights into the ecophysiology of comammox Nitrospira. *Appl. Microbiol. Biotechnol.* 1–13 (2018). doi:10.1007/s00253-018-9486-3
11. Palomo, A. *et al.* Comparative genomics sheds light on niche differentiation and the evolutionary history of comammox Nitrospira. *ISME J.* **12**, 1779–1793 (2018).
12. Dutta, C. & Paul, S. Microbial Lifestyle and Genome Signatures. *Curr. Genomics* **13**, 153–162 (2012).
13. Scheuerl, T. *et al.* Bacterial adaptation is constrained in complex communities. *Nat. Commun.* **11**, 754 (2020).
14. González-Torres, P., Rodríguez-Mateos, F., Antón, J. & Gabaldón, T. Impact of Homologous Recombination on the Evolution of Prokaryotic Core Genomes. *MBio* **10**, (2019).
15. Martinez, J. L. The role of natural environments in the evolution of resistance traits in pathogenic bacteria. *Proc. R. Soc. B Biol. Sci.* **276**, 2521–2530 (2009).
16. Klappenbach, J. A. *et al.* DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *Int. J. Syst. Evol. Microbiol.* **57**, 81–91 (2007).
17. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
18. Olm, M. R. *et al.* Consistent Metagenome-Derived Metrics Verify and Delineate Bacterial Species Boundaries. *mSystems* **5**, (2020).
19. Louca, S. *et al.* Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* (2018). doi:10.1038/s41559-018-0519-1
20. Gülay, A. *et al.* Internal porosity of mineral coating supports microbial activity in rapid sand filters for groundwater treatment. *Appl. Environ. Microbiol.* **80**, 7010–7020 (2014).
21. Costa, E., Pérez, J. & Kreft, J.-U. Why is metabolic labour divided in nitrification? *Trends Microbiol.* **14**, 213–219 (2006).
22. Palomo, A., Dechesne, A. & Smets, B. F. Genomic profiling of Nitrospira species reveals ecological success of comammox Nitrospira. *bioRxiv* (2019). doi:10.1101/612226
23. Crits-Christoph, A., Olm, M. R., Diamond, S., Bouma-Gregson, K. & Banfield, J. F. Soil bacterial populations are shaped by recombination and gene-specific selection across a grassland meadow. *ISME J.* 1–25 (2020). doi:10.1038/s41396-020-0655-x
24. Shapiro, B. J. *et al.* Population Genomics of Early Events in the Ecological Differentiation of Bacteria. *Science (80-.).* **336**, 48–51 (2012).
25. Rosen, M. J., Davison, M., Bhaya, D. & Fisher, D. S. Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. *Science (80-.).* **348**, 1019–1023 (2015).
26. Bendall, M. L. *et al.* Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *Isme J* **10**, 1589–1601 (2016).
27. Shapiro, B. J., Leducq, J.-B. & Mallet, J. What Is Speciation? *PLOS Genet.* **12**, e1005860 (2016).
28. VanLiere, J. M. & Rosenberg, N. A. Mathematical properties of the measure of linkage disequilibrium. *Theor. Popul. Biol.* **74**, 130–137 (2008).
29. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* **16**, 199–204 (2019).
30. Doroghazi, J. R. & Buckley, D. H. Widespread homologous recombination within and between Streptomyces species. *ISME J.* **4**, 1136–1143 (2010).
31. Carja, O., Liberman, U. & Feldman, M. W. Evolution in changing environments: Modifiers

of mutation, recombination, and migration. *Proc. Natl. Acad. Sci.* **111**, 17935–17940 (2014).

32. Hanage, W. P. Not So Simple After All: Bacteria, Their Population Genetics, and Recombination. *Cold Spring Harb. Perspect. Biol.* **8**, a018069 (2016).

33. Didelot, X. & Maiden, M. C. J. Impact of recombination on bacterial evolution. *Trends Microbiol.* **18**, 315–322 (2010).

34. Luo, H., Gao, F. & Lin, Y. Evolutionary conservation analysis between the essential and nonessential genes in bacterial genomes. *Sci. Rep.* **5**, 13210 (2015).

35. Dilucca, M., Cimini, G. & Giansanti, A. Essentiality, conservation, evolutionary pressure and codon bias in bacterial genomes. *Gene* **663**, 178–188 (2018).

36. Aguilar-Rodríguez, J. & Wagner, A. Metabolic Determinants of Enzyme Evolution in a Genome-Scale Bacterial Metabolic Network. *Genome Biol. Evol.* **10**, 3076–3088 (2018).

37. Zhong, C. *et al.* Pan-genome analyses of 24 Shewanella strains re-emphasize the diversification of their functions yet evolutionary dynamics of metal-reducing pathway. *Biotechnol. Biofuels* **11**, 193 (2018).

38. Petersen, L., Bollback, J. P., Dimmic, M., Hubisz, M. & Nielsen, R. Genes under positive selection in Escherichia coli. *Genome Res.* **17**, 1336–1343 (2007).

39. Rabby, A. *et al.* Identification of the positively selected genes governing host-pathogen arm race in Vibrio sp. through comparative genomics approach. *Biojournal Sci. Technol.* **2**, (2015).

40. Rodriguez-Valera, F. *et al.* Explaining microbial population genomics through phage predation. *Nat. Rev. Microbiol.* **7**, 828–836 (2009).

41. Cordero, O. X. & Polz, M. F. Explaining microbial genomic diversity in light of evolutionary ecology. *Nat. Rev. Microbiol.* **12**, 263–273 (2014).

42. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

43. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25**, 1043–55 (2015).

44. Nayfach, S., Rodriguez-Mueller, B., Garud, N. & Pollard, K. S. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res.* **26**, 1612–1625 (2016).

45. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz848

46. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).

47. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

48. Sayers, E. W. *et al.* GenBank. *Nucleic Acids Res.* **47**, D94–D99 (2019).

49. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).

50. Ogata, H. *et al.* KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27**, 29–34 (1999).

51. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–D301 (2012).

52. Huerta-Cepas, J. *et al.* eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res.* **44**, D286–D293 (2016).

53. Doron, S. *et al.* Systematic discovery of antiphage defense systems in the microbial pangenome. *Science (80-.).* **359**, eaar4120 (2018).

54. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* (2018). doi:10.1038/nbt.4229

55. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).

56. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).

57. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).

58. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

59. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

60. Robinson, D. F. & Foulds, L. R. Comparison of phylogenetic trees. *Math. Biosci.* **53**, 131–147 (1981).

61. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).

62. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–95 (2010).

63. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

64. Team, R. C. R: A language and environment for statistical computing. (2014).

65. Palarea-Albaladejo, J. & Martín-Fernández, J. A. zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.* **143**, 85–96 (2015).

66. Quinn, T. P., Richardson, M. F., Lovell, D. & Crowley, T. M. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *Sci. Rep.* **7**, 16252 (2017).

67. Wei, T. & Simko, V. R Package 'corrplot': visualization of a correlation matrix. (2017).

68. Gabor, C. & Tamas, N. The igraph software package for complex network research. *InterJournal* **Complex Sy**, 1695 (2006).

69. Oksanen, J. *et al.* Package 'vegan' Title Community Ecology Package. *Community Ecol. Packag.* **2**, 1–297 (2019).

70. Kassambara, A. & Mundt, F. factoextra: Extract and Visualize the Results of Multivariate Data Analyses. *https://CRAN.R-project.org/package=factoextra* (2017).

71. Amgarten, D., Braga, L. P. P., da Silva, A. M. & Setubal, J. C. MARVEL, a Tool for Prediction of Bacteriophage Sequences in Metagenomic Bins. *Front. Genet.* **9**, (2018).

72. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3**, e1165 (2015).

73. Uritskiy, G. V., DiRuggiero, J. & Taylor, J. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. *Microbiome* **6**, 158 (2018).

74. Bower, C. E. & Holm-Hansen, T. A Salicylate–Hypochlorite Method for Determining Ammonia in Seawater. *Can. J. Fish. Aquat. Sci.* **37**, 794–798 (1980).