Supplementary Information for:

Hierarchical mode of evolution in freshwater SAR11 driven by species dispersal and lake history

Clafy Fernandes, Yusuke Okazaki, Michaela M. Salcher

This PDF file includes:

Supplementary text

Statistical analysis details.

Supplementary references.

Supplementary Figures S1 to S9:

Supplementary Fig. 1: Phylogeny of the SAR11 order.

Supplementary Fig. 2: Locations of lake metagenomes used for analyses (*n* metagenomes = 117, *n* lakes = 21).

Supplementary Fig. 3: Scatter plots showing correlations between nucleotide diversity (π) and SNV density (counts per Mbp) for each species.

Supplementary Fig. 4: Population genetic structure and differentiation of *Fontibacterium* species.

Supplementary Fig. 5: Scale-dependent patterns of Isolation-by-Distance (IBD) for the temperate specialist, *F. temperatum*.

Supplementary Fig. 6: Isolation-by-Distance (IBD) patterns for the ubiquitous species, *F. commune*, reveal global connectivity.

Supplementary Fig. 7: Allele Frequency Spectra (AFS) for (a) *F. temperatum*, (b) *F. commune* and (c) *F. africanum* populations across freshwater lakes.

Supplementary Fig. 8: Relationship between ecosystem age and minor allele frequency (MAF) thresholds.

Supplementary Figure 9: Dynamics of purifying selection across the wide range of allele frequencies for (a) the temperate specialist (*F. temperatum*), (b) the ubiquitous species (*F. commune*), and (c) the endemic species (*F. africanum*).

Captions for Supplementary Data 1 to 10.

Supplementary Data 1: Details for newly sequenced genomes.

Supplementary Data 2: List of publicly available and new genomes used for phylogenomic tree reconstruction.

Supplementary Data 3: Details on freshwater metagenomes used in this study.

Supplementary Data 4: InStrain genome-wide metric output for each of the three *Fontibacterium* species.

Supplementary Data 5: Statistical tests for nucleotide diversity among the three *Fontibacterium* species.

Supplementary Data 6: Fixation index (F_{ST}) matrix scores calculated for *F. temperatum* and *F. commune* and *F. africanum.*

Supplementary Data 7: Metadata for isolation by distance analysis (IBD) and related statistics based on F_{ST} scores.

Supplementary Data 8: Relationship between ecosystem age and minor allele frequency (MAF) threshold.

Supplementary Data 9: pN/pS ratios and underlying polymorphism counts for all species and lake populations.

Supplementary text:

Statistical analysis details

Nucleotide diversity across the three species

All statistical analyses of nucleotide diversity (π) were performed in R (v4.3.1)[1]. Data were grouped by species (*F. commune*, *F. temperatum*, *F. africanum*). To compare π across species, data were first assessed for normality (Shapiro-Wilk test[2]) and homogeneity of variances (Levene's test[3], car::levene). Due to normality violations, Kruskal-Wallis test[4] was used followed by a Dunn's post-hoc test with Bonferroni correction[5]. A sub-analysis comparing π between Lake Malawi and Lake Tanganyika populations of *F. africanum* was conducted using a Mann-Whitney U test[6].

Hierarchical clustering of pairwise fixation index (F_{ST})

Clustering analyses based on pairwise F_{ST} data for the two species were conducted in R (v4.3.1)[1] using the subsequent packages; tidyverse was used for data processing. Hierarchical clustering was performed on the F_{ST} distance matrix using the hclust function with the "ward.D2" method. Bootstrap support (1,000 replicates) was used to assess node reliability. The resulting clusters were confirmed by calculating the average silhouette width using the silhouette function (cluster package) and visualized with fviz_silhouette (factoextra package). Results were visualized as a dendrogram (fviz_dend) and heatmap (pheatmap), with colour palettes managed by the R ColorBrewer package.

Population Genetic Structure and Differentiation Analysis

SNV Data Pre-processing and Aggregation

To prepare data for population-level analyses, we first aggregated single nucleotide variant (SNV) information from per-sample output files generated by inStrain[7]. For each of the three representative species, an initial processing script was used to load all *_SNVs.tsv files. These files were filtered to retain only high-confidence SNVs (inStrain class of "SNV" or "con_SNV") at sites with minimum per-sample coverage of ≥30×. Metadata was parsed from filenames to map each sample to its corresponding lake population. This process yielded a unified table containing the raw allele counts (A, C, G, T) for every qualifying SNV site in every sample.

Site Filtering

A step-wise filtering was applied to the aggregated allele count table to ensure that only informative and high-quality polymorphic sites were used for F_{ST} and PCA, minimizing biases from sequencing errors or low-coverage data. Per-sample allele counts were first pooled to create a single set of counts for each lake population. This step averages out intra-lake sample variation to estimate population-level allele frequencies. Further, we applied a set of filters to this lake-pooled dataset to select for polymorphic sites suitable for comparing populations; sites were required to be covered (depth ≥30×) in at least two distinct lake

populations to be considered for pairwise comparisons. To ensure sites represented genuine, established polymorphisms, we required a minor allele count of at least 6 reads when pooled across all lakes. This removed sites where polymorphism is driven by very rare alleles or potential singleton errors.

To focus on sites representing established intra-population diversity, we defined a site as polymorphic if its Nei's gene diversity (He) was ≥ 0.01. This filter was used to exclude sites that are effectively monomorphic or contain only rare, uninformative alleles (singletons or sequencing errors), to ensure that downstream analyses of population structure are based on robustly polymorphic loci[8]. To regularize allele frequency estimates and prevent issues with zero-count alleles, especially at sites with low coverage, a pseudocount of 0.5 was added to each of the four allele counts (A, C, G, T) for every site in every lake.

Calculation of Pairwise Fixation Index (F_{ST})

To quantify genetic differentiation between all pairs of lake populations, we calculated the pairwise fixation index (F_{ST}). The calculations were performed on the full multiallelic data from the filtered set of SNV sites to retain all available genetic information. We used the Hudson's estimator, formulated as a ratio of sums over all qualifying loci[9]:

$$\mathrm{FST} = 1 - \left(\frac{\Sigma\, \pi_{\mathrm{within}}}{\Sigma\, \pi_{\mathrm{between}}}\right)$$

where $\Sigma\, \pi_{within}$ is the sum of the average within-population diversity across two populations and $\Sigma\, \pi_{between}$ is the sum of the between-population diversity, both calculated over all shared loci between the pair of populations. The final F_{ST} values represent the proportion of total genetic variance at filtered sites that is explained by differences between populations.

Principal Component Analysis (PCA)

To visualize the genetic variation and population structure, principal component analysis (PCA) was performed on the exact same set of filtered SNV sites used for the F_{ST} calculations. A matrix of samples (columns) by alleles (rows) was constructed from the per-sample allele counts. Each feature represented a specific non-reference allele at a given site ("scaffold position: C"), and its value was the frequency of that allele in a given sample. Alleles absent in more than 50% of the samples were removed. To ensure that the remaining features contribute meaningfully to variance, they were further filtered to retain only those with a minimum across-sample MAF (minor allele frequency; see below) of 0.02. Missing values in the final matrix were imputed with the allele's mean frequency across all samples in which it was present. The resulting matrix was centered and scaled before principal component analysis using the prcomp function in R.

Isolation-by-Distance (IBD) Analysis

To investigate geographic patterns of genetic differentiation, we performed an Isolation-by-Distance (IBD) analysis. All statistical analyses were conducted in R (v4.3.1)[1]. Geographic distances between lakes were calculated using the Haversine great-circle formula (R package

geosphere). To linearize the relationship between distance and differentiation, geographic distances were log10-transformed, and pairwise F_{ST} values were transformed using Slatkin's linearization (F_{ST} / (1 - F_{ST})). F_{ST} values were capped at 0.9999 to prevent division by zero. To assess scale-dependent patterns, we classified each lake pair into one of three geographic categories: (1) Local (<100 km), (2) within-continent (≥100 km, same continent), and (3) between-continents.

Statistical Testing and Slope Estimation

We used two approaches to analyse the IBD relationship. The primary statistical test for a significant correlation between the genetic and geographic distance matrices was a permutation-based Mantel test (R package vegan), which accounted for the non-independence of pairwise data. Significance was determined based on 1,000 permutations. For these tests, any lakes with fewer than three pairwise connections were excluded to ensure stable correlation estimates.

To estimate the slope of the IBD relationship and visualize trends, we fitted linear regression models (R function lm). However, since the standard Ordinary Least Squares (OLS) regression is sensitive to outliers and assumes data independence, we conducted a comprehensive robustness analysis to validate our slope estimates. First, we identified influential outliers in the OLS model using Cook's distance with a threshold of 4/n (R package car). Second, we compared the standard OLS slope to estimates from four alternative models, i.e., OLS with influential outliers removed, Weighted Least Squares (WLS), with weights set as the inverse of Cook's distance, and two robust regression models less sensitive to outliers: Huber's M-estimation (rlm in MASS) and MM-estimation (lmrob in robustbase). Finally, non-parametric 95% confidence intervals for the OLS slopes were generated using 1,000 bootstrap replicates (R package boot). The consistency of a slope's sign and significance across this suite of methods was used to confirm the robustness of our conclusions (see Supplementary Table 7 and Supplementary Figs. 5 and 6).

Allele Frequency Spectrum, Strain Diversity, and Selection Analysis

Extraction of Minor Allele Frequencies (MAF)

To construct allele frequency spectra (AFS) and assess within-population diversity, we first extracted per-sample minor allele frequencies (MAFs) from inStrain[7] SNV output files (*.SNVs.tsv). A custom bash and awk pipeline was used to process all samples. This applied a stringent filtering to each SNV site; sites were required to have a minimum total read depth of ≥30×. For an allele to be considered valid, a base (A, C, G, or T) required support from a minimum of 3 reads. This step helped minimize the impact of sequencing errors on low-frequency variant calls (allele support filter; b-gate). Only sites with exactly two alleles passing the b-gate filter were retained (biallelic constraint). This ensures that the MAF is clearly defined to simplify downstream analyses by excluding complex multiallelic sites. For each site passing these filters, the MAF was calculated as the count of the second-most abundant allele divided by the total coverage. This process produced a comprehensive table of all high-confidence biallelic SNVs and their corresponding MAFs for every sample.

Selection of minor allele frequency (MAF) threshold to infer population structure

MAF filtering substantially affects population genetic inferences, with different windows capturing distinct evolutionary timescales[10], we tested which frequency range best captured ecosystem-age dynamics. We excluded rare variants (MAF < 0.2) to minimize sequencing errors and focus on stable variation. We calculated the fraction of SNPs within three overlapping MAF windows: (1) MAF 0.2-0.4 (intermediate frequencies), (2) MAF 0.2-0.5 (intermediate to common), and (3) MAF 0.4-0.5 (near-balanced polymorphisms).The primary metric included the fraction of SNPs at each MAF window, defined as the proportion of all polymorphic sites within a population that fall in the respective range. We defined the SNP fraction (denoted frac $_{\text{MAF\_window}}$) as:

$$frac_{\text{MAF\_window}} = \frac{number\ of\ SNPs\ within\ the\ MAF\ window}{total\ number\ of\ SNPs\ with\ MAF > 0}$$

This metric was computed separately for each sample and then averaged across biological replicates within each lake to obtain lake-level estimates.

To test the hypothesis that ecosystem age influences within-population genetic structure, we analysed the relationship between lake age and fraction of SNPs at each MAF window. Lake age data (in years) was compiled from literature and categorized as ancient, post-glacial, post-mining, or reservoir. Lake age was $\log_{10}$ transformed prior to all statistical analyses to normalize the distribution and linearize the potential relationship. A non-parametric Spearman's rank correlation (cor.test function in R), was used to assess the relationship between variables without assuming linearity. This analysis was conducted on the full dataset combining all species at each MAF threshold window. Statistical significance was set at $p < 0.05$.This was visualized using a scatter plot created with the ggplot2 package in R. To illustrate the overall trend, an Ordinary Least Squares (OLS) linear regression line (95% confidence interval) was overlaid on the data.

pN/pS Data Calculation and Selection Analysis

The raw SNV files were prefiltered to retain the mutation-type annotation information with the same filtering criteria as above (depth ≥30×, b-gate count ≥3, biallelic-only). For each qualifying SNV, the mutation type field from the inStrain[7] output was parsed to classify the polymorphism as either "synonymous" or "non-synonymous." SNVs with ambiguous or missing annotations were classified as "unknown" and excluded from pN/pS calculations. This process generated a table containing the MAF and mutation type for every filtered SNV in every sample. We used the annotated MAF table to infer selective pressures on protein-coding genes via the ratio of non-synonymous to synonymous polymorphisms (pN/pS). The analysis was conducted using a custom R script.

For each lake population, we calculated a single genome-wide pN/pS ratio. This was computed as the total count of non-synonymous SNPs divided by the total count of synonymous SNPs across all filtered polymorphic sites in that population. Further, to resolve the selection dynamics in more detail, we also calculated pN/pS within distinct MAF bins. All

filtered SNPs were partitioned into five bins: (0–0.05), (0.05–0.10), (0.10–0.20), (0.20–0.40), and (0.40–0.50). The pN/pS for each bin were calculated as the ratio of the number of non-synonymous SNPs to synonymous SNPs falling within that frequency range.

Supplementary references:

1 Giorgi, F. M., Ceraolo, C. & Mercatelli, D. The R language: an engine for bioinformatics and data science. *Life* **12**, 648 (2022).

2 Razali, N. M. & Wah, Y. B. Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics* **2**, 21-33 (2011).

3 Brown, M. B. & Forsythe, A. B. Robust tests for the equality of variances. *Journal of the American statistical association* **69**, 364-367 (1974).

4 McKight, P. E. & Najab, J. Kruskal‐wallis test. *The corsini encyclopedia of psychology*, 1-1 (2010).

5 Elliott, A. C. & Hynan, L. S. A SAS® macro implementation of a multiple comparison post hoc test for a Kruskal–Wallis analysis. *Computer methods and programs in biomedicine* **102**, 75-80 (2011).

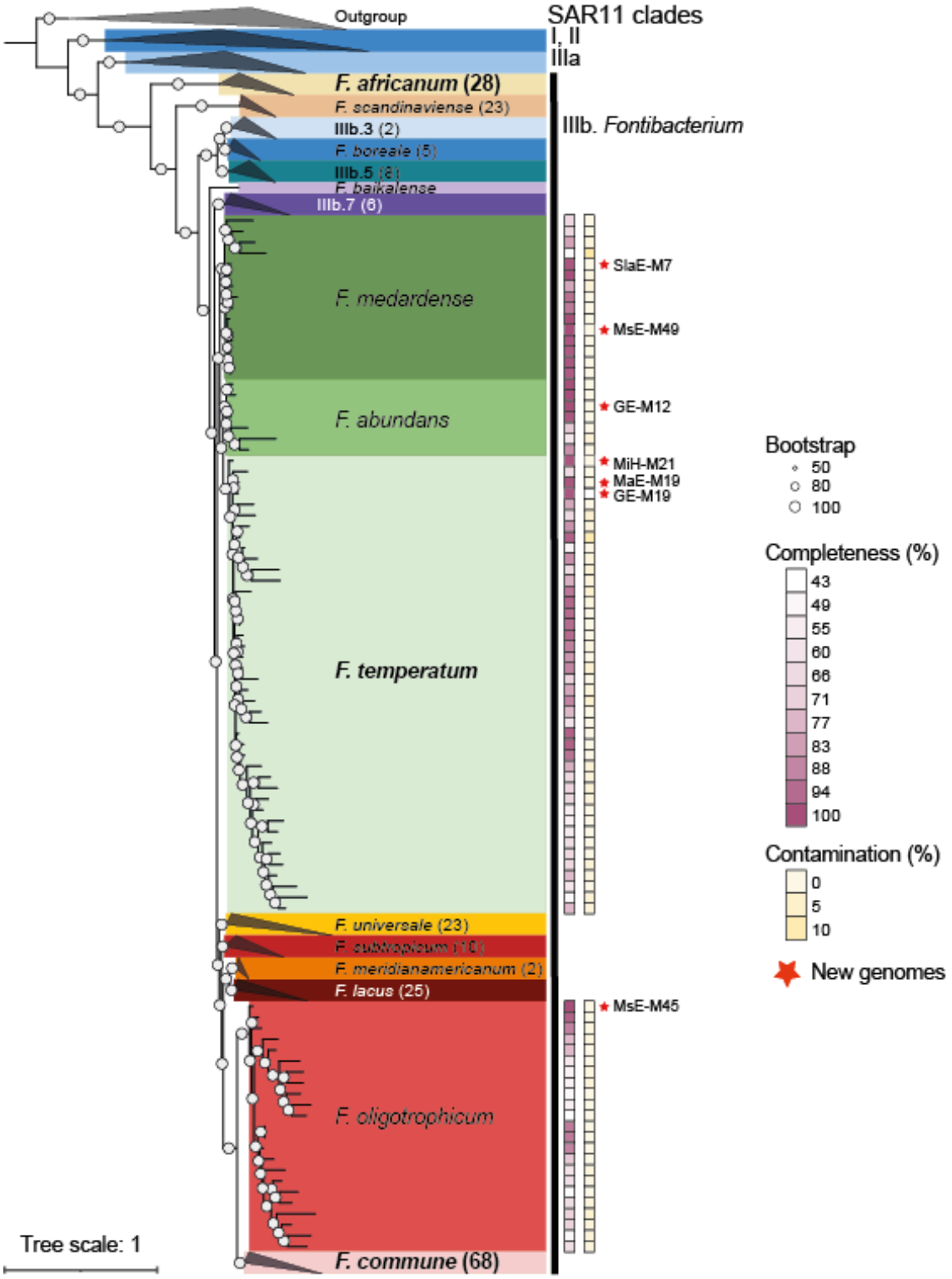
6 McKnight, P. E. & Najab, J. Mann‐whitney U test. *The Corsini encyclopedia of psychology*, 1-1 (2010).

7 Olm, M. R. *et al.* inStrain profiles population microdiversity from metagenomic data and sensitively detects shared microbial strains. *Nature Biotechnology* **39**, 727-736 (2021).

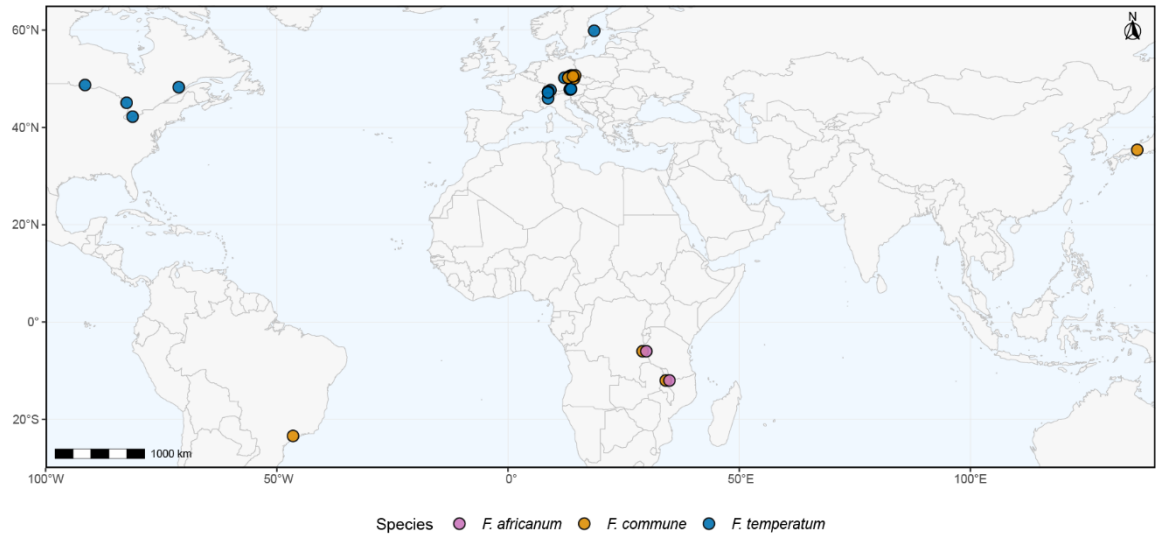
8 Nei, M. (Columbia University Press, 1987).

9 Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome research* **23**, 1514-1521 (2013).

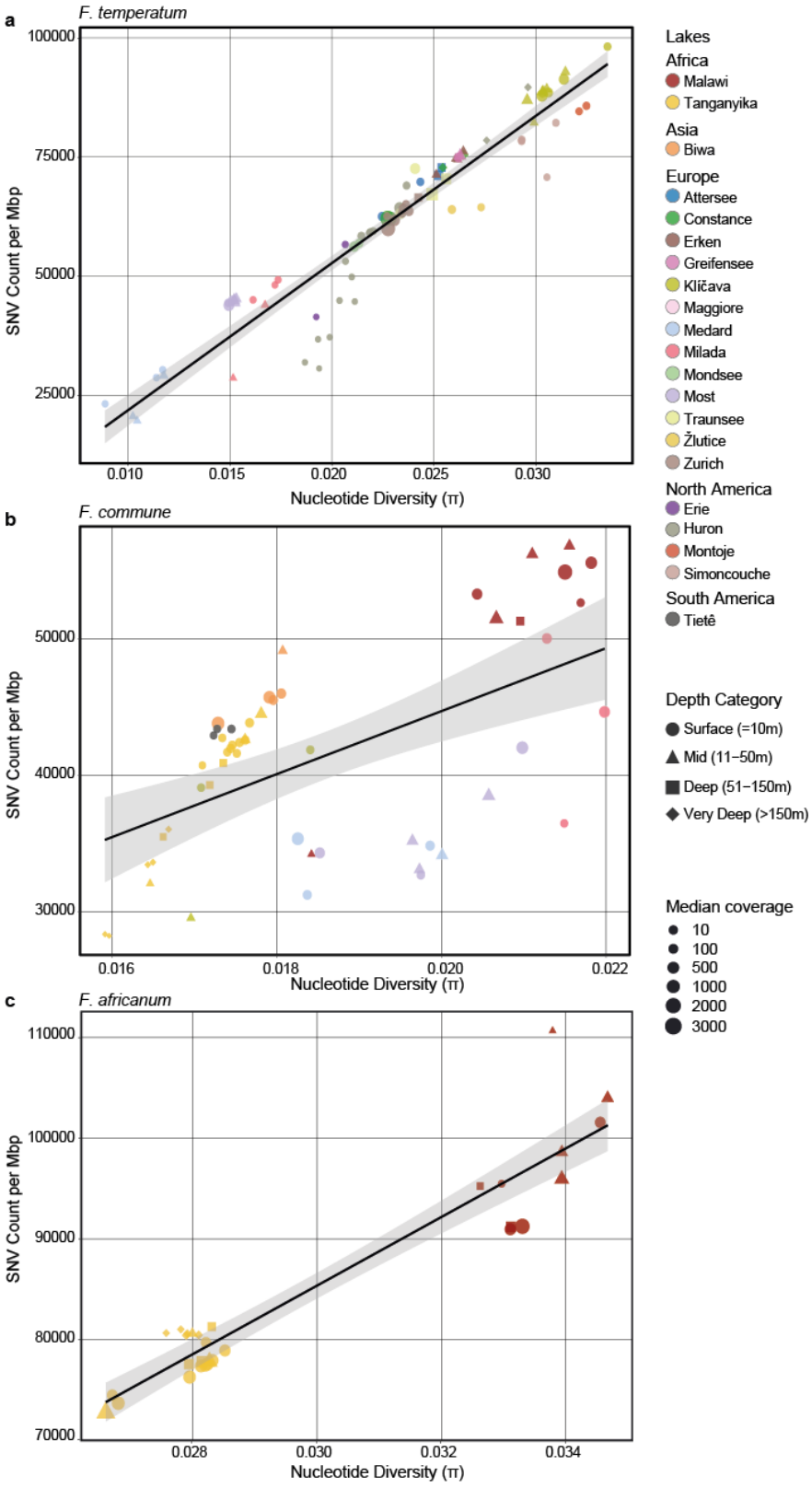
10 Linck, E. & Battey, C. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources* **19**, 639-647 (2019).



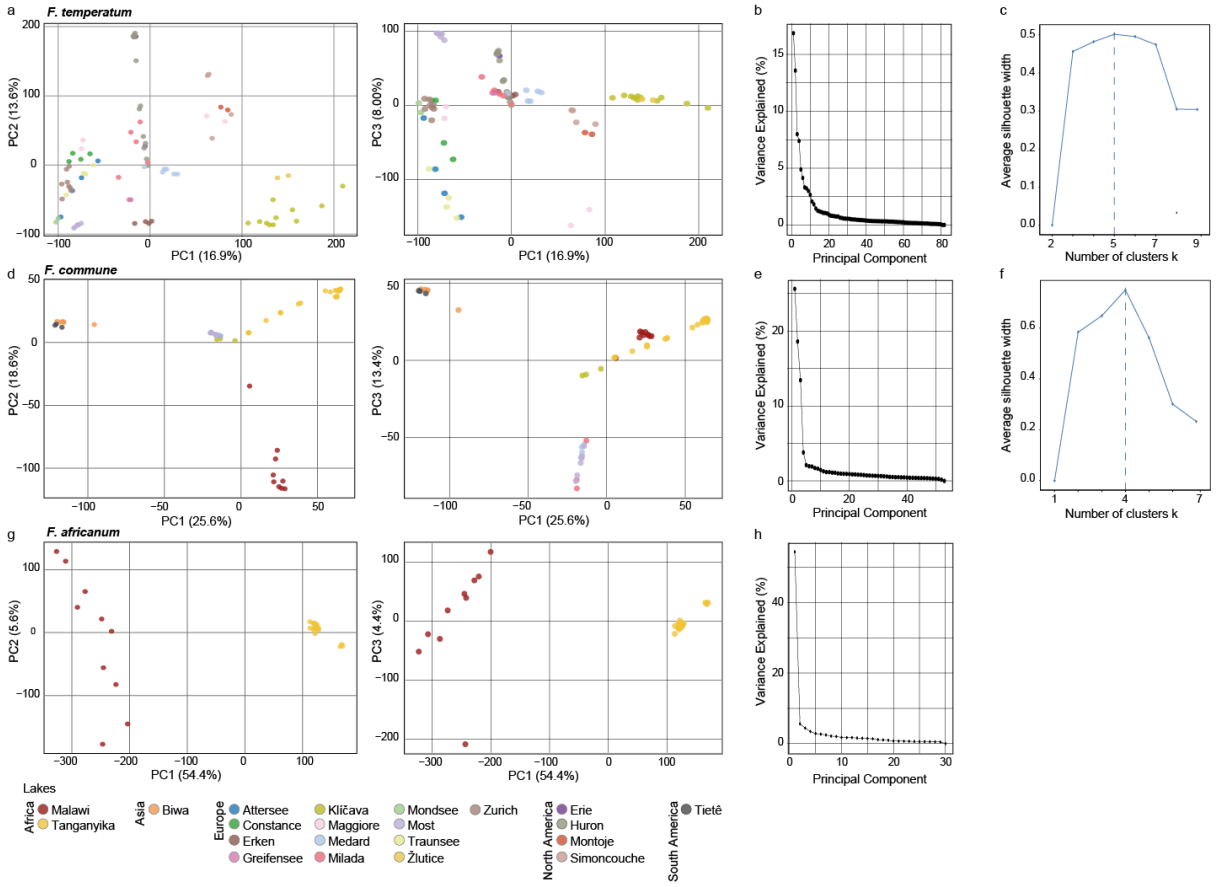
Supplementary Fig. 1: Phylogeny of the SAR11 order. Phylogenomic tree with ultrafast bootstrapping using 751 single-copy marker protein sequences and representatives of SAR11-IV and V as outgroup. Different clades of SAR11-I, II, IIIa and species of SAR11-IIIb (*Fontibacterium*) are displayed in different colours, numbers of genomes in each collapsed branch is given in brackets. New genomes sequenced in this study are highlighted by red stars followed by strain names (n =7), the three species used for SNV analyses are highlighted in bold.



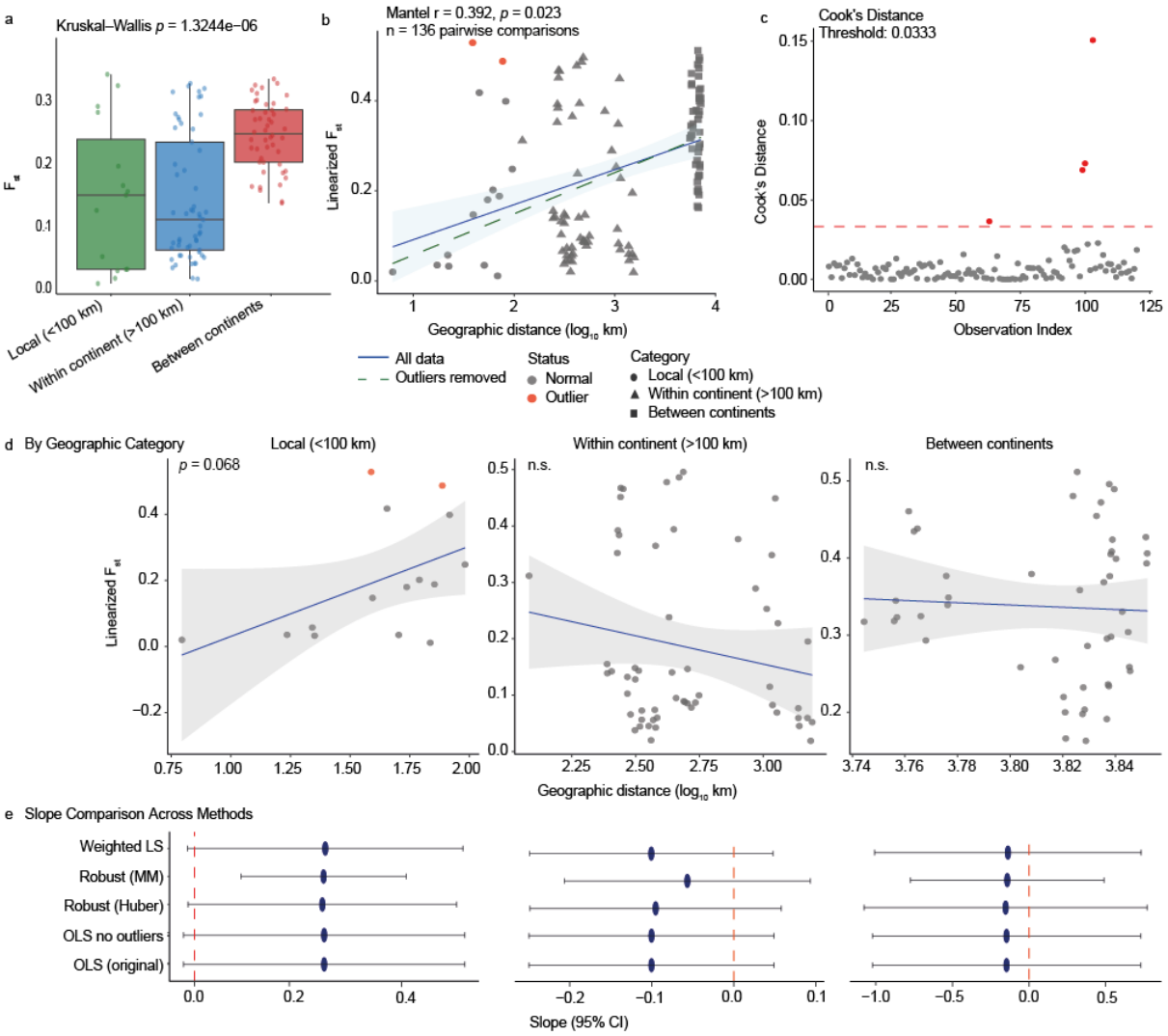
Supplementary Fig. 2: Locations of lake metagenomes used for analyses (*n* metagenomes = 117, *n* lakes = 21). Points are colored by *Fontibacterium* species occurring in a specific system.



Supplementary Fig. 3: Scatter plots showing correlations between nucleotide diversity (π) and SNV density (counts per Mbp) for each species. Points are colored by lake systems, sized by median coverage and shaped by depth category. Grey bands indicate 95% confidence intervals.

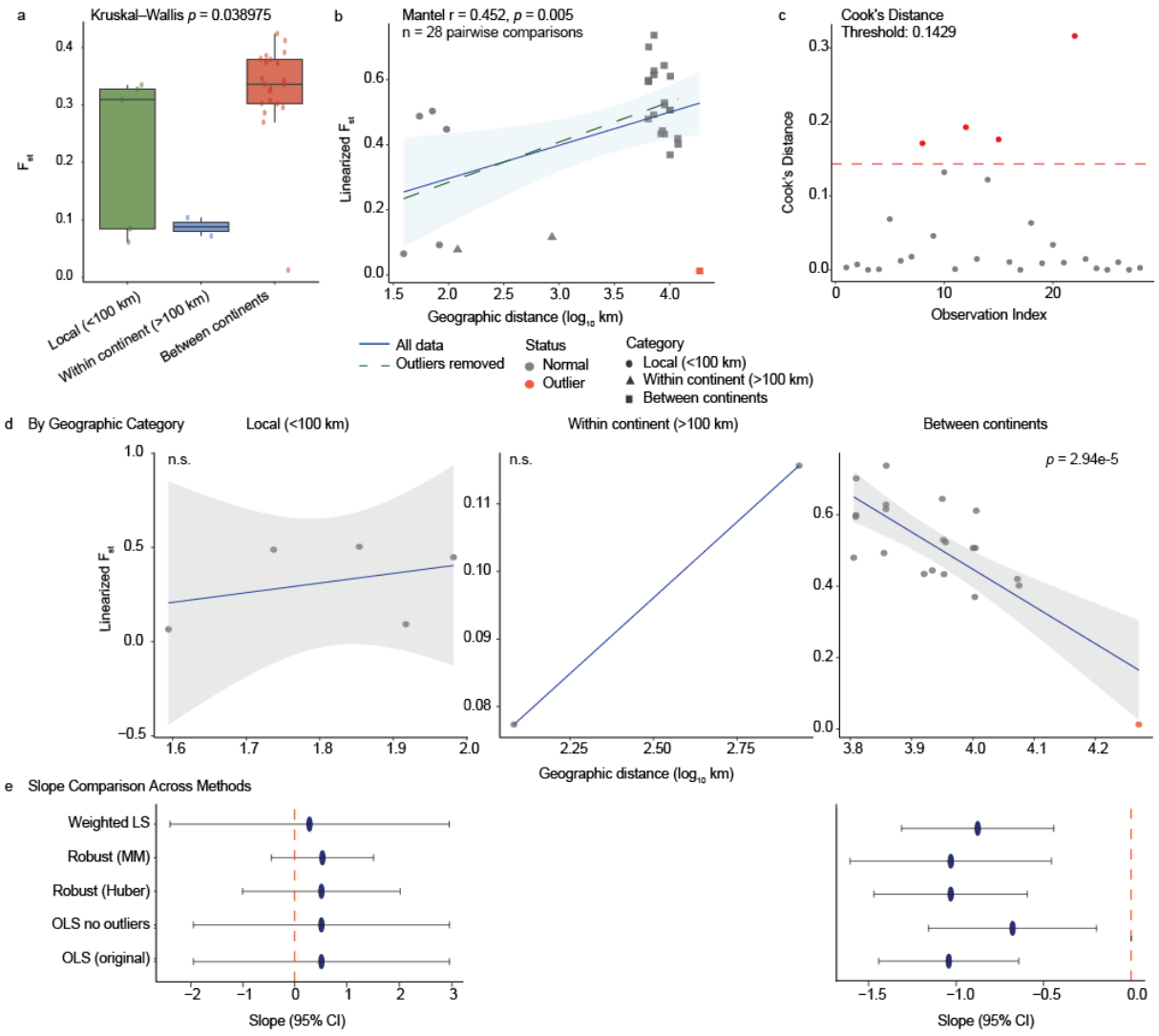


Supplementary Fig. 4: Population genetic structure and differentiation of *Fontibacterium* species. **a, b, e, f, i, j:** Principal component analysis (PCA) showing genetic differentiation among *Fontibacterium* populations from lakes across five continents, shown are variances explained by PC1, PC2 and PC3; **c, g, k:** explained variance of PCs; **d, h:** Silhouette analysis for optimal clusters observed in Fig. 2. **a-d:** *F. temperatum* populations (n=80) from 17 European and North American lakes. **e-h:** F. commune populations (n=53) from nine lakes in Africa, Asia, Europe, and South America. **i-k:** *F. africanum* populations (n=30) from African Great Lakes.



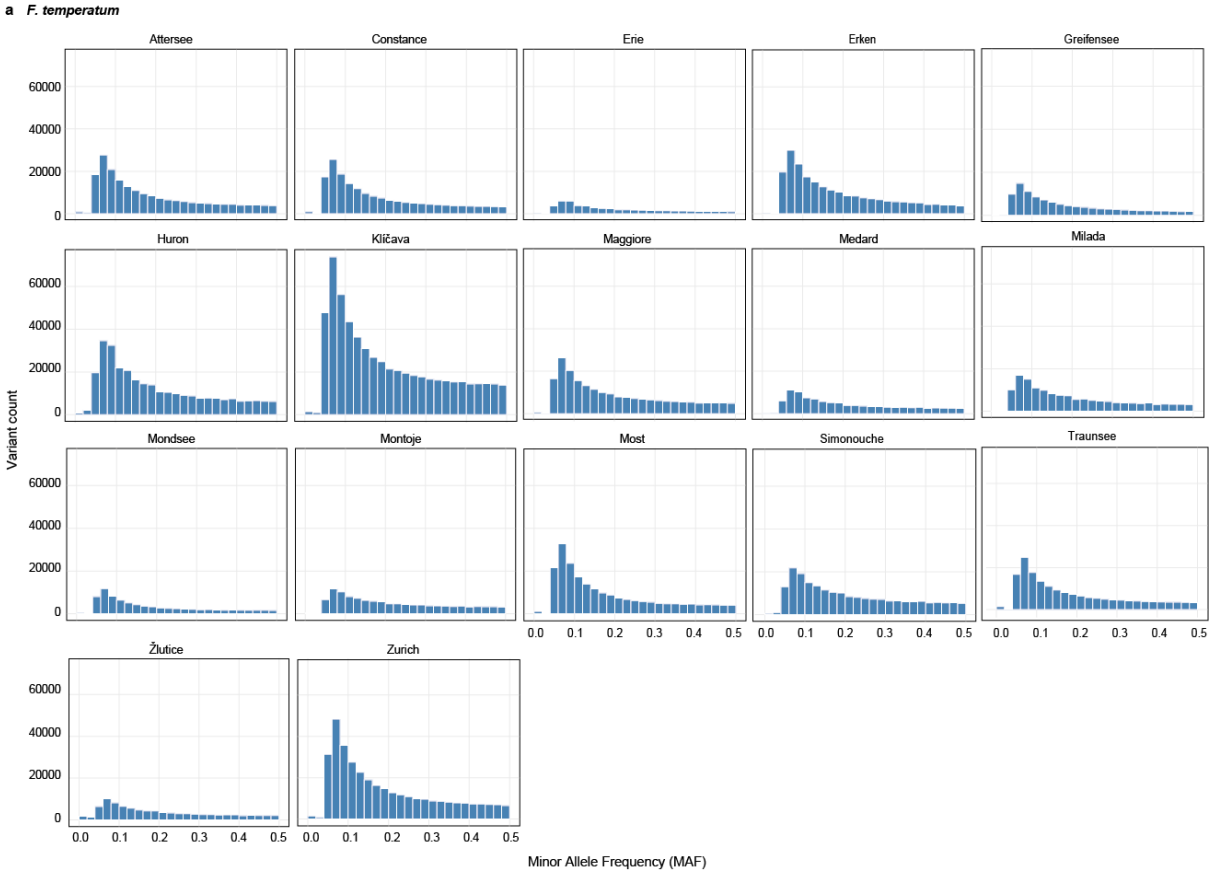
Supplementary Fig. 5: Scale-dependent patterns of Isolation-by-Distance (IBD) for the temperate specialist, *F. temperatum*. Genetic differentiation (F_{ST}) among 17 lake populations and its relationship with geographic distance. **(a)** F_{ST} distribution by geographic category. Boxplots showing the distribution of pairwise F_{ST} values into three geographic scales: local (<100 km), within continent (>100 km), and between continents. The overall variation among categories was statistically significant (Kruskal-Wallis $p = 1.32 \times 10^{-6}$), with inter-continental pairs showing the highest differentiation. **(b)** Scatterplot of linearized F_{ST} against $\log_{10}$ transformed geographic distance for all 136 pairwise comparisons. The analysis revealed a significant, weak positive overall correlation (Mantel r = 0.392, $p = 0.023$). The OLS regression on all data indicated by solid blue line and the regression after removing influential outliers by green line. **(c)** Outlier diagnosis with Cook's distance for the OLS regression in (b). Points exceeding the dashed red threshold (4/n = 0.0333) are identified as influential outliers. **(d)** IBD analysis by geographic scale. The IBD relationship was modelled separately for each category. A marginally significant positive trend observed only at the local scale ($p =$

0.068, bootstrap significant), indicating dispersal limitation over short distances. No significant relationship between genetic and geographic distance was found at within or between continents scales. **(e)** Comparison of the estimated IBD slopes and their 95% confidence intervals across five different statistical methods for each geographic category. The vertical dashed red line at zero indicates the null hypothesis of no relationship. This confirmed the robustness of the findings in (d); the positive slope at the local scale is consistent across methods, while the slopes for the broader geographic scales consistently overlap zero, indicating a lack of significant IBD.

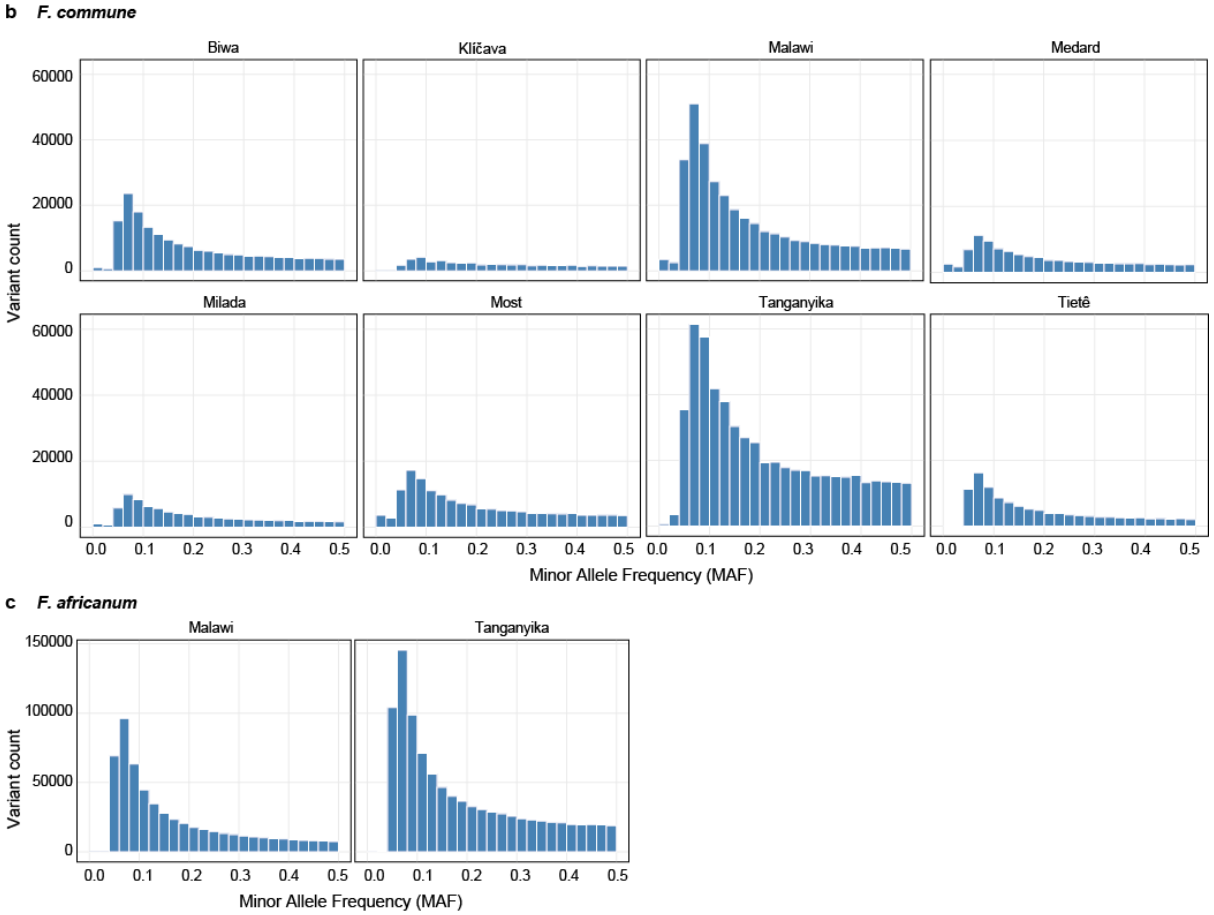


Supplementary Fig. 6: Isolation-by-Distance (IBD) patterns for the ubiquitous species, *F. commune*, reveal global connectivity. Genetic differentiation (F_{ST}) among eight globally distributed lake populations and its relationship with geographic distance. **(a)** F_{ST} distribution by geographic category. Boxplots showing the distribution of pairwise F_{ST} values by geographic scale. The overall variation among categories was statistically significant (Kruskal-Wallis $p = 0.038975$). **(b)** Scatterplot of linearized F_{ST} against $\log_{10}$ transformed geographic distance for all 28 pairwise comparisons. The overall trend is weakly positive but is heavily influenced by the distinct patterns at different scales (Mantel r = 0.452, $p = 0.005$). The solid blue line shows the OLS regression on all data; the dashed green line shows the regression after removing influential outliers. **(c)** Outlier diagnosis with Cook's distance for the OLS regression in (b). Several points exceeded the dashed red threshold (4/n = 0.1429), identifying them as highly influential outliers. This included the genetically similar but geographically distant Lake Biwa (Japan) and Tietê Reservoir (Brazil) populations. **(d)** IBD analysis by geographic scale. The relationship between genetic and geographic distance is modelled separately for each category. At the local and within continent scales, no significant IBD pattern is detected. In contrast, at the between continents scale, there is a strong and highly significant negative IBD relationship ($p = 2.94 \times 10^{-5}$). This indicates that the most geographically distant

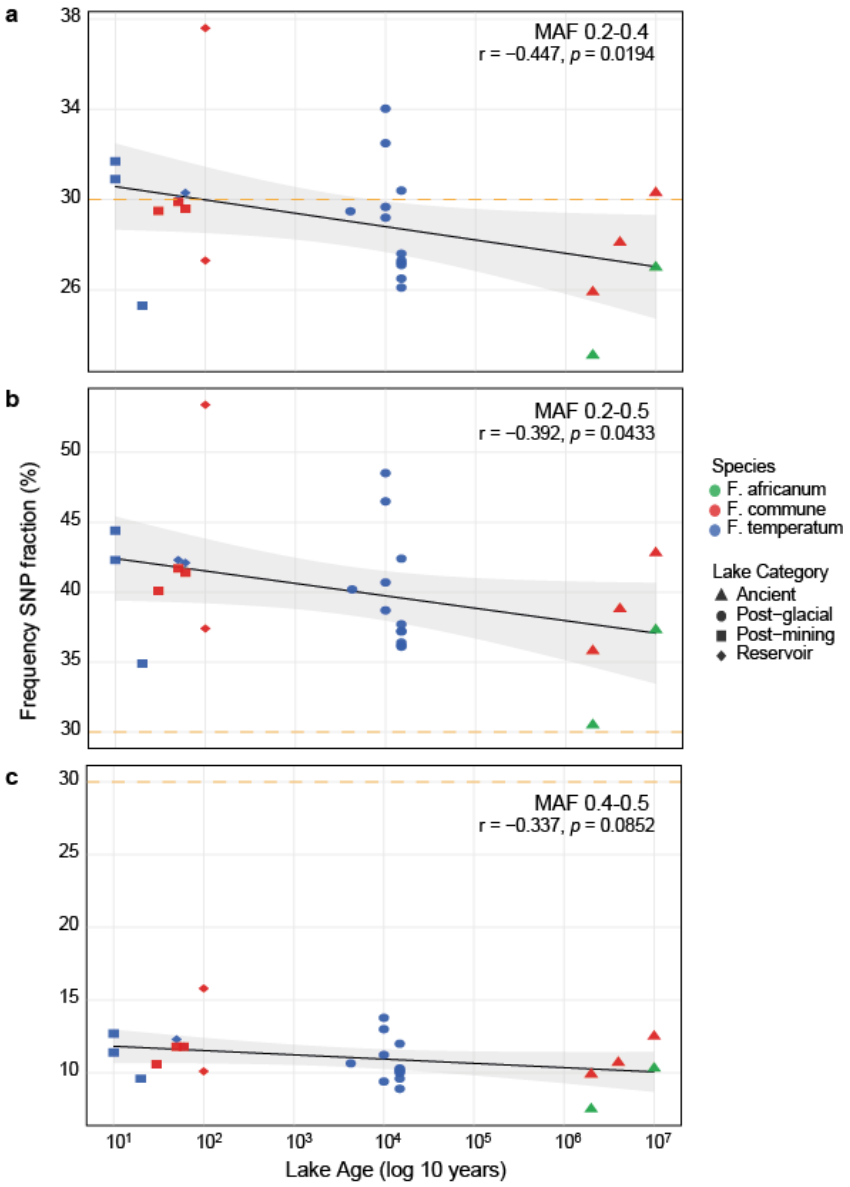
populations are the most genetically similar. **(e)** Comparison of the IBD slope estimates and their 95% confidence intervals across five different statistical methods. This confirmed that the negative slope at the intercontinental scale is a robust finding, with the confidence interval remaining well below zero regardless of the statistical model used.



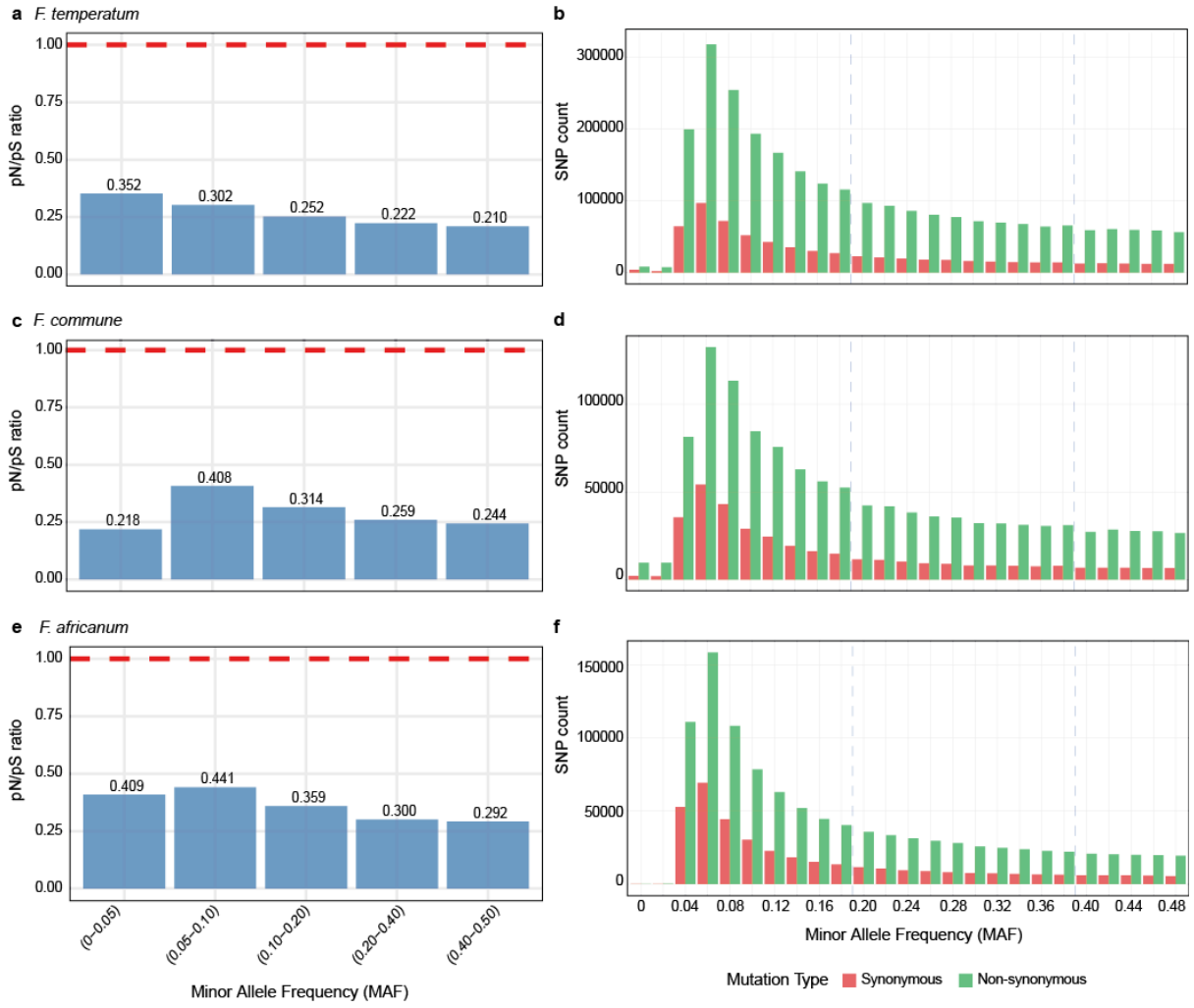
Supplementary Fig. 7a: Allele Frequency Spectra (AFS) for *F. temperatum* populations across freshwater lakes. Each histogram displays the distribution of single nucleotide polymorphisms (SNPs) binned by minor allele frequency (MAF, 0.0 to 0.5) for each lake population. These spectra were constructed by pooling all per-sample SNP data for a given species within each lake. The displayed variants were subjected to a stringent filtering pipeline to ensure high quality. Each SNV site was required to have a minimum read depth of ≥30× in the sample from which it was called and be biallelic, with the minor allele supported by a minimum of 3 reads.



Supplementary Fig. 7 continued: Allele Frequency Spectra (AFS) for (b) *F. commune* and (c) *F. africanum* populations across freshwater lakes. Each histogram displays the distribution of single nucleotide polymorphisms (SNPs) binned by minor allele frequency (MAF, 0.0 to 0.5) for each lake population. These spectra were constructed by pooling all per-sample SNP data for a given species within each lake. The displayed variants were subjected to a stringent filtering pipeline to ensure high quality. Each SNV site was required to have a minimum read depth of ≥30× in the sample from which it was called and be biallelic, with the minor allele supported by a minimum of 3 reads.



Supplementary Fig. 8: Relationship between ecosystem age and minor allele frequency (MAF) thresholds. The fraction of single nucleotide polymorphisms (SNPs) at different minor allele frequencies windows was plotted against lake age. Each point represents a single lake population, colored by species (*F. africanum*, green; *F. commune*, red; *F. temperatum*, blue) and shaped by lake category (Ancient, triangle; Post-glacial, circle; Post-mining, square; Reservoir, diamond). Solid black lines represent the Ordinary Least Squares (OLS) regression trend, with grey shaded areas indicating the 95% confidence interval. Horizontal dashed orange lines indicate the 30% threshold for high strain diversity used as a baseline in this study. **(a)** SNP fraction at intermediate frequencies (0.2-0.4 MAF) with the strongest negative correlation (Spearman's ρ = -0.447, p = 0.019) between ecosystem age and genetic heterogeneity. **(b)** SNP fraction at intermediate to common frequencies (0.2-0.5 MAF) also showed significant negative correlation with lake age (Spearman's ρ = -0.392, p = 0.043). **(c)** SNP fraction at near-balanced frequencies (MAF 0.4-0.5) showed no significant trend (Spearman's ρ = -0.337, p = 0.085).



Supplementary Figure 9: Dynamics of purifying selection across the wide range of allele frequencies for (a) the temperate specialist (*F. temperatum*), (b) the ubiquitous species (*F. commune*), and (c) the endemic species (*F. africanum*). All single nucleotide polymorphisms (SNPs) were pooled across all lake populations for each species to generate these patterns. (**a-c)** The ratio of non-synonymous to synonymous polymorphisms (pN/pS) calculated within distinct minor allele frequency (MAF) bins. The red dashed line indicates the neutral expectation (pN/pS = 1.0). (**d-f)** The right panels show the absolute counts of non-synonymous (green) and synonymous (red) SNPs across the allele frequency spectra. Dashed blue lines indicate the intermediated frequency threshold (MAF; 0.2-0.4).