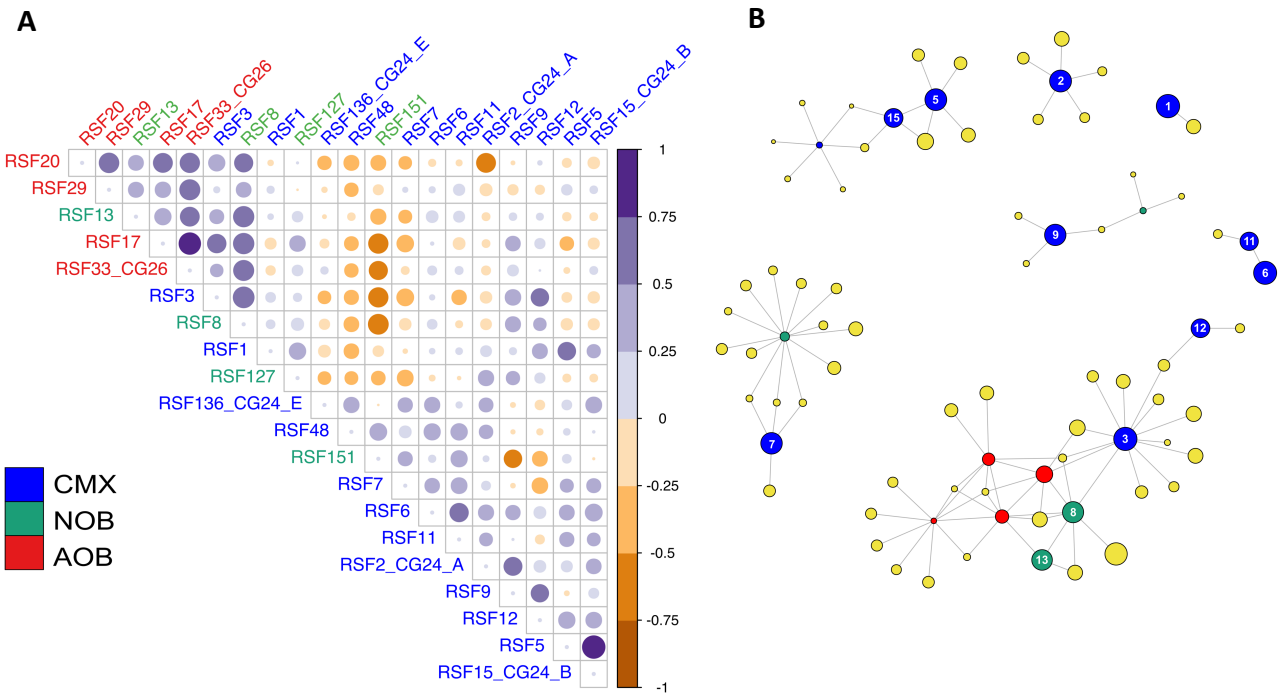
Supplementary Information

Evolutionary ecology of natural comammox *Nitrospira* populations

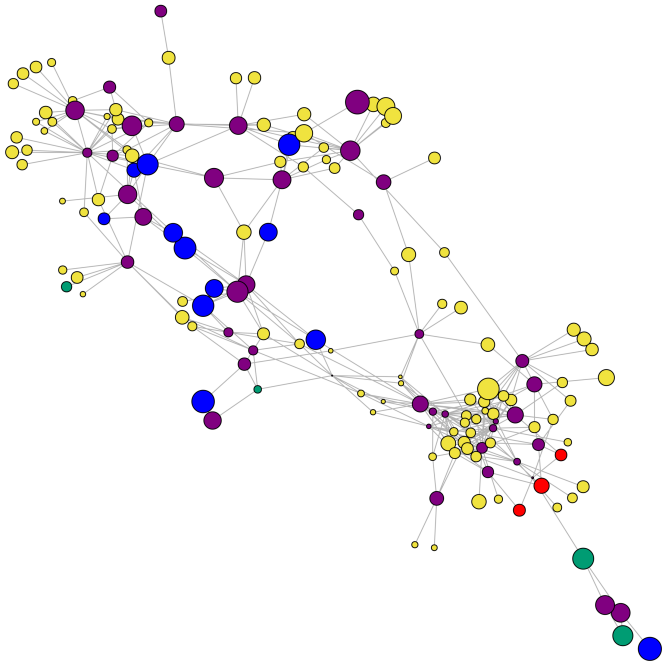
Alejandro Palomo[1], Arnaud Dechesne[1], Otto X. Cordero[2] and Barth F. Smets[1]
[1]Department of Environmental Engineering, Technical University of Denmark, Kgs Lyngby, Denmark
[2]Ralph M. Parsons Laboratory for Environmental Science and Engineering, Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

Supplementary Information includes:
Supplementary Fig. 1 to 14

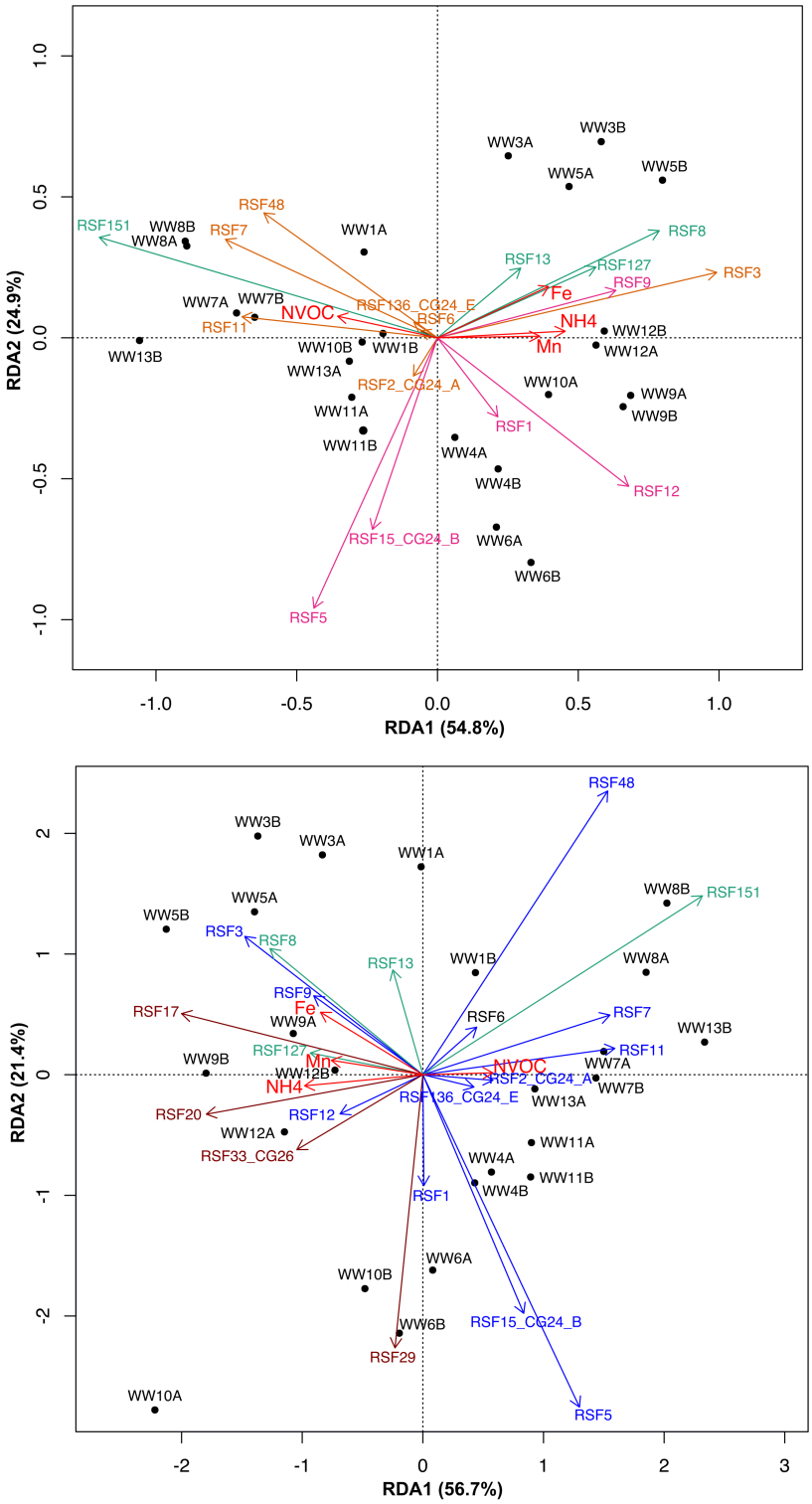
Other supplementary materials for this manuscript include the following:
Supplementary Table 1 to 11



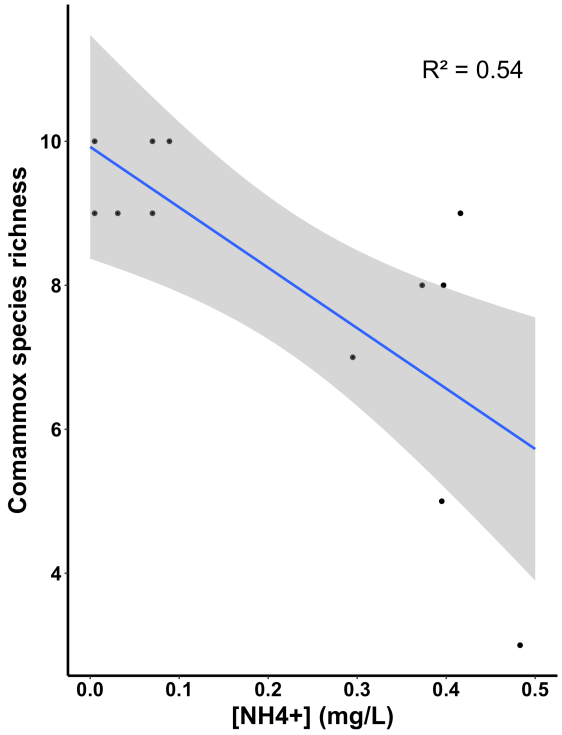
Supplementary Figure 1. A) Correlogram showing proportionality for centred log-transformed abundances of the *Nitrospira* and AOB species across the 12 waterworks. Colour indicates whether the correlation is positive (purple) or negative (brown). Size and darkness of the circles indicate the strength of the proportionality, with stronger proportionality being larger and darker than weaker ones. B) Network analysis revealing the co-occurrence patterns among species present in the studied waterworks. The nodes were coloured according to species type (lineage II canonical *Nitrospira*, green; comammox, blue; AOB, red; other bacteria, yellow). A connection represents a strong proportionality (ρ>0.55 and FDR < 0.05). The size of each node is proportional to the average species log-transformed abundance. Only nodes connected with *Nitrospira* species are shown.



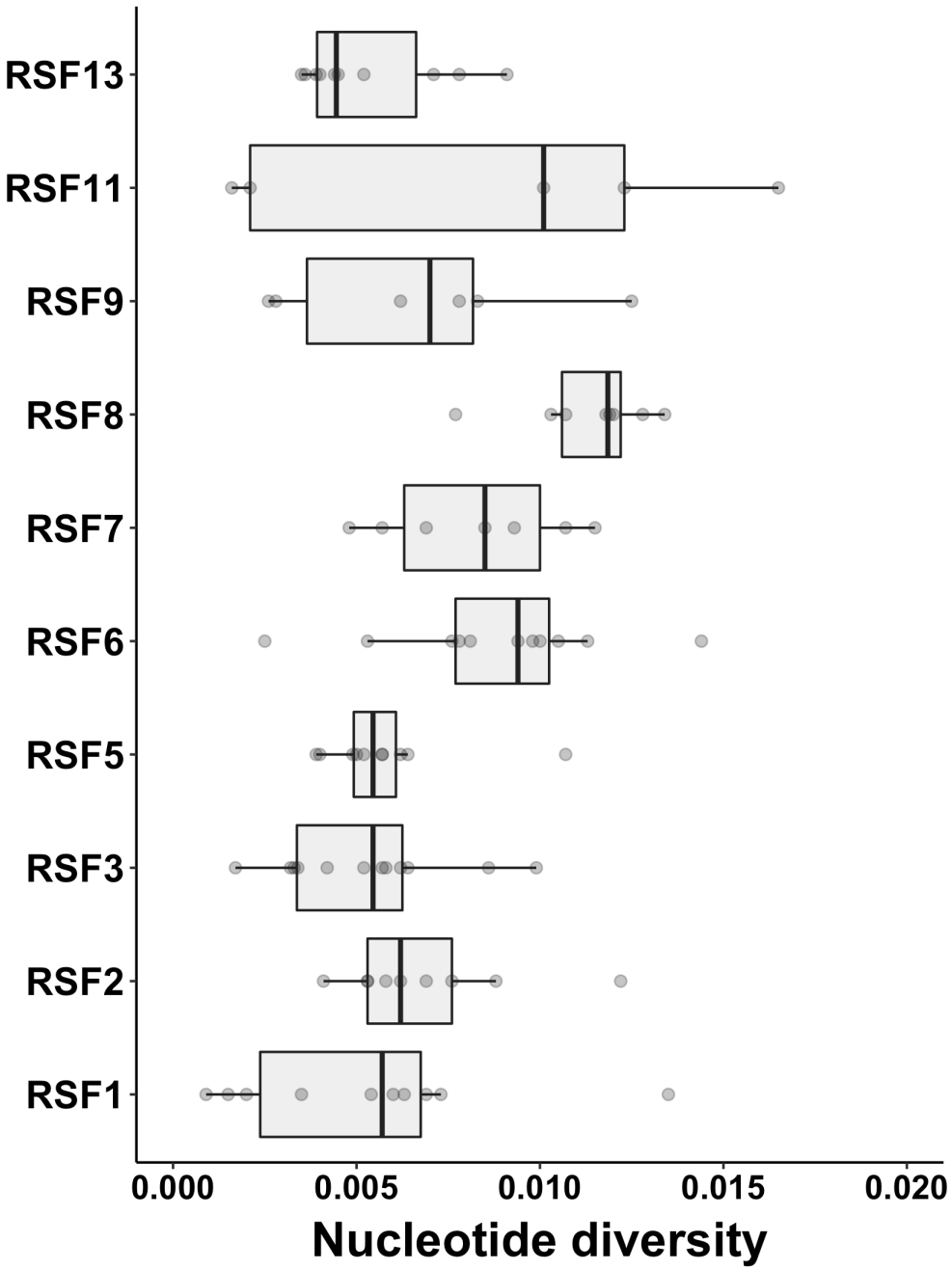
Supplementary Figure 2. Network analysis revealing the co-occurrence patterns among species present in the studied waterworks. The nodes were coloured according to species type (phage, purple; lineage II canonical *Nitrospira*, green; comammox, blue; AOB, red; other bacteria, yellow). A connection represents a strong proportionality (ρ>0.50 and FDR < 0.05). The size of each node is proportional to the average species log-transformed abundance. Only nodes connected with phages are shown.



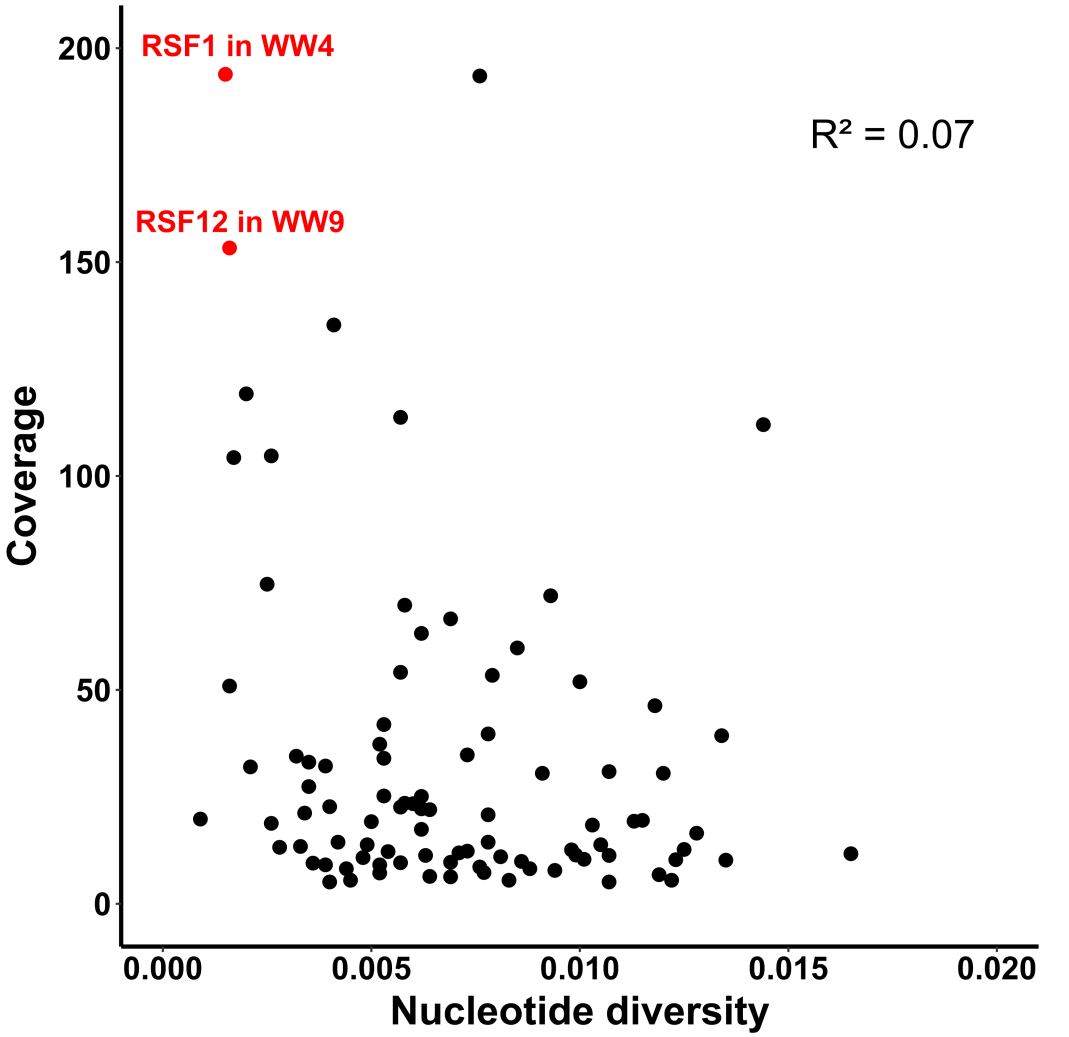
Supplementary Figure 3. Redundancy analysis (RDA) of variation in species abundance explained by variation in water chemistry variables across 12 waterworks. Arrows show the association of individual species (different colours) and water chemistry variables (red arrows) with each axis. Dots positions represent waterworks scores in relation to each axis. Top: individual species colour code: pink: comammox clade A; orange: comammox clade B; green: canonical *Nitrospira*. Bottom: blue: comammox; brown: canonical AOB; green: canonical *Nitrospira*.



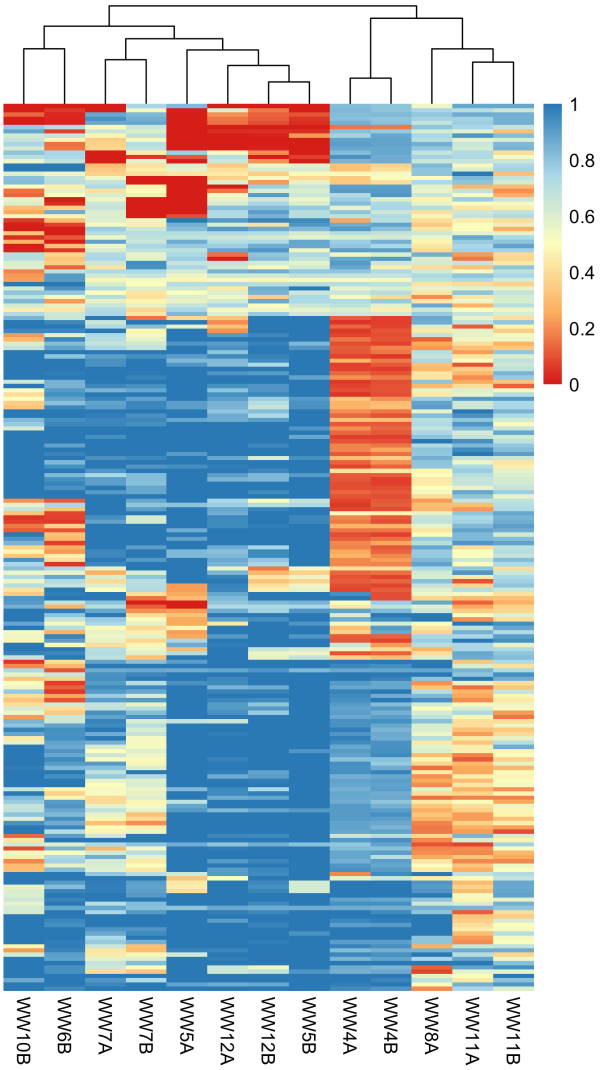
Supplementary Figure 4. Relationship between the ammonium concentration and the comammox species richness in each waterworks (number of species with abundances > 0.5%). Blue line shows the linear regression with shadowed region indicating 95% confidence intervals for the slope (p value for $R^2 < 0.01$).

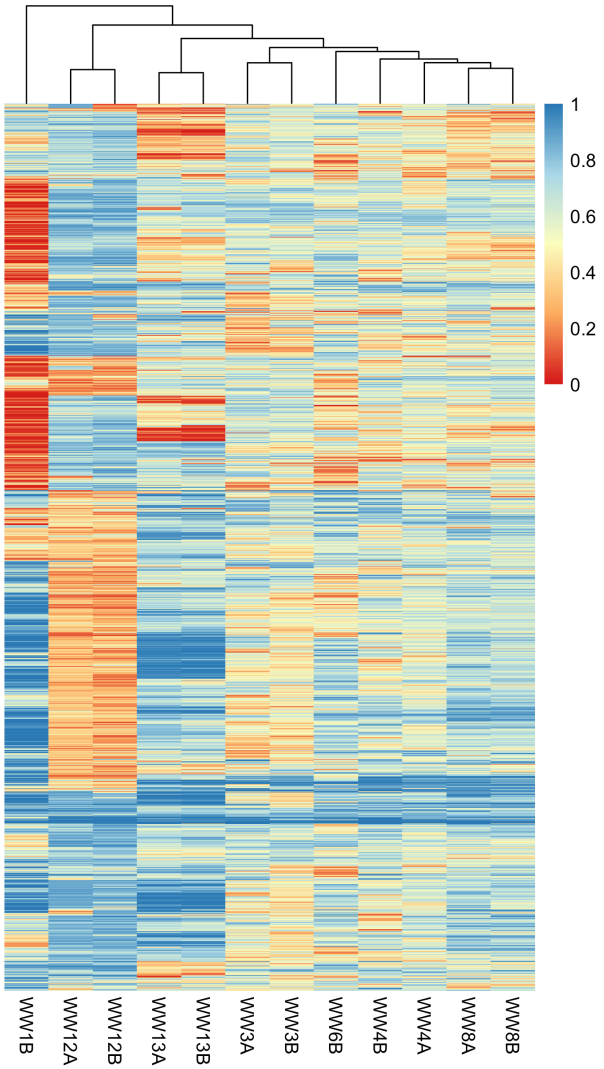


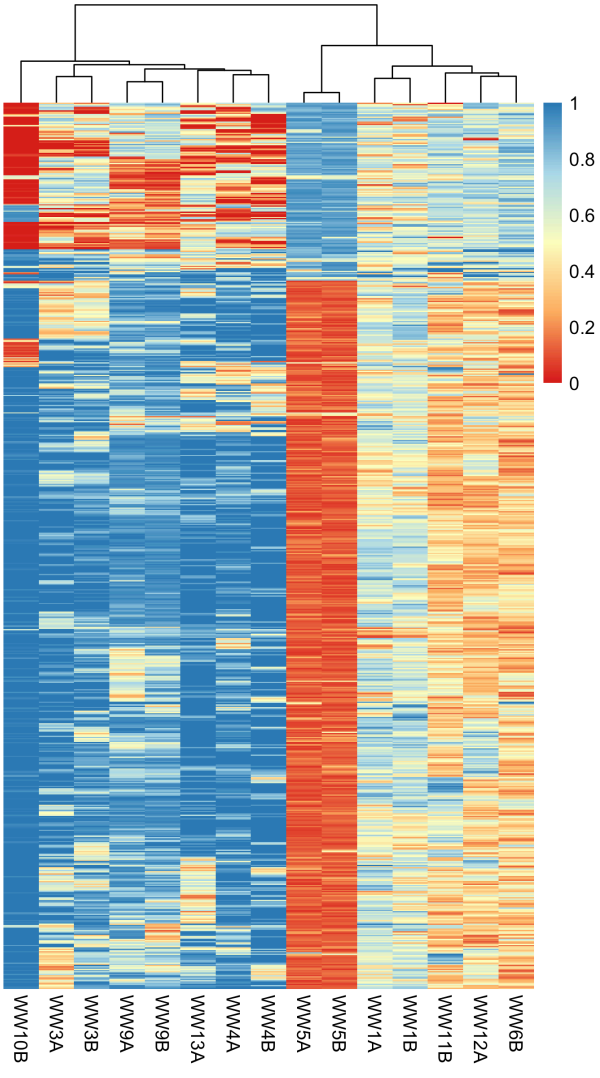
Supplementary Figure 5. Relationship between the nucleotide diversity of the studied comammox *Nitrospira* populations and their coverage in each waterworks. Each dot represents a sample-genome mapping pairing, with the mean nucleotide diversity of the genome in that sample plotted against the mean coverage of the genome in that sample. Red colour

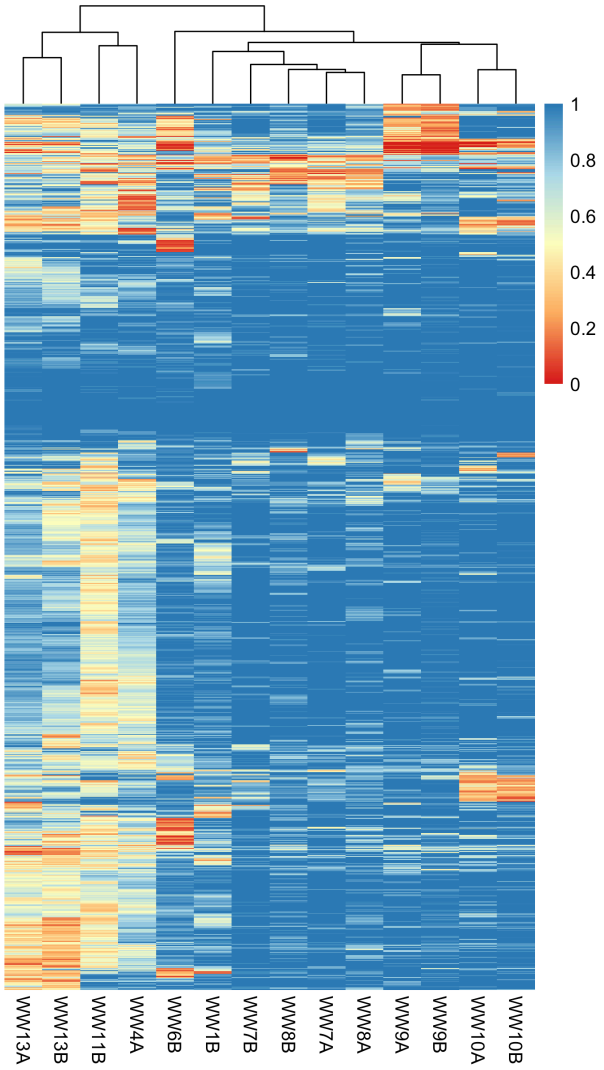


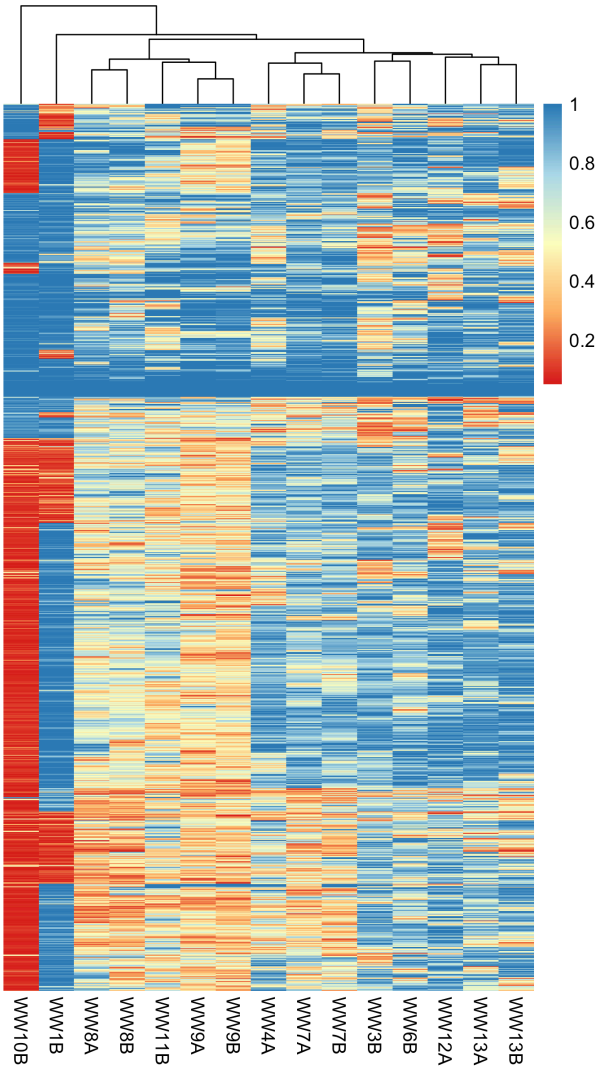
Supplementary Figure 6. The relationship between nucleotide diversity and coverage for each *Nitrospira* population in each waterworks. The mean nucleotide diversity and the mean coverage across species in each waterworks is shown. Local populations with high coverage and low nucleotide diversity are coloured in red. (linear regression; $R2 = 0.07$; $p > 0.01$).

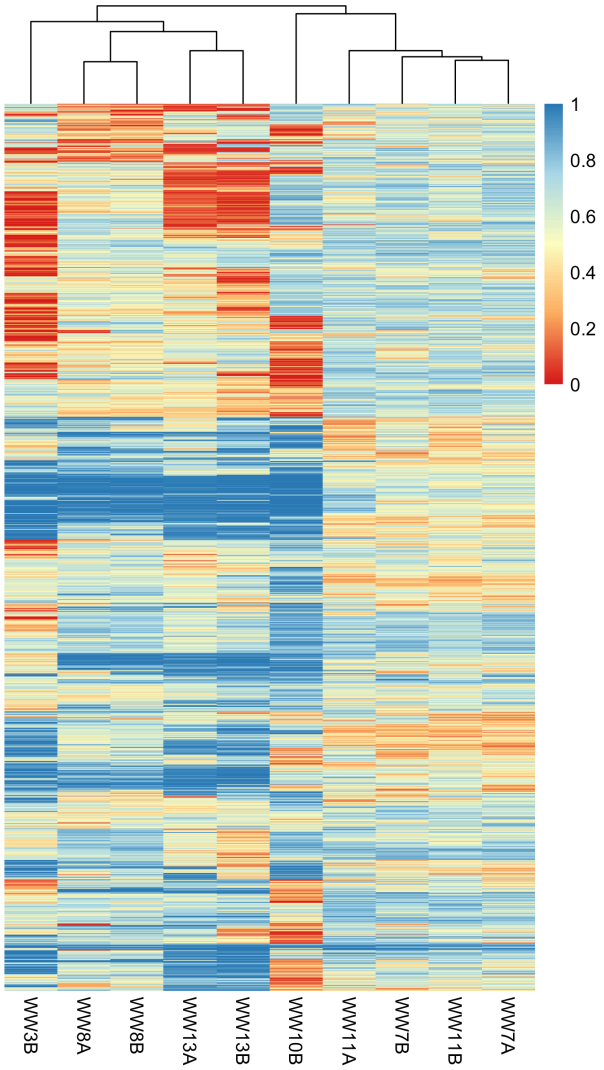

1
0.8
0.6
0.4
0.2
0
WW10B
WW6B
WW7A
WW7B
WW5A
WW12A
WW12B
WW5B
WW4A
WW4B
WW8A
WW11A
WW11B

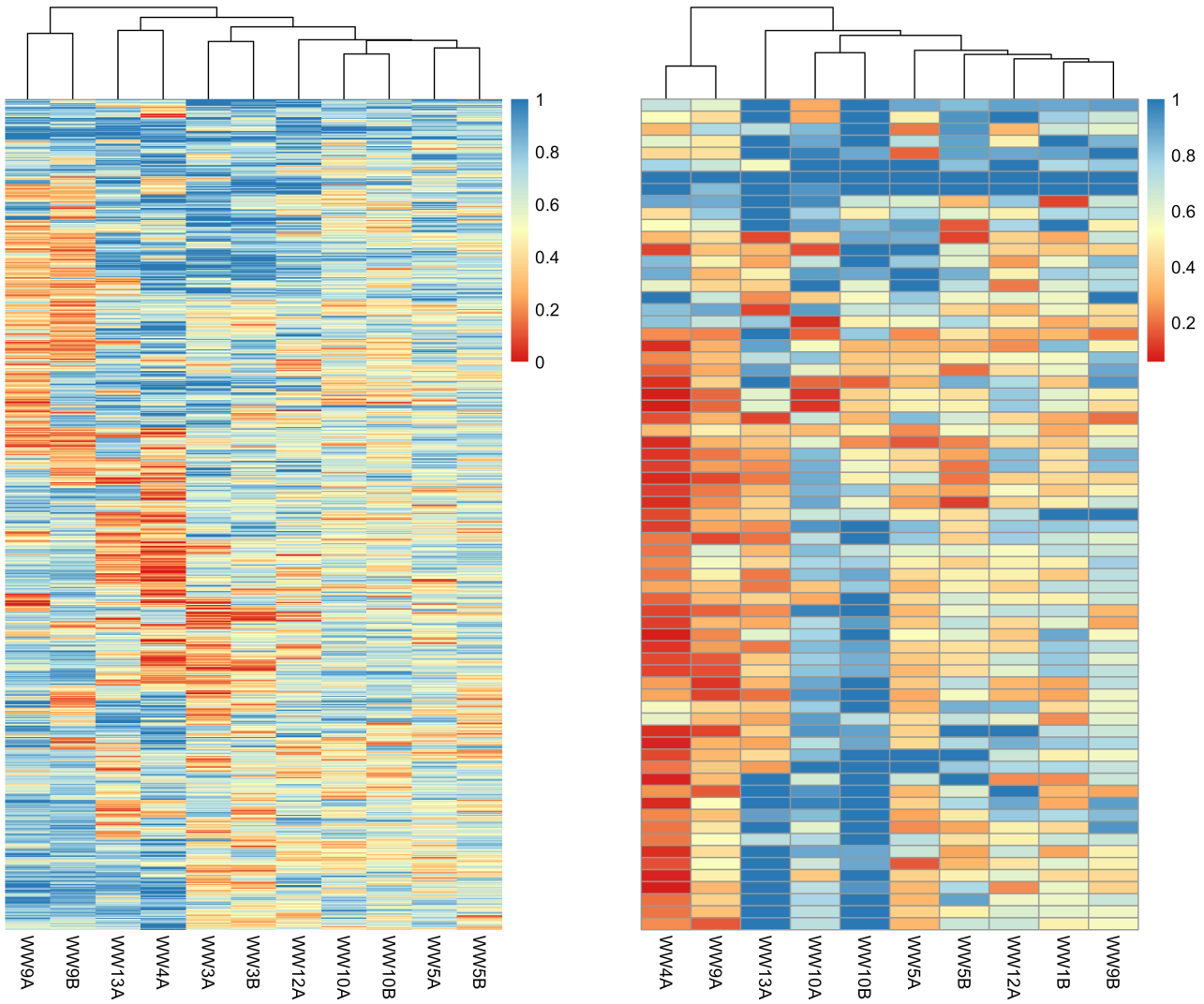

1
0.8
0.6
0.4
0.2
0
WW1B
WW12A
WW12B
WW13A
WW13B
WW3A
WW3B
WW6B
WW4B
WW4A
WW8A
WW8B


1
0.8
0.6
0.4
0.2
0
WW10B
WW3A
WW3B
WW9A
WW9B
WW13A
WW4A
WW4B
WW5A
WW5B
WW1A
WW1B
WW11B
WW12A
WW6B

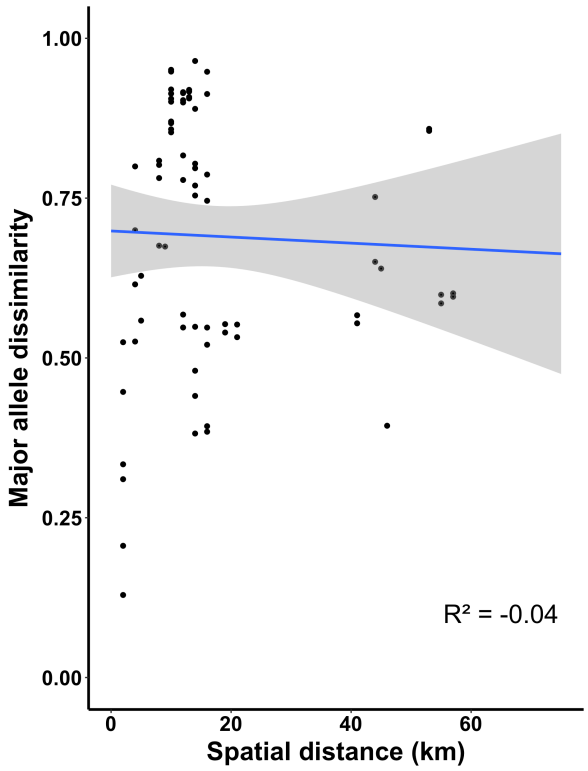

1
0.8
0.6
0.4
0.2
0
WW13A
WW13B
WW11B
WW4A
WW6B
WW1B
WW7B
WW8B
WW7A
WW8A
WW9A
WW9B
WW10A
WW10B

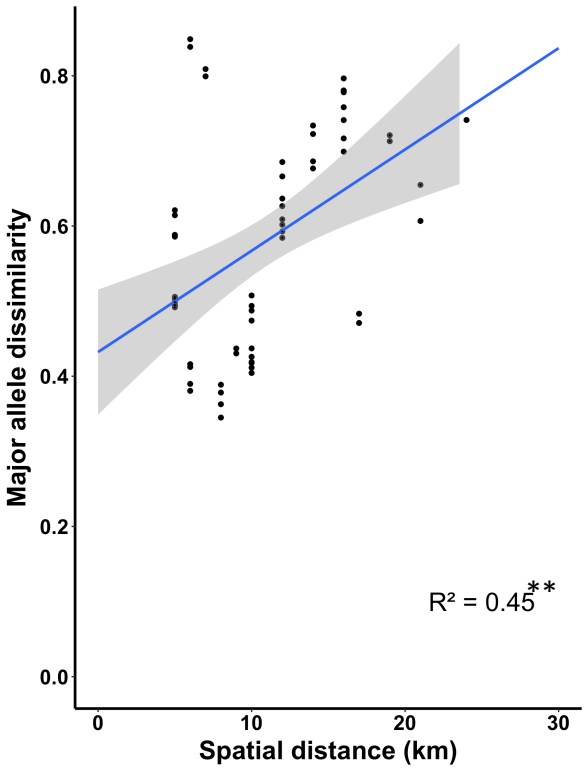

1
0.8
0.6
0.4
0.2
WW10B
WW1B
WW8A
WW8B
WW11B
WW9A
WW9B
WW4A
WW7A
WW7B
WW3B
WW6B
WW12A
WW13A
WW13B

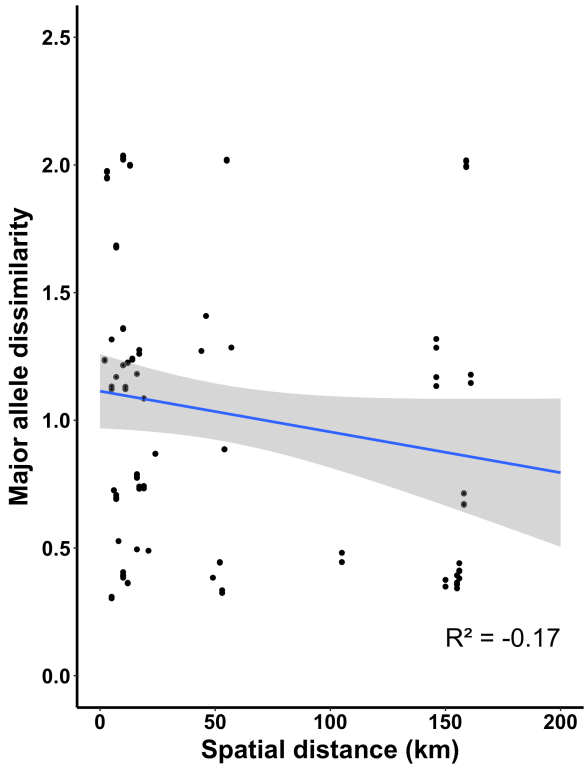

1
0.8
0.6
0.4
0.2
0
WW3B
WW8A
WW8B
WW13A
WW13B
WW10B
WW11A
WW7B
WW11B
WW7A

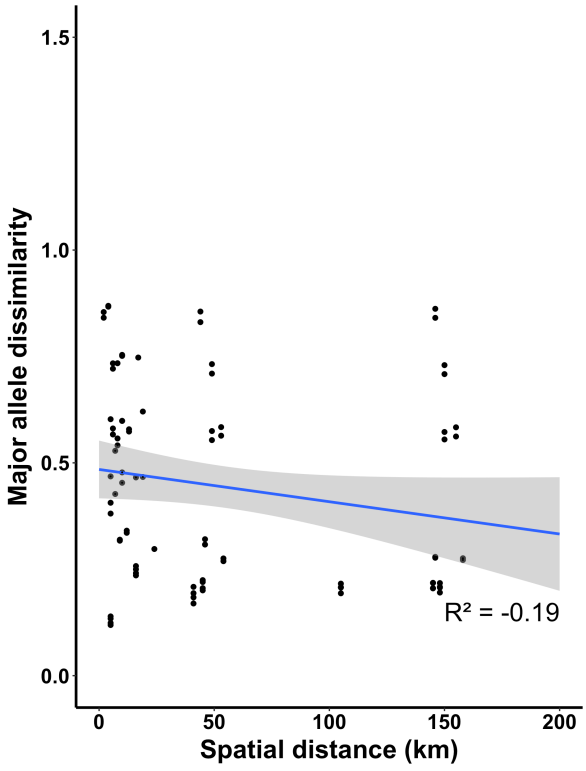


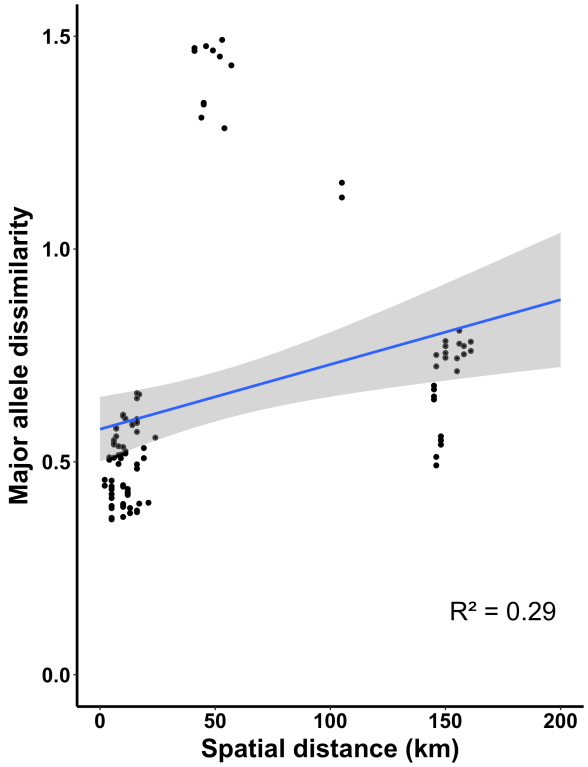
Supplementary Figure 7. Major allele frequency of common SNPs across the studied waterworks (the ones where RSF has a coverage > 5) for *Nitrospira* population. (This analysis is done for all the RSF with a coverage > 5 in at least 10 samples). Dendrogram is built based on Euclidean distance. Rows represent each SNP and columns represent each waterworks. The order of the plots from left to right and top to bottom corresponds to: RSF1, RSF2, RSF3, RSF5, RSF6, RSF7, RSF8, RSF13.

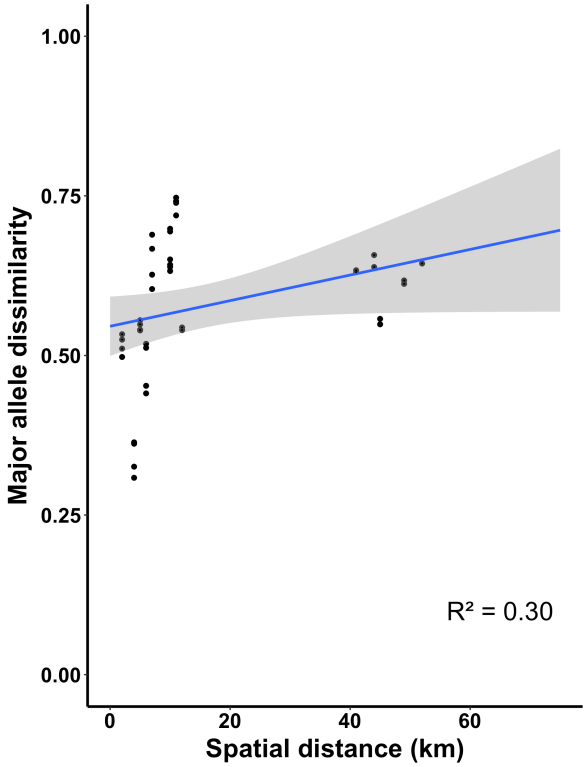

Major allele dissimilarity
1.00
0.75
0.50
0.25
0.00
0
20
40
60
Spatial distance (km)
R² = -0.04

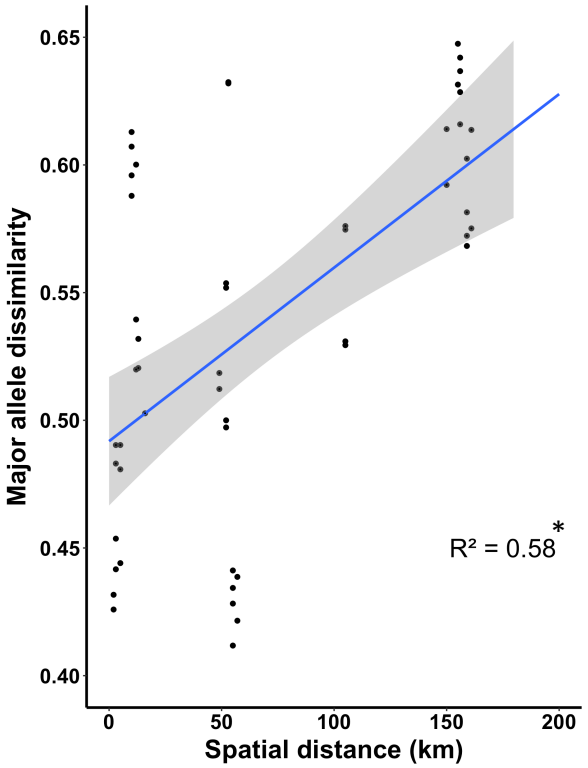

Major allele dissimilarity
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Spatial distance (km)
R² = 0.45**

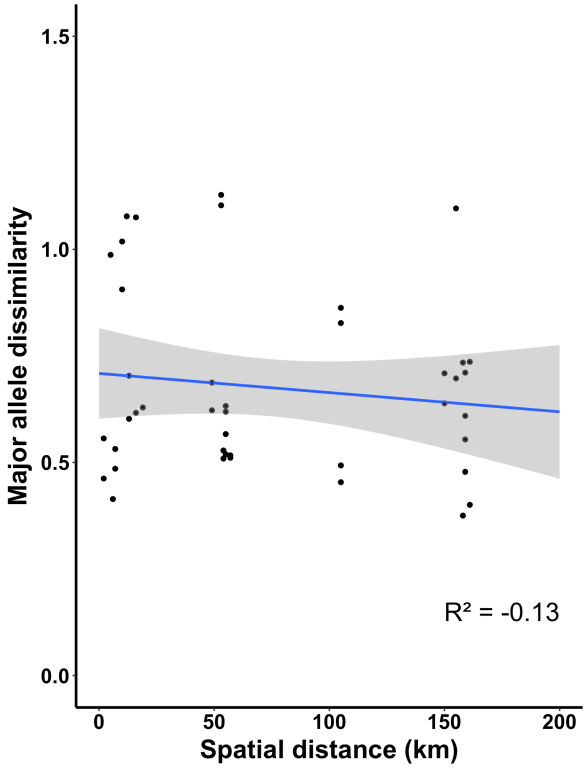

Major allele dissimilarity
2.5
2.0
1.5
1.0
0.5
0.0
0
50
100
150
200
Spatial distance (km)
R² = -0.17


Major allele dissimilarity
1.5
1.0
0.5
0.0
0
50
100
150
200
Spatial distance (km)
R² = -0.19

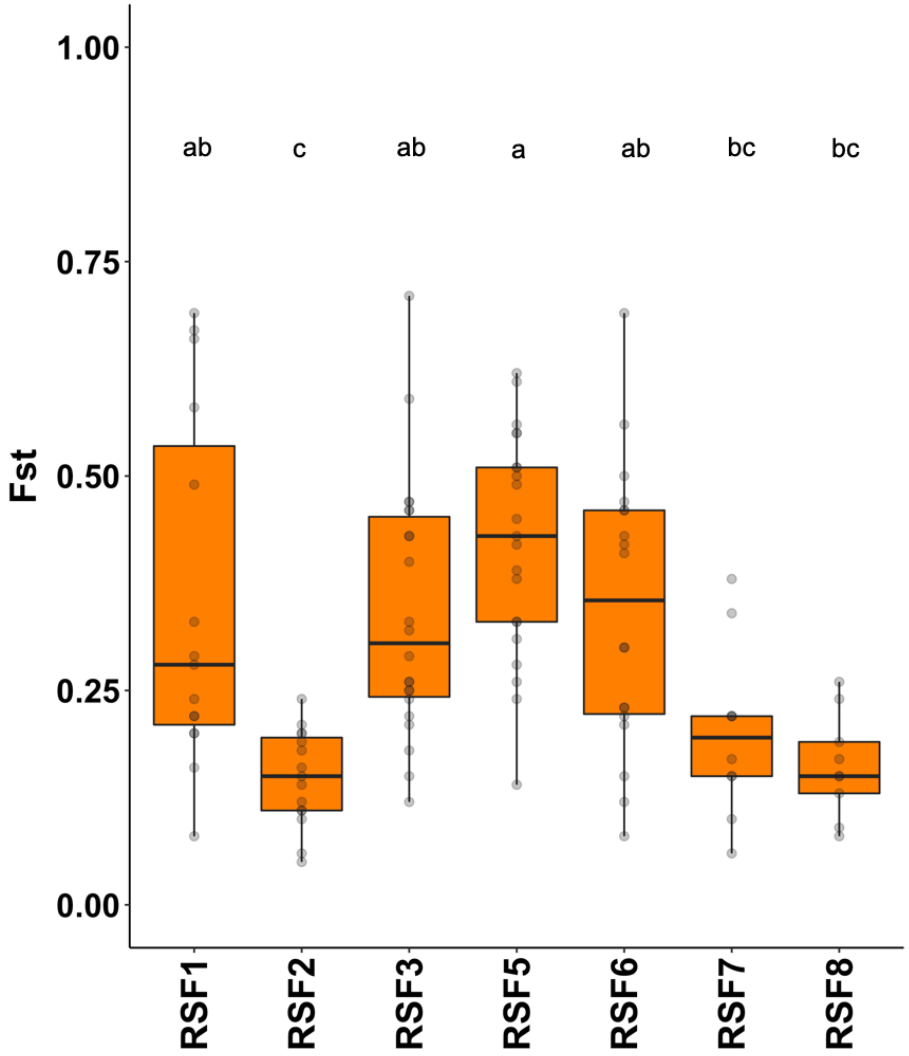

Major allele dissimilarity
1.5
1.0
0.5
0.0
0
50
100
150
200
Spatial distance (km)
R² = 0.29


Major allele dissimilarity
1.00
0.75
0.50
0.25
0.00
0
20
40
60
Spatial distance (km)
R² = 0.30

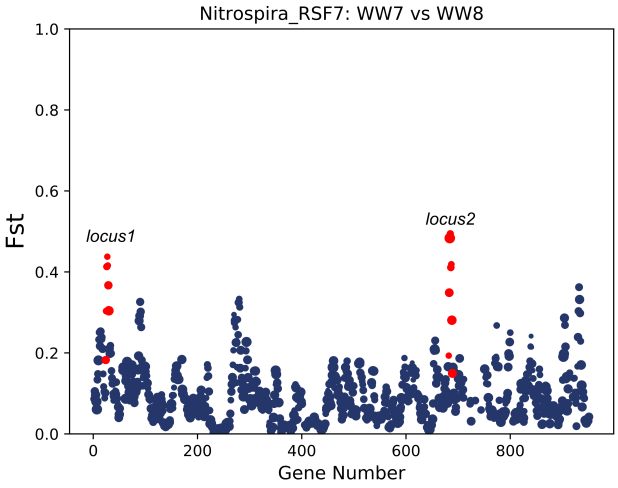

Major allele dissimilarity
0.65
0.60
0.55
0.50
0.45
0.40
0
50
100
150
200
Spatial distance (km)
R² = 0.58*

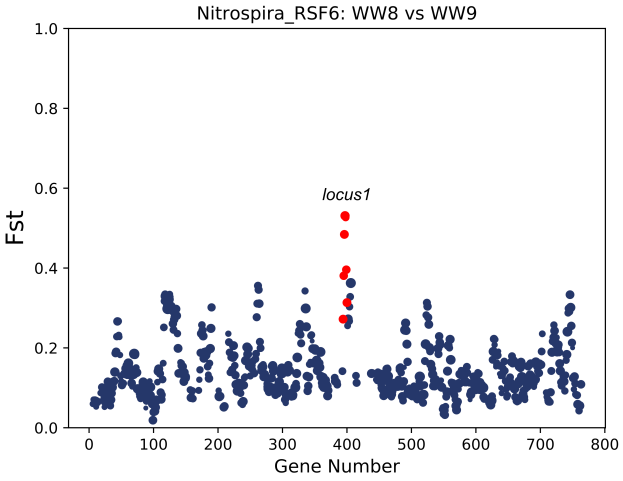

Major allele dissimilarity
1.5
1.0
0.5
0.0
0
50
100
150
200
Spatial distance (km)
R² = -0.13

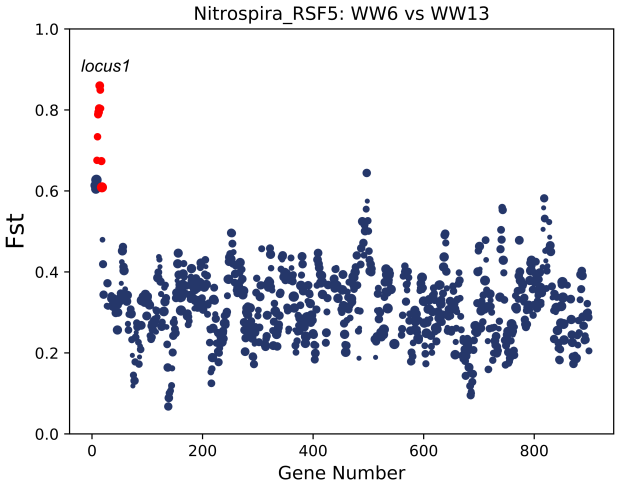
Supplementary Figure 8. Relationship between waterworks dissimilarity based on major allele frequency of common SNPs and the geographic distance of the waterworks. The dissimilarities between populations in pairs of waterworks are calculated using the Jaccard index from a matrix of major allele frequencies for common SPNs across the waterworks: the value 0 means that the two population in the two waterworks have the same allele profile. The Mantel test was used to test the strength and significance of correlations (R2 denotes the Mantel statistic *r;* * denotes $p < 0.05$ and ** denotes $p < 0.01$). Blue line shows the linear regression with shadowed region indicating 95% confidence intervals for the slope. The order of the plots from left to right and top to bottom corresponds to: RSF1, RSF2, RSF3, RSF5, RSF6, RSF7, RSF8, RSF13.

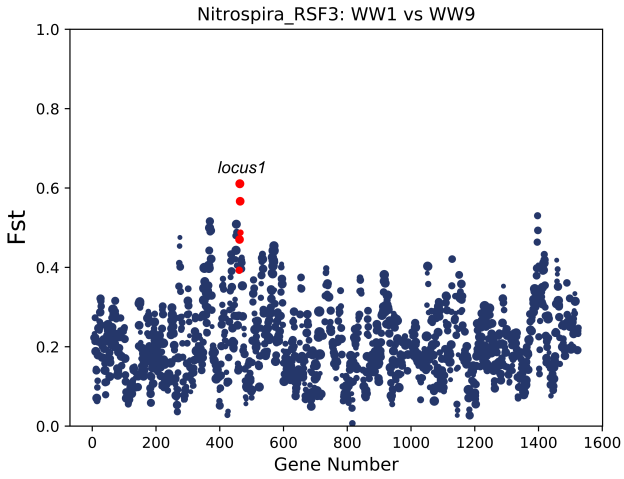


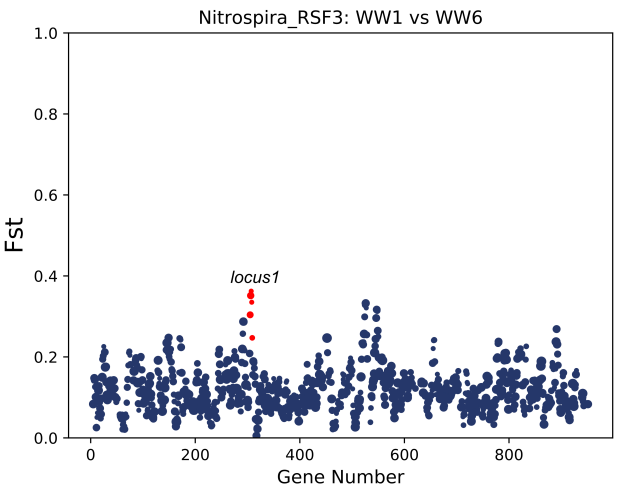
Supplementary Figure 9. Boxplot showing the F_{ST} values for each species (each dot represents the F_{ST} value measured as the differences in allele frequencies between populations of the same species found in two distinct waterworks). Only species with more than two data points are shown. Differences between the mean F_{ST} were assessed by a Dunn's test; same letter have means not significantly different from each other ($p < 0.01$).

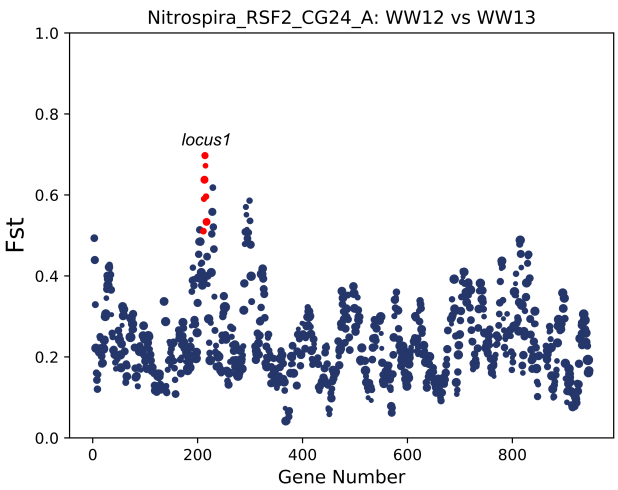

Nitrospira_RSF7: WW7 vs WW8
Fst
Gene Number
locus1
locus2

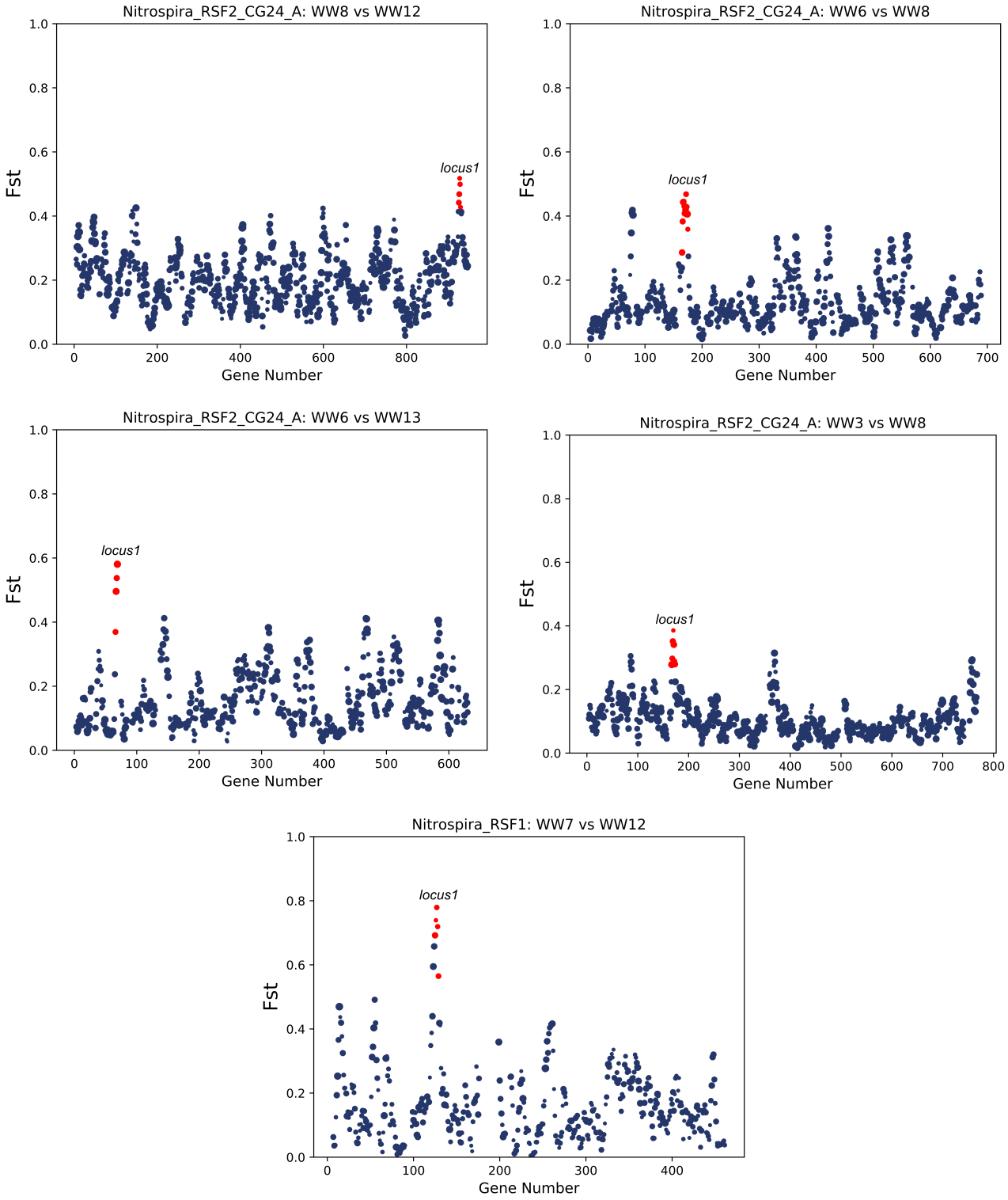

Nitrospira_RSF6: WW8 vs WW9
Fst
Gene Number
locus1


Nitrospira_RSF5: WW6 vs WW13
Fst
Gene Number
locus1

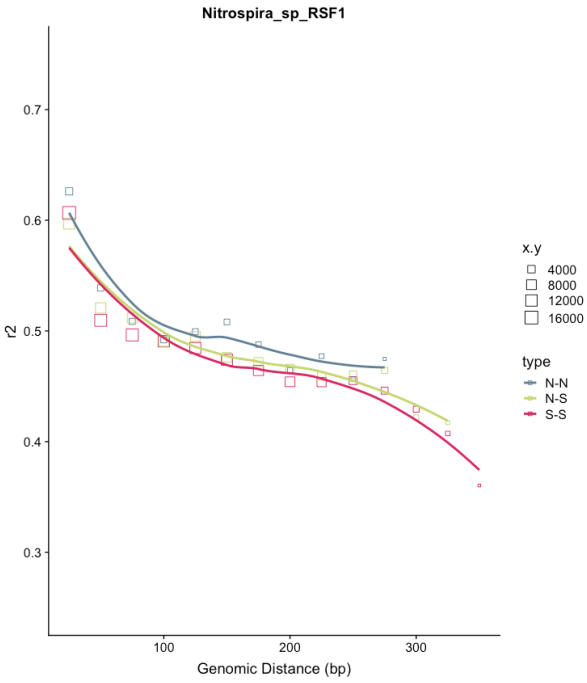

Nitrospira_RSF3: WW1 vs WW9
Fst
Gene Number
locus1

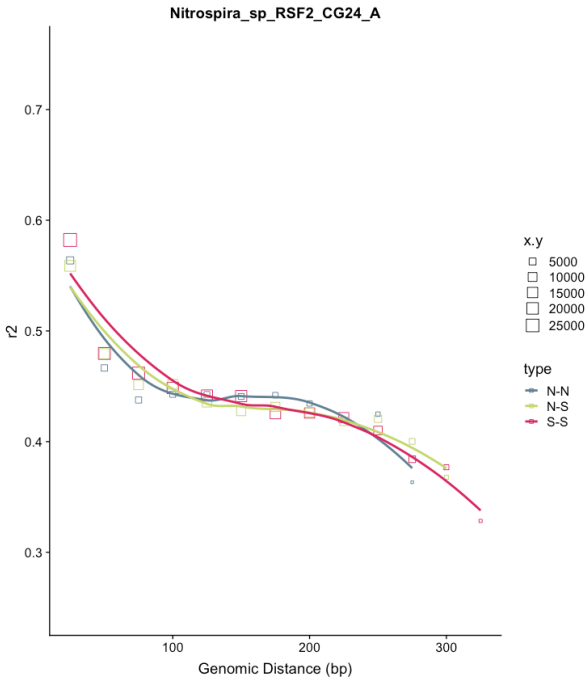

Nitrospira_RSF3: WW1 vs WW6
Fst
Gene Number
locus1

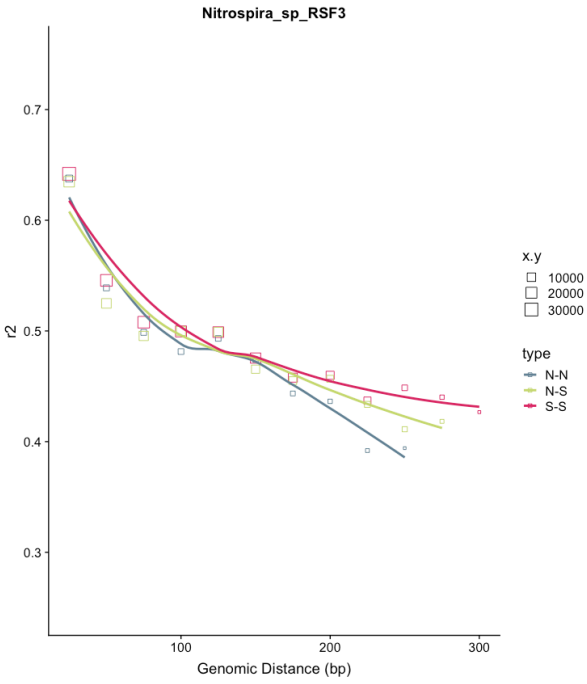

Nitrospira_RSF2_CG24_A: WW12 vs WW13
Fst
Gene Number
locus1

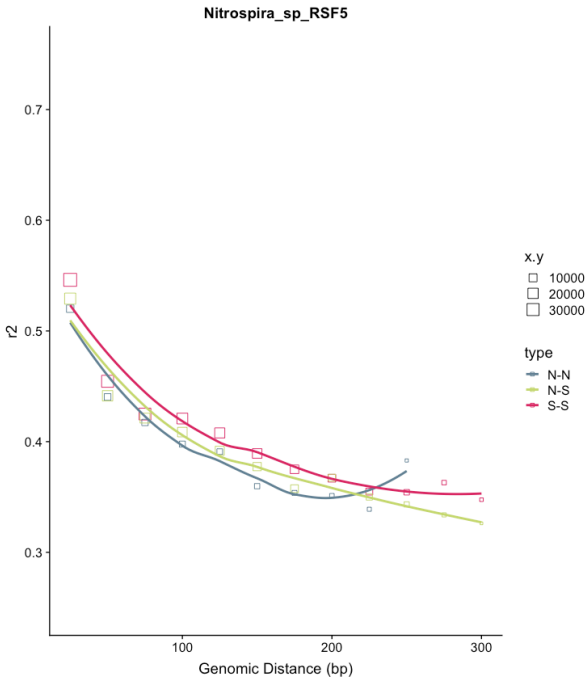


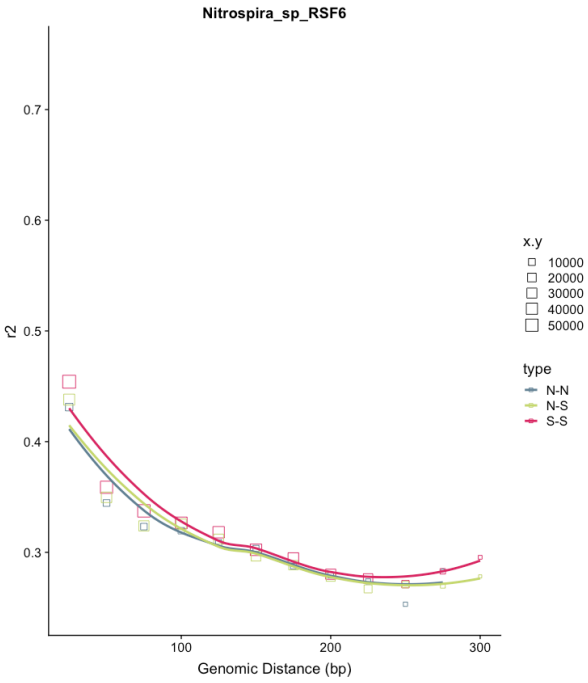
Supplementary Figure 10. Highly different genomic loci between waterworks. Values of F_{ST} for genes across the genomes of the bacterial populations. Each point is a gene, and the size of the point is determined by the number of SNPs within that gene. Loci with significantly higher F_{ST} than the background are highlighted in red. (Here are shown all the comparisons containing loci with significantly higher $F_{ST,}$ and one example without it). These plots were produced using the script provided by Crits-Christoph *et al*. (2020)[1].

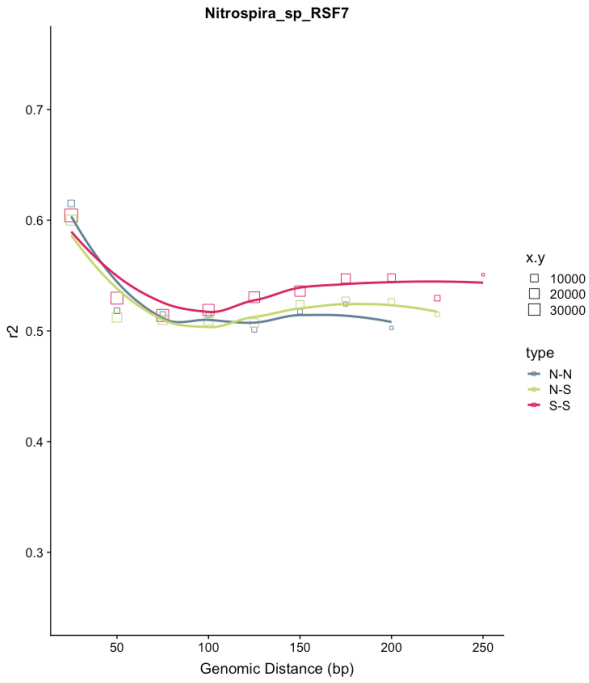

Nitrospira_sp_RSF1
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
x.y
4000
8000
12000
16000
type
N-N
N-S
S-S

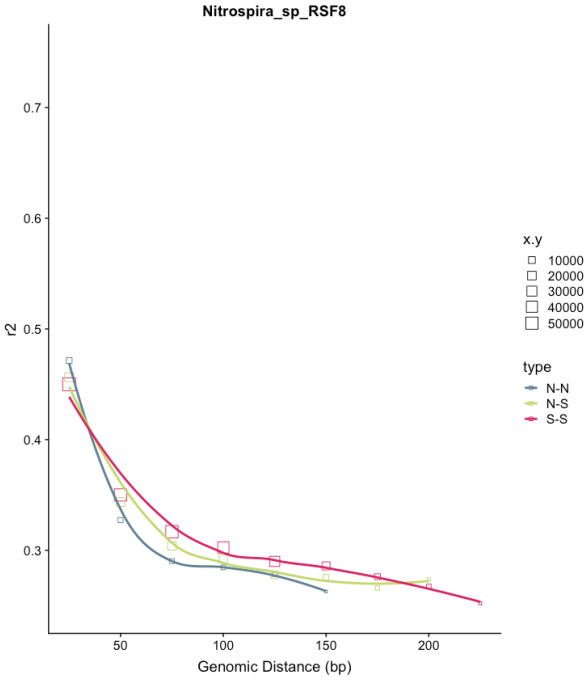

Nitrospira_sp_RSF2_CG24_A
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
x.y
5000
10000
15000
20000
25000
type
N-N
N-S
S-S

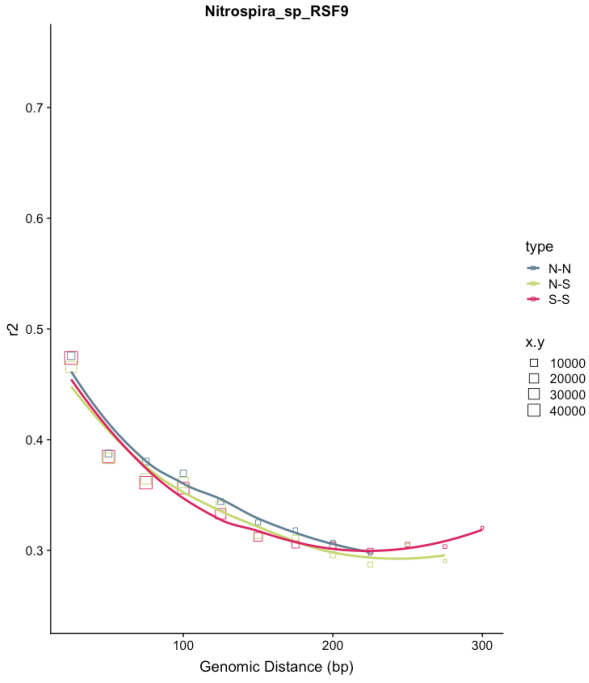

Nitrospira_sp_RSF3
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
x.y
10000
20000
30000
type
N-N
N-S
S-S

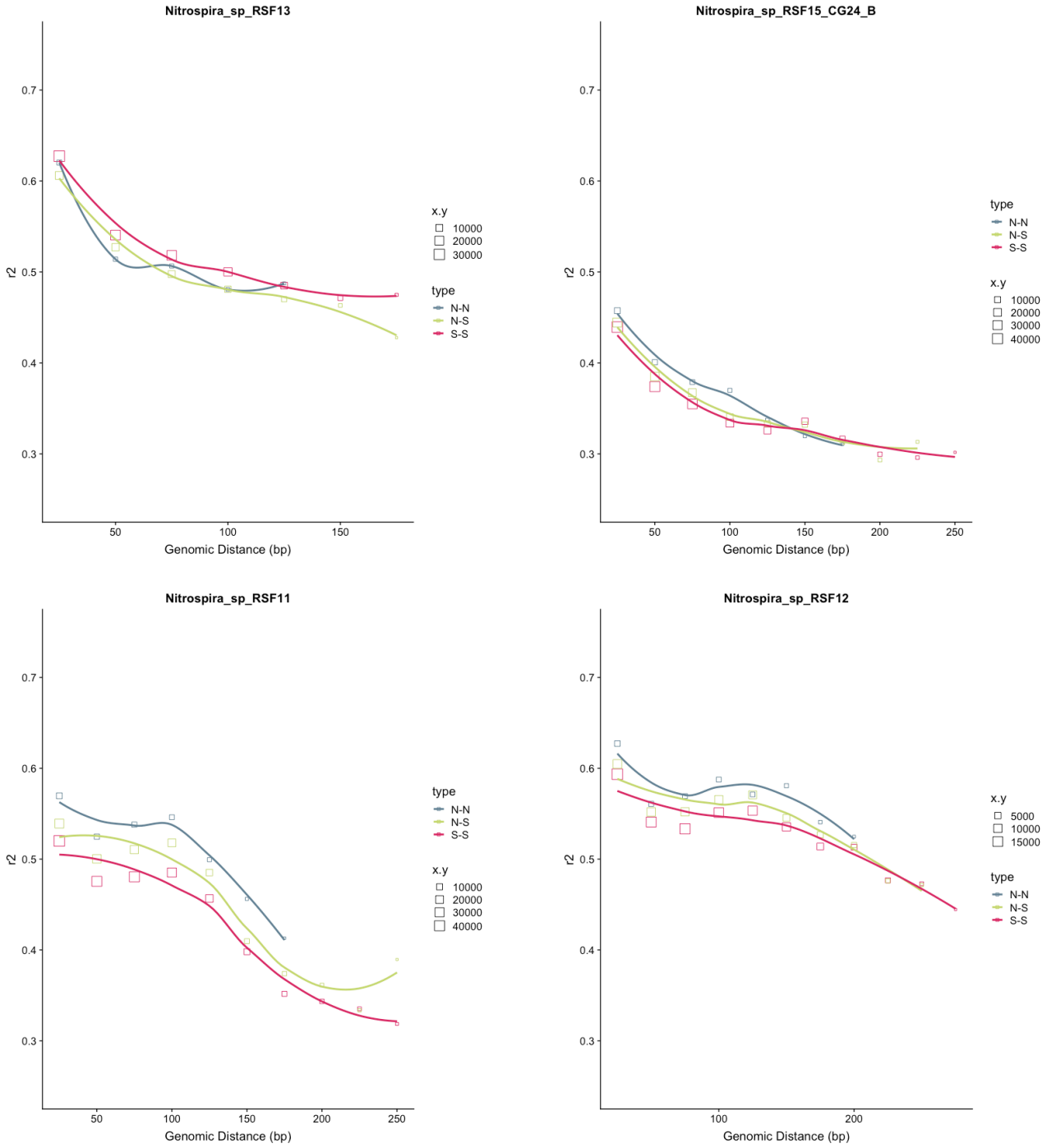

Nitrospira_sp_RSF5
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
x.y
10000
20000
30000
type
N-N
N-S
S-S


Nitrospira_sp_RSF6
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
x.y
10000
20000
30000
40000
50000
type
N-N
N-S
S-S

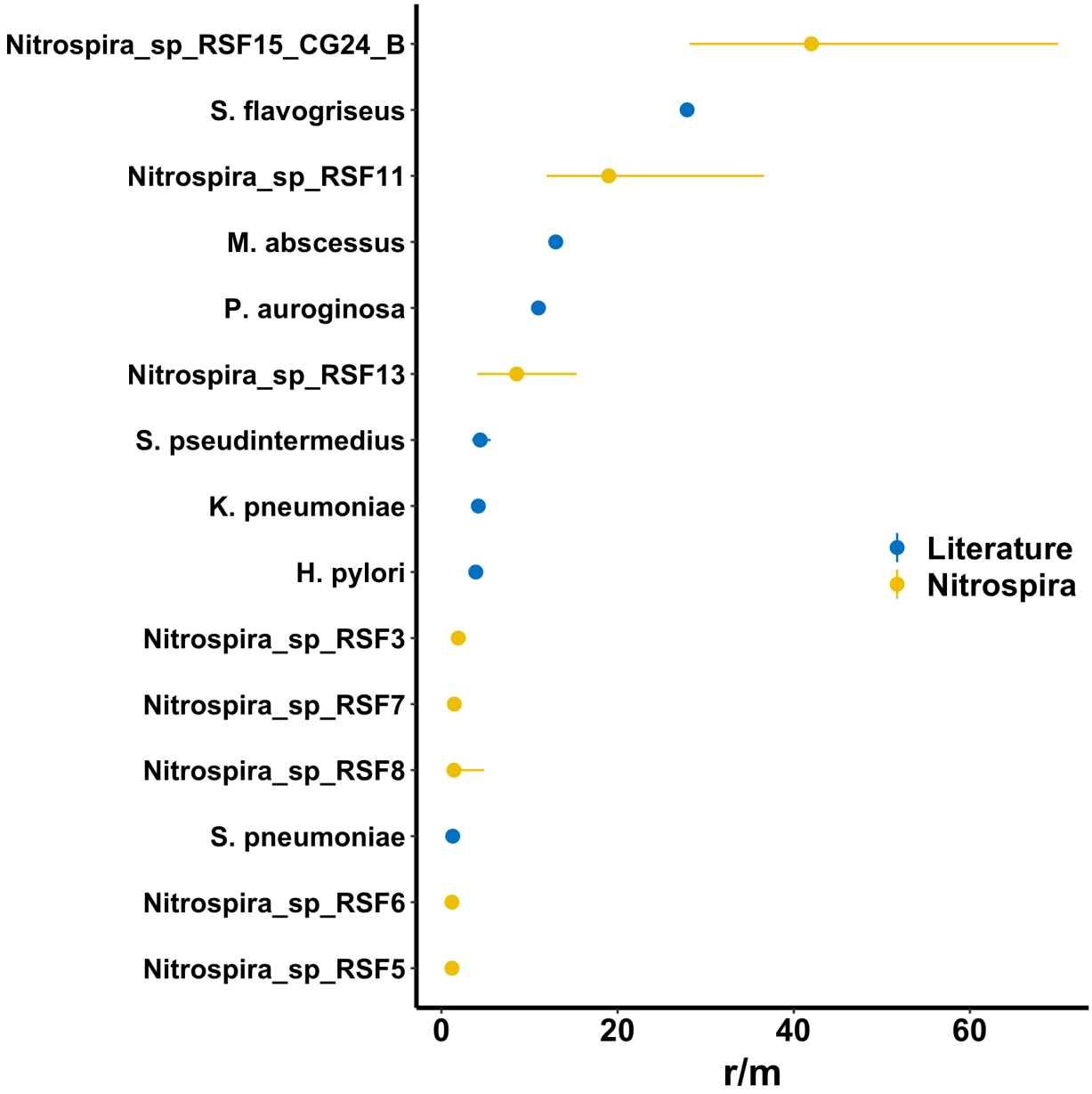

Nitrospira_sp_RSF7
r2
0.7
0.6
0.5
0.4
0.3
50
100
150
200
250
Genomic Distance (bp)
x.y
10000
20000
30000
type
N-N
N-S
S-S


Nitrospira_sp_RSF8
r2
0.7
0.6
0.5
0.4
0.3
50
100
150
200
Genomic Distance (bp)
x.y
10000
20000
30000
40000
50000
type
N-N
N-S
S-S

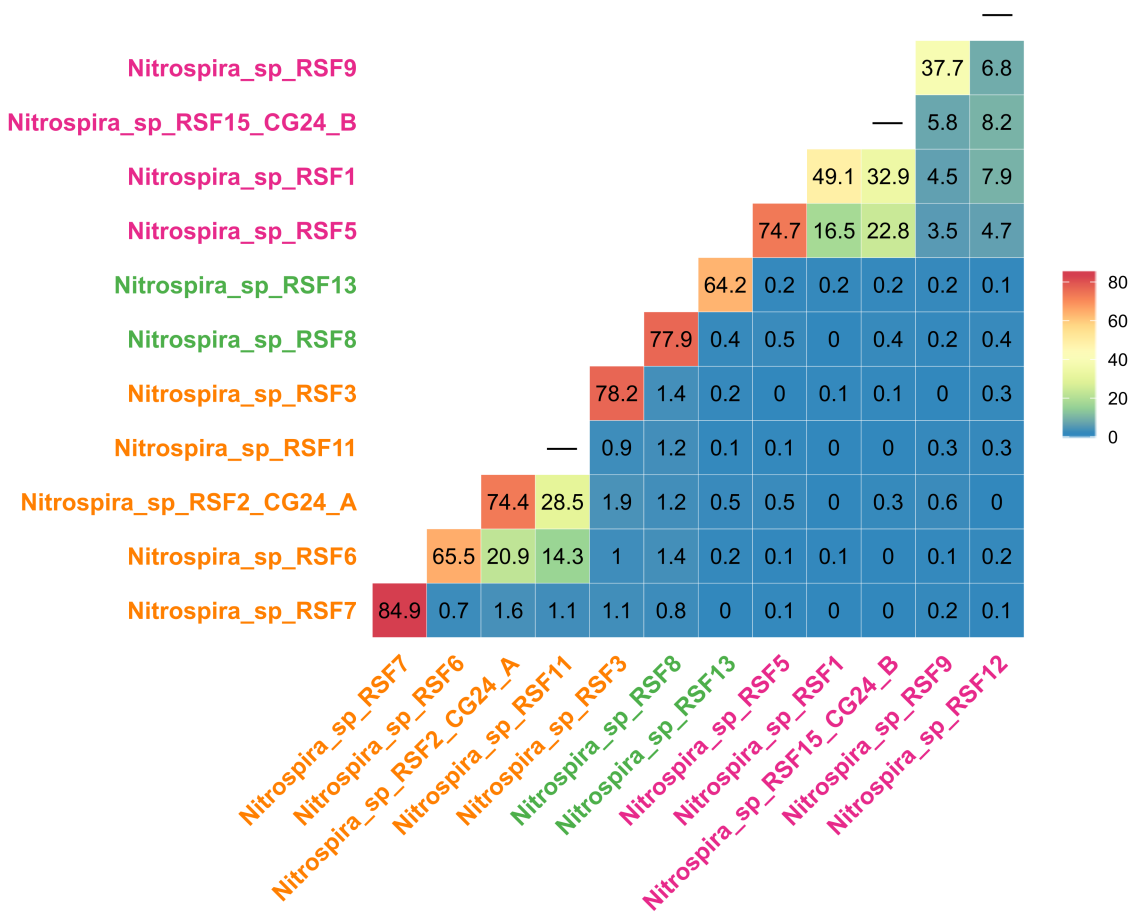

Nitrospira_sp_RSF9
r2
0.7
0.6
0.5
0.4
0.3
100
200
300
Genomic Distance (bp)
type
N-N
N-S
S-S
x.y
10000
20000
30000
40000



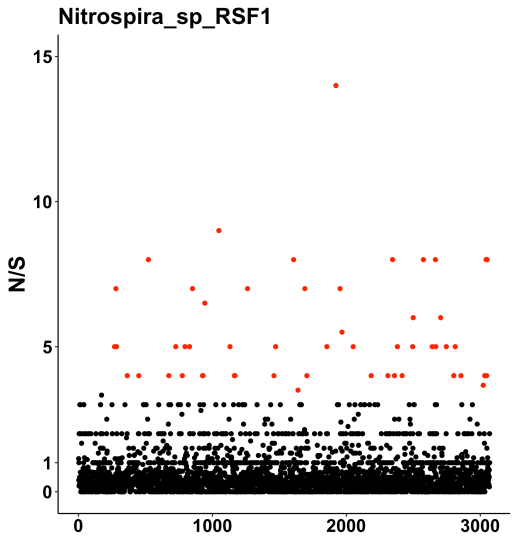
Supplementary Figure 11. Linkage decay of homologous recombination (r^2) for pairs of loci within the *Nitrospira* populations. Each square is an average of pairs of biallelic sites at that distance, with the area of the square point proportional to the number of pairs of biallelic sites that went into the mean. Haplotypes (site pairs) are binned by the predicted function of the mutations of each of the paired SNPs (nonsynonymous: N, synonymous: S). These plots were produced using the script provided by Crits-Christoph *et al*. (2020)[1].

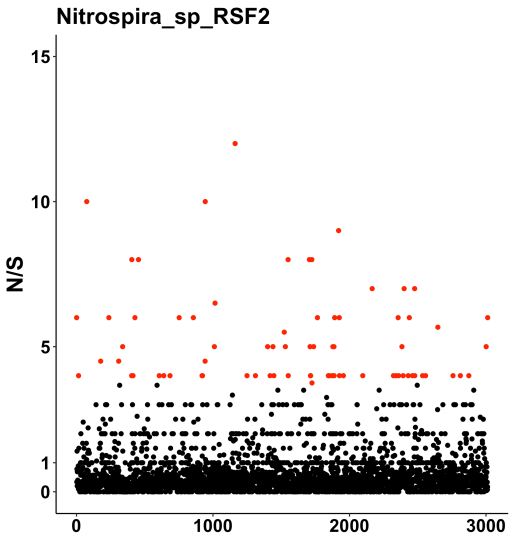


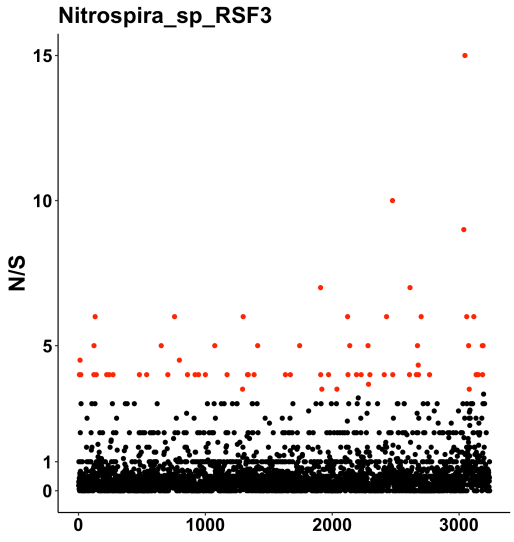
Supplementary Figure 12. Relative rates of recombination to mutation (*r/m*) calculated across the 12 waterworks for the *Nitrospira* populations (yellow) on synonymous third position codon sites, compared with previous values (blue) reported by Lin and Kussell[2], the value reported for Streptomyces flavogensis by Doroghazi and Buckley[3], and the value reported for Staphylococcus pseudintermedius by Smith et al. (2020)[4]. Error bars represent the 95% confidence interval across 1000 bootstraps.

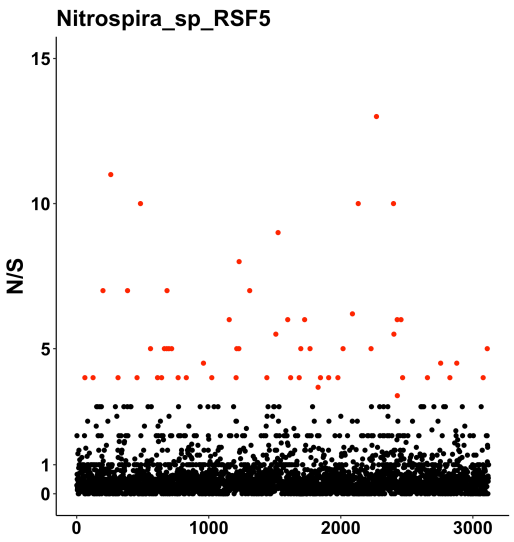


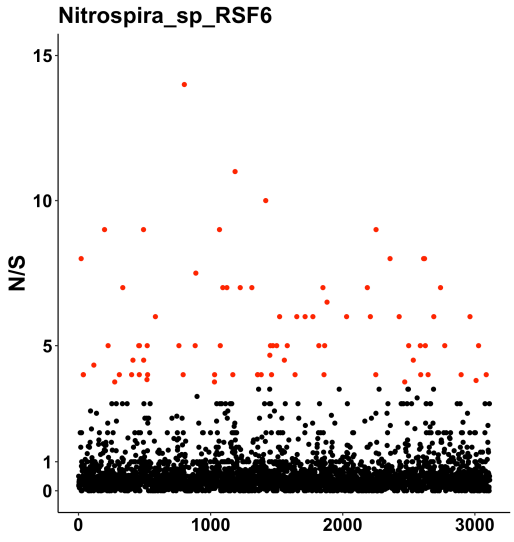
Supplementary Figure 13. Heatmap showing the percentage of single-copy orthologous genes phylogenetic trees which do not support the species phylogenetic tree topology. The analysis was done using tree quartets (four from the same species for within same species analysis; and two from one species and two from another one for between species analysis). Lineages and sublineages are denoted with different colours (lineage II canonical *Nitrospira*, green; comammox clade A, pink; comammox clade B, orange).

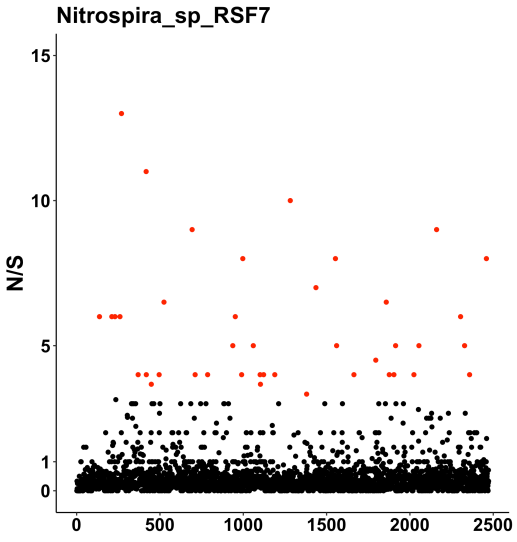

Nitrospira_sp_RSF1
N/S
15
10
5
1
0
0
1000
2000
3000

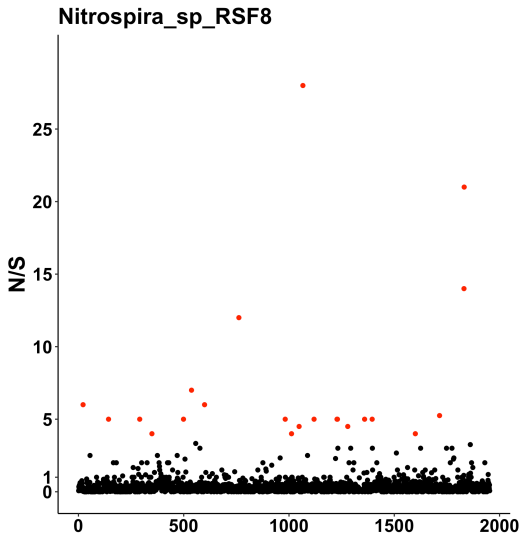

Nitrospira_sp_RSF2
N/S
15
10
5
1
0
0
1000
2000
3000

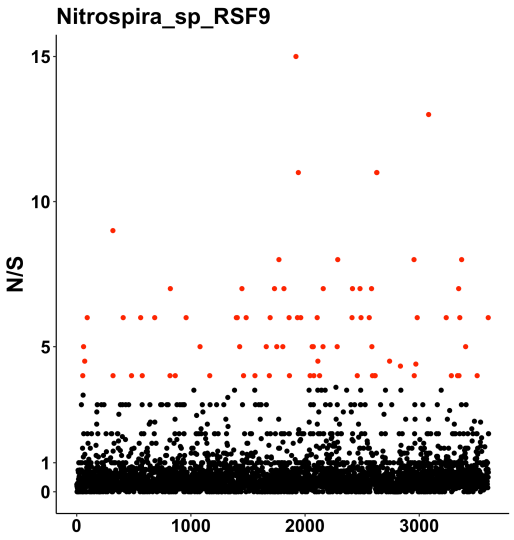

Nitrospira_sp_RSF3
N/S
15
10
5
1
0
0
1000
2000
3000

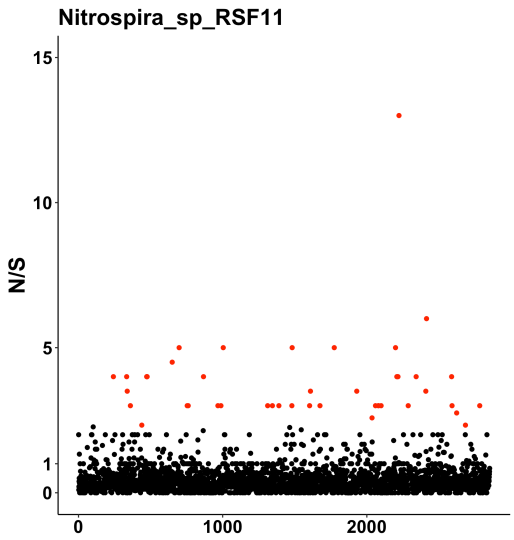

Nitrospira_sp_RSF5
N/S
15
10
5
1
0
0
1000
2000
3000

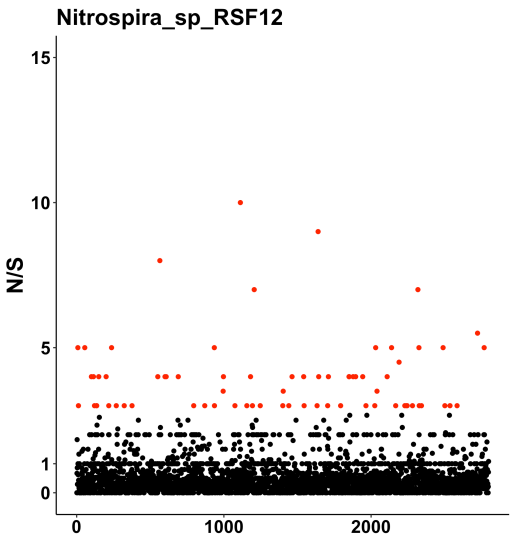

Nitrospira_sp_RSF6
N/S
15
10
5
1
0
0
1000
2000
3000

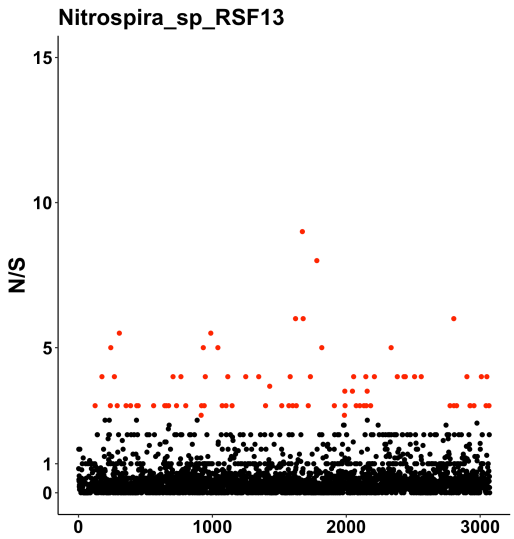

Nitrospira_sp_RSF7
N/S
15
10
5
1
0
0
500
1000
1500
2000
2500

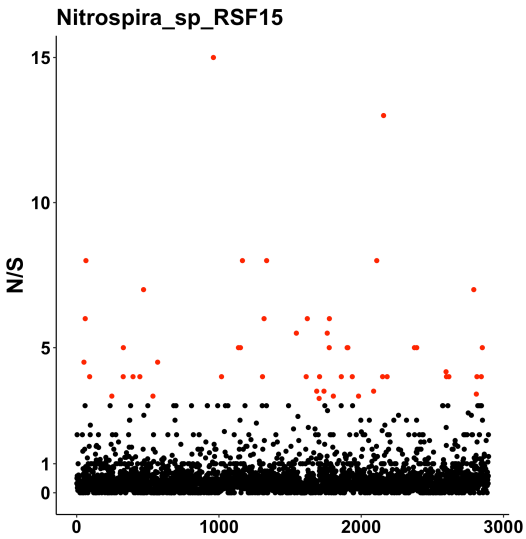

Nitrospira_sp_RSF8
N/S
25
20
15
10
5
1
0
0
500
1000
1500
2000


Nitrospira_sp_RSF9
N/S
15
10
5
1
0
0
1000
2000
3000


Nitrospira_sp_RSF11
N/S
15
10
5
1
0
0
1000
2000


Nitrospira_sp_RSF12
N/S
15
10
5
1
0
0
1000
2000


Nitrospira_sp_RSF13
N/S
15
10
5
1
0
0
1000
2000
3000


Nitrospira_sp_RSF15
N/S
15
10
5
1
0
0
1000
2000
3000

Supplementary Figure 14. Values of pN/pS ratios for genes across the genomes of 12 *Nitrospira* populations. Each point is a gene, and genes with significantly higher (more than three standard deviation than the mean) pN/pS value are highlighted in red.

References:

1. 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
2. Lin, M. & Kussell, E. Inferring bacterial recombination rates from large-scale sequencing datasets. *Nat. Methods* **16**, 199–204 (2019).
3. Doroghazi, J. R. & Buckley, D. H. Widespread homologous recombination within and between Streptomyces species. *ISME J.* **4**, 1136–1143 (2010).
4. Smith, J. T. *et al.* Population genomics of Staphylococcus pseudintermedius in companion animals in the United States. *Commun. Biol.* **3**, 282 (2020).