Continental-Scale Gene Flow Prevents Allopatric Divergence of Pelagic Freshwater Bacteria

Matthias Hoetzinger[1,2,*], Alexandra Pitt[1], Andrea Huemer[1], and Martin W. Hahn[1]

[1]Research Department for Limnology, University of Innsbruck, Mondseestrasse 9, A-5310 Mondsee, Austria

[2]Department of Aquatic Sciences and Assessment, Swedish University of Agricultural Sciences, Uppsala SE-75651, Sweden

*Corresponding author: E-mail: matthias.hoetzinger@uibk.ac.at.

Accepted: 25 January 2021

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

Abstract

Allopatric divergence is one of the principal mechanisms for speciation of macro-organisms. Microbes by comparison are assumed to disperse more freely and to be less limited by dispersal barriers. However, thermophilic prokaryotes restricted to geothermal springs have shown clear signals of geographic isolation, but robust studies on this topic for microbes with less strict habitat requirements are scarce. Furthermore, it has only recently been recognized that homologous recombination among conspecific individuals provides species coherence in a wide range of prokaryotes. Recombination barriers thus may define prokaryotic species boundaries, yet, the extent to which geographic distance between populations gives rise to such barriers is an open question. Here, we investigated gene flow and population structure in a widespread species of pelagic freshwater bacteria, *Polynucleobacter paneuropaeus*. Through comparative genomics of 113 conspecific strains isolated from freshwater lakes and ponds located across a North–South range of more than 3,000 km, we were able to reconstruct past gene flow events. The species turned out to be highly recombinogenic as indicated by significant signs of gene transfer and extensive genome mosaicism. Although genomic differences increased with spatial distance on a regional scale (<170 km), such correlations were mostly absent on larger scales up to 3,400 km. We conclude that allopatric divergence in European *P. paneuropaeus* is minor, and that effective gene flow across the sampled geographic range in combination with a high recombination efficacy maintains species coherence.

Key words: pelagic freshwater bacteria, microbial evolutionary ecology, gene flow, homologous recombination, *Polynucleobacter*, allopatric speciation.

Significance

The relevance of allopatric speciation for prokaryotes is a controversial topic, and studies that use adequate tools to resolve population divergence on timescales decisive for speciation are lacking. We tackled this issue by establishing a large genome collection of conspecific bacteria isolated from geographically distant freshwater lakes and ponds. Comparative genomics allowed us to retrace ample gene flow across the whole geographic range that has happened within relatively short timescales (postglaciation period). This pronounced admixture counteracts allopatric divergence, which stands in stark contrast to the geographic isolation observed in extremophilic prokaryotes, and reveals unprecedented insights into the coherence of a free-living bacterial species across continental-scale distances.

Introduction

The concept of prokaryotic species is a subject of great debate. Major controversies derive from the fact that we know too little about the origin and consistency of putative species. How do new species arise, and how is the coherence of existing ones maintained? Diverse mechanisms have been proposed (Cohan 2001; Sikorski and Nevo 2005; Falush et al. 2006; Whitaker 2006; Fraser et al. 2007; Doolittle and Zhaxybayeva 2009; Koeppel et al. 2013; Shapiro and Polz 2015), which suggest that plenty of circumstances may cause

© The Author(s) 2021. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

speciation. Although prokaryotes reproduce asexually, high recombination rates within various species invoke a biological species concept (Fraser et al. 2007; Bobay and Ochman 2017; Hoetzinger and Hahn 2017). Within the framework of this concept, speciation is initiated by a lack of gene flow among conspecific populations (Mayr 1942). Gene flow in turn is controlled by at least five principal factors, which are 1) the dispersal potential of the species, 2) the connectivity among habitats populated by the species, 3) the potential of dispersed organisms to colonize a given habitat, 4) the recombination efficacy of the species, and 5) population sizes.

Dispersal limitation has long been regarded to be insignificant for free-living prokaryotes, which is reflected by the often-quoted statement "everything is everywhere; the environment selects" (Baas Becking 1934). The broad acceptance of this notion probably stems from the long-standing shortcomings to differentiate closely related prokaryotes, for example, as a consequence of using the conserved 16S rRNA gene to identify bacteria. Although examples of bacteria with identical 16S rRNA gene sequences originating from different continents reveal a potential for global dispersal (Zwart et al. 1998; Hahn and Pöckl 2005), the gene is of limited use for studying population differentiation with respect to speciation, which may happen faster than polymorphisms are fixed on individual genes (Hahn, Jezberová, et al. 2016; Antony-Babu et al. 2017). Studies employing multilocus sequencing have shown that dispersal barriers that may lead to allopatric speciation exist for thermophilic microbes (Papke et al. 2003; Whitaker et al. 2003). The low connectivity among potential habitats, that is, regions with geothermal activity, strongly limits gene flow among *Sulfolobus* populations (Whitaker et al. 2003; Anderson et al. 2017), although the species turned out to recombine frequently (Whitaker et al. 2005). A recent study investigating a species-like taxon of free-living terrestrial bacteria revealed population differentiation by gene flow discontinuities that was attributed to local adaptation as well as dispersal limitation (Chase et al. 2019). The geographic patchiness of geothermal springs or the high spatial heterogeneity of habitats for soil bacteria are contrasted by a more continuous transition between habitats of marine microorganisms (Müller et al. 2014). The conjectural absence of strong dispersal barriers in the oceans might explain why relatively few studies discuss dispersal limitation in context of population differentiation of marine microbes (Cui et al. 2015; Whittaker and Rynearson 2017). It might be worth noting that a theoretical model, disregarding natural selection but considering exclusively neutral mutations, genetic drift, and ocean currents, predicted the continuous emergence of differentiated populations in different ocean provinces (Hellweger et al. 2014).

Standing freshwater systems (lakes and ponds) display intermediate habitat connectivity compared with more intermixed oceanic habitats on the one end, and island-like hot springs on the other end of the spectrum. To date, there are no studies available that allow for well-founded estimations about the importance of allopatric speciation in such freshwater habitats. On the one hand, gene flow barriers may separate geographically distant populations, and thus, extant species may represent transient taxa arisen from frequent speciation and extinction in numerous geographically separated regions. On the other hand, extensive gene flow across large spatial distances may counteract population differentiation and sustain species coherence. Where a given taxon resides within this spectrum depends on its specific characteristics regarding the five factors mentioned above. To enable reasonable estimations about speciation in different taxa, studies on selected species are essential. Large culture collections of conspecific isolates from spatially distant sites allow for the implementation of such studies, but such culture collections of free-living bacteria are scarce.

Here, we investigated *Polynucleobacter paneuropaeus*, a species of pelagic freshwater bacteria previously described on the basis of six strains, isolated from lakes and ponds located along a 3,400 km cross-section across Europe (Hoetzinger et al. 2019). Mainly by targeted isolation from habitats of the same geographic range, we increased the total number of strains affiliated to this species to 120. Through comparative genomics of 113 strains, we analyzed geographic patterns in population structure and retraced recent gene transfer among the genomes. We further used a cultivation-independent method, that is, amplicon sequencing of a high-resolution marker gene, to reveal ecological preferences of the species, and assess differences in population densities among regions. Results from isolated strains and cultivation-independent environmental data were compiled to assess the factors that determine gene flow and shape genomic variation among populations. We finally discuss the potential of dispersal barriers for giving rise to discontinuous lineages and its implications for allopatric speciation of freshwater microbes.

Materials and Methods

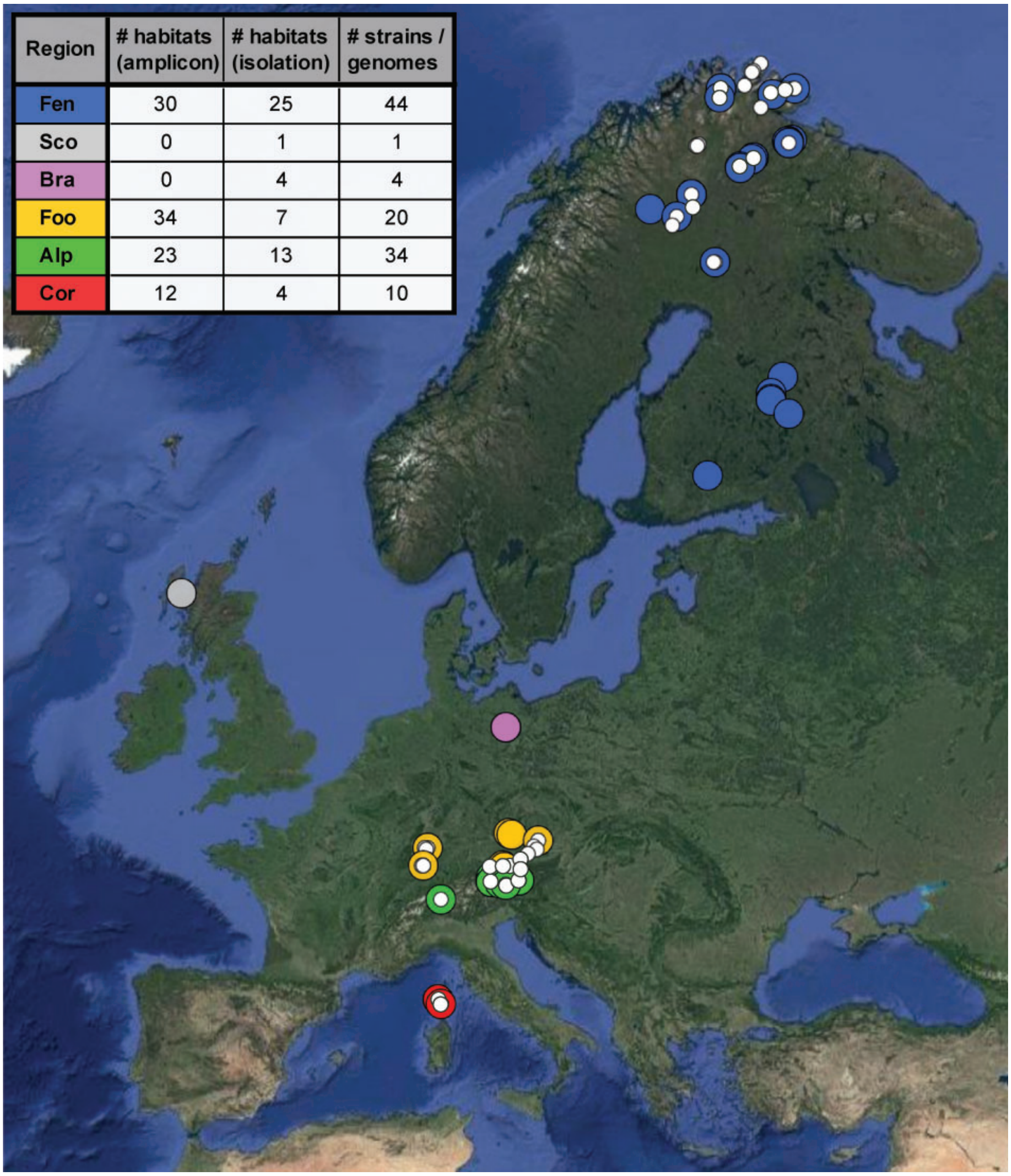
Terminology Used in This Paper

Habitat: lake or pond | population: group of *P. paneuropaeus* strains isolated from the same geographic region, that is, Corsica (Cor), Alps (Alp), Foothills of the Alps (Foo), Brandenburg (Bra), Scotland (Sco), and Fennoscandia (Fen); demarcation of regions was informed by population genomic and phylogenetic analyses (see below) | metapopulation: all *P. paneuropaeus* strains isolated from the sampled geographic range| PnecC: species-rich *Polynucleobacter* subcluster comprising several described species, including *P. paneuropaeus.*

Distribution of *Polynucleobacter paneuropaeus*

The distribution of *P. paneuropaeus* in freshwater habitats (lakes, ponds, and streams) located along a European

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021



Region	# habitats (amplicon)	# habitats (isolation)	# strains / genomes
Fen	30	25	44
Sco	0	1	1
Bra	0	4	4
Foo	34	7	20
Alp	23	13	34
Cor	12	4	10

Fig. 1.—Sampling sites for amplicon sequencing and isolation of *P. paneuropaeus* strains. The sites were grouped into four regions (Fennoscandia [Fen], blue; Foothills of the Alps [Foo], yellow; Alps [Alp], green; Corsica [Cor], red; Brandenburg [Bra], light purple; Scotland [Sco], gray) under consideration of population structure (see fig. 4). Sites from which isolates were obtained are indicated by big colored dots (see inset for color code) and sites from which amplicon samples were retrieved are shown as small white dots. Note that samplings sites for isolation and amplicon sequencing partially overlap. The number of samples and number of isolates/genomes obtained from each region are given in the inset. The map was adopted from Google, ©2020 NASA, TerraMetrics.

North–South (71°N–42°N) cross-section spanning a geographic range of 3,400 km from Lapland (Northern Fennoscandia) to the island of Corsica located in the Mediterranean Sea (fig. 1) was investigated by a cultivation-independent approach. High-throughput sequencing of amplicons of the single-copy gene encoding the primosomal replication protein N (*priB*) was performed. In contrast to 16S rRNA genes, this marker provides a species-specific taxonomic resolution. The 114 samples from 99 habitats selected for this investigation represent a broader ecological diversity (e.g., pH 4.2–8.6) than those habitats selected for isolation experiments. Surface water (0.1–0.5 m depth) samples were taken from the shoreline or from piers if available and filtered through 0.2 µm Nucleopore filters. Filters were preserved in absolute ethanol and stored in a mobile refrigerator upon arrival in the laboratory. DNA was extracted as described previously (Hahn et al. 2015). Amplicons were obtained by primers (priBinnFd 5′-YGGCGTTGAATCATTTMAC-3′ and

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

priBinnRd 5′-TTCCAAACGCCATGRTGATT-3′) specific for *priB* genes of strains affiliated with the *Polynucleobacter* subcluster PnecC (Hahn 2003). The primers were tagged with Illumina adaptors and sample-specific tags. The amplicons were paired-end sequenced (300 bp) by Illumina MiSeq. Reads were processed using QIIME2 (Bolyen et al. 2019), which included demultiplexing, trimming of adapter and primer sequences, joining of paired reads, denoising (DADA2), exclusion of too long (>288 bp) and too short (<285 bp) sequences, removal of sequences with additional stop codons, removal of sequences with copy number <10 present only in single samples, and rarefication to 25,230 sequences per sample. Three environmental samples were excluded due to too small read numbers. Reads obtained from the remaining 114 samples were clustered into operational taxonomic units (OTUs) by employing a 98% sequence similarity threshold, which was found to best represent the gANI threshold of 95% between isolated strains (see supplementary fig. S2, Supplementary Material online) that has proven useful to delineate bacterial species in general (Konstantinidis et al. 2006; Richter and Rosselló-Móra 2009), as well as *Polynucleobacter* species (Hahn, Schmidt, et al. 2016; Hoetzinger and Hahn 2017). OTUs were taxonomically classified by employing a reference set of *priB* sequences obtained from *Polynucleobacter* strains.

Isolation of Strains

Strains have been isolated between August 2003 and June 2017 (see supplementary table S1, Supplementary Material online) from surface water samples (0.1–0.5 m depth) of lakes and ponds spanning the same geographic range than the habitats sampled for amplicon sequencing (fig. 1). Twenty-four strains were obtained previously by using the acclimatization method, which provides a slow transition from the low environmental substrate concentrations to the high concentration of standard microbial media (Hahn et al. 2004), without employing strategies for species-specific targeted isolation (Hoetzinger et al. 2019). For further isolation, we focused on the geographic regions defined by isolation sites of these initial 24 strains. Ninety-six additional strains were obtained by targeted isolation. Large numbers of liquid cultures were established in 24- or 96-well cell culture plates by using the acclimatization method. Potential *Polynucleobacter* cultures were screened for *P. paneuropaeus* using species-specific PCR primers (det-F15-hemF-F2 5′-GTCGAATACAACTTGATTTACGATC-3′, det-F15-hemF-R 5′-TTCCGGAGAGCCCGC-3′) targeting the *hemF* gene encoding the oxygen-dependent coproporphyrinogen-III oxidase of *P. paneuropaeus* by quantitative PCR with SYBR Green detection. The affiliation of obtained strains to the species *P. paneuropaeus* was confirmed for the genome-sequenced strains by genome-wide average nucleotide identity analyses (gANI >95%, see below).

Genome Sequencing, Assembly, and Annotation

Genomic DNA of strains was extracted as described previously (Hahn et al. 2017). Apart from six previously published genomes (Hoetzinger et al. 2019), which were sequenced and assembled by a hybrid approach (Roche GS FLX and Illumina), genomes were sequenced by an Illumina MiSeq (paired-end, 300 bp) or an Illumina HiSeq instrument (paired-end, 150 bp) by Eurofins Genomics, Ebersberg, Germany. Reads which passed the Illumina chastity filter (average Phred score of 34.1 with 86.9% of all reads with Phred >Q30) were assembled using SPAdes v.3.10 (Bankevich et al. 2012) in the mismatch careful mode (including read error correction) with six k-mer lengths (21, 33, 55, 77, 99, and 127). Assembled genome sequences were annotated using the Integrated Microbial Genomes-Expert Review (IMG/ER) annotation pipeline v.4.15.1–v.5.0.3 (Markowitz et al. 2012). Genome completeness was assessed by following the lineage-specific workflow in CheckM v.1.0.12 (Parks et al. 2015).

Core Gene Alignment

Protein-coding genes shared among the 113 strains were identified using Roary v.3 (Page et al. 2015), applying a BlastP identity threshold of 95%. The software was run with the "-s" option, that is, paralogs were not split but grouped into the same protein cluster. Amino acid and nucleotide sequences of genes that did not contain any paralogs and were present in all 113 strains were aligned with MAFFT v.7 (Katoh and Standley 2013) using the "ginsi" option. The individual gene alignments were concatenated using the catfasta2phyml.pl script (https://github.com/nylander/catfasta2-phyml, last accessed February 15, 2021), to obtain the core gene alignments of amino acid and nucleotide sequences, respectively.

Genome Similarities and Dereplication

Genome-wide average nucleotide identities (gANI) were calculated using the Integrated Microbial Genomes-Expert Review (IMG/ER) analysis system (Markowitz et al. 2012). The number of pairwise nucleotide and amino acid differences between strains were calculated from the core gene alignment using the MEGA X software (Kumar et al. 2018). Groups of strains with more than 99.9% gANI were reduced to one reference genome if they were isolated from the same sample. For each such group, the genome containing the lowest number of scaffolds was used as reference. The 113 genomes were thereby dereplicated to a set of 90 genomes (supplementary table S1, Supplementary Material online), which was used in all further analyses.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

Phylogenetics

Phylogenetic trees were calculated with RAxML v.8 (Stamatakis 2014) using the "-f d" option (new rapid hill-climbing algorithm) and the GTRGAMMA substitution model. Bootstrap support was calculated for 200 replicates using "-b" and "-N" with the same settings than for calculating the initial tree. Bootstrap values were then drawn on the initial tree using the "-f b" option. Division of the core gene alignment into windows containing 1,000 and 5,000 segregating sites each, respectively, was done using scripts of the "core gene sweeps" module of the PopCOGenT software package (Arevalo et al. 2019). Scripts of the same module were used to test for monophyly of the Alp population in the generated trees. To quantify incongruence among phylogenetic trees, topological distances were calculated according to the Kuhner–Felsenstein branch length method (Kuhner and Felsenstein 1994) using the dist.topo function of ape v.5.3 (Paradis and Schliep 2019) in R (R Core Team 2019). Trees were midpoint-rooted for visualization using phangorn v.2.5.5 (Schliep 2011) and plotted using ggtree v.2.0.3 (Yu et al. 2017).

Population Genomics

A shortened alignment containing only the segregating sites was extracted from the core gene nucleotide alignment using SNP-sites (Page et al. 2016), resulting in an alignment length of 79,886 bp. Fixation indices (F_{ST}) were calculated from this alignment with Arlequin v.3.5 (Excoffier and Lischer 2010) using 10,000 permutations to test the significance.

Population structure was inferred from the core gene nucleotide alignment using Structure v.2.3.2 (Pritchard et al. 2000). The xmfa2struct program was used to convert the nucleotide alignment into the input file format of Structure. The software was run under the admixture model and the correlated allele frequency model, using the default $\lambda=1$. To estimate the number of populations (K), the Markov chain Monte Carlo (MCMC) scheme was run two times testing K from 1 to 8, using 50,000 iterations for burn-in and data collection each. As variability of estimated likelihoods between the runs was relatively high compared with variability among estimated likelihoods for different K, three further runs with 100,000 iterations for burn-in and data collection each were executed. Figure 4*B* shows the results of a run with $K=5$, visualized using Structure Plot v.2.0 (Ramasamy et al. 2014). The value of $K=5$ is not supposed to represent the "correct" number of populations, but was chosen because it appeared to capture most of the structure in the data and yielded the highest mean estimated likelihood among the tested values (supplementary table S2, Supplementary Material online). The populations inferred by Structure are referred to as genomic groups in the Results section in order to distinguish them from the populations demarcated by geographic origin.

Recombination and Gene Flow

Three complementary programs, PopCOGenT (Arevalo et al. 2019), ClonalFrameML (Didelot and Wilson 2015), and ConSpeciFix (Bobay et al. 2018) were used to assess recombination and reconstruct gene flow among *P. paneuropaeus* strains.

PopCOGenT was designed to infer recent gene transfer and identify gene-flow discontinuities that delineate populations. As it performs pairwise genome comparisons, gene transfer is detected across all genome regions shared by any two strains, that is, both core and flexible gene pool components are considered. The main measure of gene transfer calculated by this software is the so-called length bias. It is assumed that recent transfers between two genomes are reflected by stretches of identical genomic regions, which are longer than would be expected under clonal evolution. The length bias quantifies this difference between the observed distribution and its expectation under a clonal model. Importantly, the signal of length bias quickly decays due to mutational accumulation, that is, within the time it takes for 0.001 mutations to accumulate per site per genome according to simulations (Arevalo et al. 2019). This suggests that the length bias is an appropriate measure to differentiate recent from historical gene flow. PopCOGenT was run on the 90 genomes as described at https://github.com/philarevalo/PopCOGenT (last accessed February 15, 2021). To illustrate the results in a network, nodes (genomes or populations, respectively) were arranged according to calculated length biases using the ForceAtlas2 algorithm in Gephi (Bastian et al. 2009). Gene flow within and among populations was quantified as the average length bias from the respective genome pairs.

In contrast to PopCOGenT reconstructing gene flow between pairs of genomes by identifying identical genome segments, ClonalFrameML reconstructs imports from an external source by identifying segments containing a higher than expected number of substitutions. ClonalFrameML infers such recombination events on the branches of a presumed clonal genealogy using a maximum likelihood algorithm (Didelot and Wilson 2015). The starting tree (initial genealogy) for the ClonalFrameML algorithm was generated using RAxML as described above under phylogenetics. The underlying alignment of the *P. paneuropaeus* genomes was generated using progressiveMauve v.2.4.0 (Darling et al. 2010) with default settings. The stripSubsetLCBs script that is part of the Mauve package was used to leave only alignment blocks longer than 500 bp, yielding an alignment of 1,356,560 bp. ClonalFrameML was first run using the standard model assuming that recombination parameters are the same for all branches of the tree, that is, using default settings and the number of simulations set to 100 using the "-emsim" flag. To allow for heterogeneity in the recombination process and estimate recombination parameters per branch,

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

ClonalFrameML was run a second time using the per-branch model, that is, setting "-embranch true," "-embranch_dispersion 0.1," and specifying initial values of R/θ, 1/δ, and ν according to the results from the first run by using the "-initial_values" flag. Results from the per-branch run were plotted using the cfml_results.R script with the packages phangorn v.2.5.5 (Schliep 2011) and ape v.5.3 (Paradis and Schliep 2019) in R (R Core Team 2019). The ratio of substitutions introduced by recombination relative to mutation (*r/m*) across all branches of the tree was calculated from the posterior means of R/θ (recombination rate relative to mutation rate), δ (mean length of recombined segments), and ν (mean number of substitutions per site in recombined segments) obtained from the per-branch run using the formula $r/m = (R/\theta) \times \delta \times \nu$.

Similar to ClonalFrameML, ConSpeciFix analyzes homologous recombination in the core genome, yet, recombination events across the population under study rather than imports from external sources are considered. Recombination rates are assessed by inference of homoplasic polymorphisms (*h*) relative to nonhomoplasic polymorphisms (*m*). Homoplasies are polymorphisms that are incompatible with vertical inheritance from a single ancestor, that is, resulted either from recombination or convergent mutations. To illustrate the proportion attributable to convergent mutations, expected *h/m* ratios when homoplasies are generated in the absence of recombination are calculated. The program was run on the reduced set of 90 *P. paneuropaeus* genomes as described at https://github.com/Bobay-Ochman/ConSpeciFix (last accessed February 15, 2021) under "Personal Comparison". To illustrate interspecies recombination barriers, the set of 90 genomes was tested in another run together with the genome of *P. asymbioticus* strain QLW-P1DMWA-1^T (Hahn, Schmidt, et al. 2016). To compare pan-European and regional gene flow, the program was additionally run on the genomes of each population separately.

Biogeography

Spearman rank correlations between genetic distances among the 90 *P. paneuropaeus* strains and spatial distances, pH, or temperature differences among their sites of origin were assessed in Mantel tests with 1,000 permutations using the vegan v.2.5-6 package (Oksanen et al. 2019) in R (R Core Team 2019). Genomic distances were represented by nucleotide differences in the core gene alignment (see above). Analogous analyses were conducted using the *priB* amplicon data, with genetic distance represented by F_{ST} values between samples calculated using Arlequin v.3.5 (Excoffier and Lischer 2010). Here, samples were reduced to a set that yielded only significant ($P < 0.05$) F_{ST} values, resulting in 42 samples with 411–149,022 (median 23,601) *P. paneuropaeus* reads retained for the analysis. Partial Mantel tests were computed to control for potential effects of pH, temperature, and conductivity differences on the correlation between genetic and spatial distance. Mantel Correlograms (Oden and Sokal 1986; Borcard and Legendre 2012) were computed to assess the correlations between genomic and spatial distances in 20 equidistant distance classes. Here, *P* values were corrected for multiple testing using the Holm method.

Results

In Situ Abundance of *Polynucleobacter paneuropaeus*

Amplicon sequencing was performed in order to obtain insights into the relative importance of *P. paneuropaeus* populations within the 3,400 km wide cross-section through Europe (fig. 1) and along a pH gradient (fig. 2*A*). The primers used for generating the amplicons are specific for the species-rich subcluster PnecC of the genus *Polynucleobacter* (Hahn, Jezberová, et al. 2016). The obtained reads were clustered into species-like OTUs by using a 98% threshold. The *priB* sequences of the 113 cultured *P. paneuropaeus* strains share sequence similarities in the range of 98.3–100% (average 99.6%). Amplicons of 117 samples representing 99 habitats were analyzed. In total, 600 species-like *Polynucleobacter* OTU$_{98\%}$ were detected in the samples, among which *P. paneuropaeus* is the most abundant taxon contributing 11.3% of all reads. The relative abundance (% of OTU-specific *priB* reads among all *Polynucleobacter priB* reads) of *P. paneuropaeus* in the samples ranges from 0% (0 reads) to 99.6% (25,126 reads) (fig. 2*B*). Importantly, regional-specific differences in relative abundances of *P. paneuropaeus* were observed (fig. 3*A*). The vast majority of *priB* reads of *P. paneuropaeus* originate from habitats in the pH range 4.5–7.0 (fig. 2*A*), however, isolates affiliated with this species were obtained from a slightly broader pH range (supplementary fig.S1*C*, Supplementary Material online). On average, in that pH range, reads of *P. paneuropaeus* comprised 23.8% (SD 31.2%) of reads of all members of the *Polynucleobacter* subcluster PnecC. Based on data obtained previously by fluorescent in situ hybridization (FISH) from lakes and ponds of a broad pH range (Jezbera et al. 2012), which even include some samples investigated here by *priB* amplicon sequencing, the absolute abundance of *P. paneuropaeus* can be roughly estimated. Jezbera et al. reported relative abundances of the *Polynucleobacter* subcluster PnecC of 10–20% for the pH range 4.5–7.0, and absolute bacterial abundances of about 2×10^6 cells/ml. Thus, for this pH range, average absolute abundances between 5×10^4 and 1×10^5 *P. paneuropaeus* cells/ml could roughly be estimated. In other words, across this pH range preferred by the species, *P. paneuropaeus* is assumed to represent about 2–5% of bacterioplankton cells.

Isolation of Strains and Genome Characteristics

In total, we obtained 120 *P. paneuropaeus* strains from 53 different freshwater lakes and ponds located within the

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

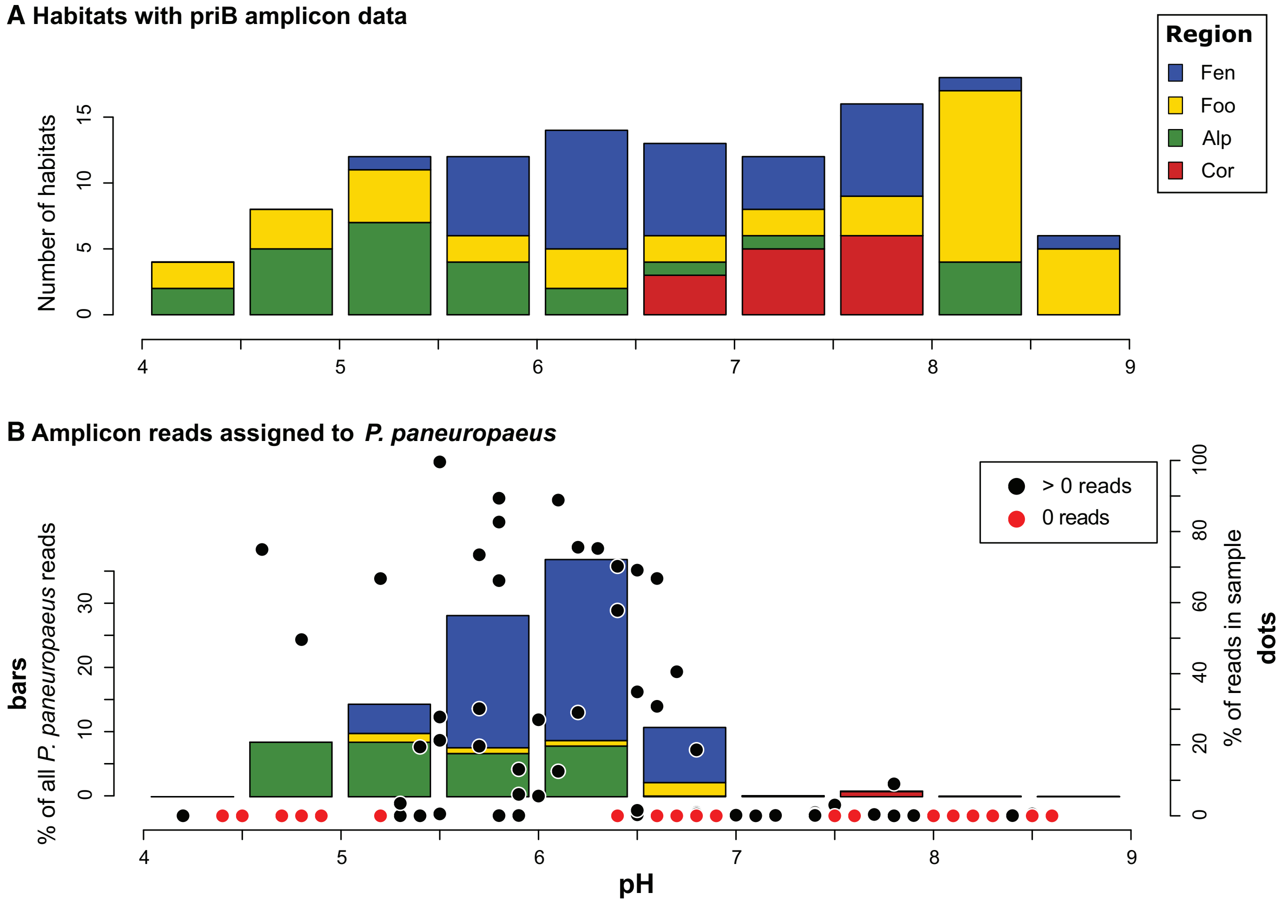


Fig. 2.—pH distribution of *P. paneuropaeus*. Relative abundances in 99 freshwater habitats were determined by *priB* amplicon sequencing. The regions of origin of samples and strains are color coded (compare fig. 1). (*A*) The investigated water samples cover a pH gradient ranging from pH 4.2 to 8.6. (*B*) Relative *P. paneuropaeus* abundance determined for individual samples (dots, right hand *y* axis), and average values for pH categories (bars, left hand *y* axis). These relative abundance data represent the relative contribution of reads assigned to *P. paneuropaeus* among the total number of reads contributed by the whole subcluster PnecC. The species constituted a major fraction of PnecC communities in several slightly acidic habitats (pH 4.5–6.7), especially in the Fen region.

North–South cross-section ranging from Northern Fennoscandia to the Mediterranean island Corsica (fig. 1 and supplementary table S1, Supplementary Material online). The pH of samples from which the strains were obtained ranged from 4.4–7.8, conductivity from 3.9–151.3 µS/cm, water temperatures from 1.5–26.4 °C, and altitudes from 13–2,103 m (supplementary table S1, Supplementary Material online). Seven of the 120 isolated strains were not considered for genome sequencing because they shared identical partial glutamine synthetase gene (*glnA*) sequences with other strains obtained from the same water sample, thus may represent identical clones.

Genome sequencing of 113 strains resulted in 44 closed genomes (supplementary table S1, Supplementary Material online). The remaining genome sequences consist of 2–23 scaffolds (median 5). The 113 genome assemblies showed average coverage values of 91.0× (range 30×–311×, median 75.7×). Assessment of genome completeness using CheckM (Parks et al. 2015) suggests that completeness of the 113 genomes ranged from 91.8 to 93.6% (supplementary table S1, Supplementary Material online). Even the 44 genomes that assembled into a single closed scaffold scored completeness values of only 93.4% on average (SD 0.16%). This includes four genomes that were obtained from hybrid assemblies based on paired-end sequenced shotgun as well as Long Jumping Distance Libraries of 8-kb fragment size (Hoetzinger et al. 2019), thus, with high confidence in genome completeness. The incompleteness predicted by CheckM is most likely reflecting evolutionary reduction of gene content (genome streamlining), as *Polynucleobacter* bacteria typically harbor markedly smaller genomes compared with other members of the *Burkholderiaceae* family (Hahn et al. 2012). Considering that the lowest obtained CheckM completeness values are only 1.7% below the average of the closed genomes suggests that all 113 assembled genomes are near-complete. Genome sizes range from 1.61–

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

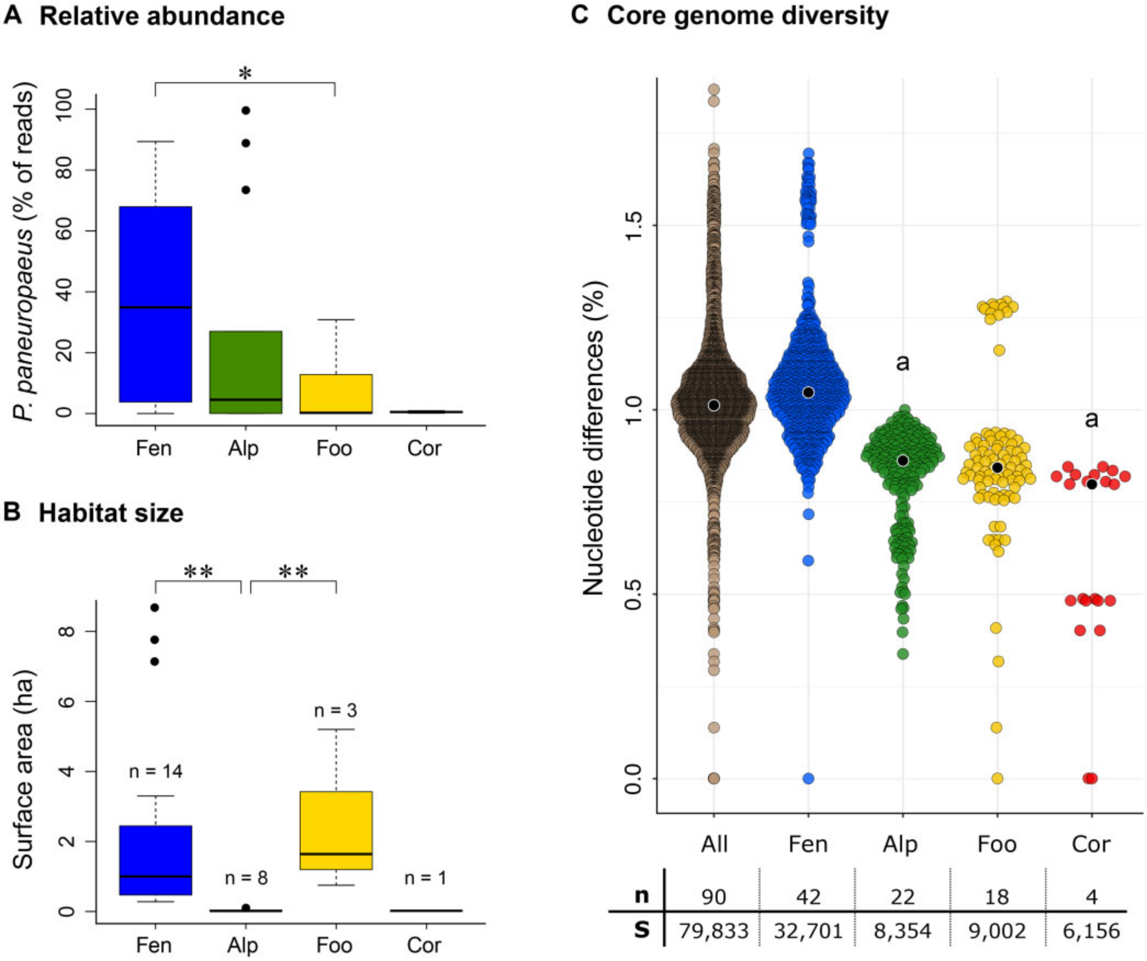


Fig. 3.—Association between population size and diversity. (*A*) Box plot depicting the relative abundance of *P. paneuropaeus* in water samples originating from different regions. Only data of samples in the pH range 5–7 were considered. Although *P. paneuropaeus* dominated the PnecC communities in many habitats in Fennoscandia, the species represented only a minor fraction of the communities in the Corsican habitats. Significant differences between populations are indicated on top of the figure (Kruskal–Wallis rank sum test: $*P < 0.05$, $**P < 0.01$). (*B*) Box plot analyzing the surface area of habitats that showed *P. paneuropaeus* abundances >1%. Besides the high relative abundances of the species in samples from Fen shown in (*A*), suitable habitats for the species are relatively large in this region, which suggests higher population sizes in Fen as compared with the other regions. Significant differences between populations are indicated on top of the figure (Kruskal–Wallis rank sum test: $*P < 0.05$, $**P < 0.01$). (*C*) Nucleotide differences of all pairwise genome comparisons and comparisons within the populations. Median values for each group are shown as black dots with white border. The number of strains (*n*) and the number of segregating sites (*S*) in each population are given below the figure. Populations that are not significantly different (Dunn's test: P_{adj}>0.05) are marked with a common letter. Fen, which is estimated to represent the largest of the studied populations (see *A* and *B*), also represents the most diverse population.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

1.93 Mb, and gene counts from 1,647 to 2,020 genes (supplementary table S1, Supplementary Material online). The minimum gANI between genomes is 96.6% (supplementary table S3, Supplementary Material online), that is, above commonly applied thresholds for species delineation of ≈ 95–96% (Konstantinidis et al. 2006; Richter and Rosselló-Móra 2009). Grouping highly similar genomes (gANI >99.9%) together resulted in 12 clusters containing multiple genomes from single samples, respectively. After reducing these clusters to one reference genome each, we obtained a set of 90 genomes (supplementary table S1, Supplementary Material online) that was used in further analyses. Gene content analysis using Roary revealed an open pan-genome containing 7,255 protein clusters (supplementary fig. S3, Supplementary Material online), 1,151 of which are present in all isolates. Eighty of the 1,151 core genes are paralogs, that is, present in multiple copies in at least one strain, and 16 genes exhibit major size differences among strains. After excluding these, 1,055 genes were obtained to generate the core gene alignments. The concatenated nucleotide alignment comprises 1,007,813 bp, including 79,886 segregating sites, and a maximum pairwise difference of 18,831 bp. The amino acid alignment contains 335,948 total and 18,597 segregating sites, and a maximum pairwise difference of 2,860 amino acids (supplementary table S4, Supplementary Material online).

Core Genome Phylogeny and Genome Mosaicism

The phylogenetic tree calculated from the core gene alignment (1,007,813 bp) shows that strains tend to cluster according to their geographic origin (fig. 4*A*). Strains isolated from the Alps even appear as monophyletic group, whereas the other geographic groups are polyphyletic. However, bootstrap values for the deeper nodes are low throughout the tree. The weak support for a single phylogenetic tree is reflected by totally differing phylogenies of trees calculated from alignments of different regions of the genome. This is exemplified by phylogenetic trees representing the core gene alignment divided into eight windows containing 5,000 segregating sites each, or 42 smaller windows of 1,000 segregating sites each, which reveal extensive genome mosaicism (supplementary fig.S4, Supplementary Material online). In contrast to the tree representing the whole core genome, strains isolated from the Alps are not monophyletic in any of the trees calculated from the split alignment, and clustering according to geographic region is less pronounced. Topological distance scores among the trees are given in supplementary table S5, Supplementary Material online. The incongruence among the trees indicates genome mosaicism and points to pronounced homologous recombination and admixture among geographic regions.

Population Structure and Gene Flow

In line with the weak population structure and admixture suggested by phylogenetic analyses, the Structure software (Pritchard et al. 2000) did not reveal distinct genomic groups but assigned multiple groups to most strains (fig. 4*B*). Nevertheless, four of the five genomic groups inferred by the structure analysis could be roughly assigned to geographic regions, that is, Corsica (Cor), Alps (Alp), Foothills of the Alps (Foo), and Fennoscandia (Fen). The fifth group is mainly represented by three strains exhibiting exceptionally long branches in the phylogenetic tree, that is, strains AP-Hunz-20-A7, TSB-Sco10W11, and JS-Vitta2-80-C6. Strains isolated from Fen appear to be particularly diverse, that is, different genomic groups contribute substantial variability to all of those strains (fig. 4*B*). This diversity is also reflected in high intrapopulation nucleotide diversities. Average nucleotide differences of pairwise comparisons within Fen (1.07%) are in the range of differences among the whole metapopulation (1.03%, fig. 3*C*). Average differences within the other populations are significantly lower (Foo: 0.87%, Alp: 0.83%, Cor: 0.61%).

Calculation of fixation indices (F_{ST}) rejected the null hypothesis of no population differentiation for all pairwise comparisons ($P < 0.005$), however, the obtained F_{ST} values ranging from 0.07 to 0.27 are relatively low (supplementary table S6, Supplementary Material online). Highest F_{ST} values resulted from comparisons with the least diverse population Cor, yet, Cor is markedly less differentiated from Fen (0.14) than from the other populations (0.25 and 0.27). Lowest F_{ST} values concern comparisons with Fen (0.07–0.14).

Population differentiation might possibly be blurred by historical gene transfer, that is, events that occurred before gene flow barriers separated diverging populations. To test for this possibility, we used the software package PopCOGenT that infers population structure by retracing only recent gene flow events (Arevalo et al. 2019). In contrast to the phylogenetic and population structure analyses described above that refer to the core genome, gene flow in both the core and the flexible gene pool is considered here, as the method does not rely on a multigenome alignment but uses pairwise genome comparisons. Based on the length bias parameter, this software detected plenty of recent gene transfers between all strain pairs (supplementary table S7, Supplementary Material online), and consequently assigned all strains to a single gene flow unit. This distinguishes *P. paneuropaeus* from, for example, a terrestrial bacterial species, where even certain co-occurring populations were differentiated into different gene flow units based on PopCOGenT (Chase et al. 2019). Three pairs of genomes (AP-Matth-20-A10/AP-Matth-80-A6, FUKU-SW-23/FUKU-NW11, and SM1-W8/AP-RamGG-20A-D9) exhibited exceptionally high length biases and were further assigned to separate subclusters by the program. The networks constructed based on length bias illustrate that gene flow tends to be more pronounced within than between geographic regions (fig. 5*A* and *B*). Yet, gene flux within populations calculated as average length bias of the respective strain pairs are in a similar range than between populations (fig. 5*B*). Effective gene flow across the whole geographic range is further demonstrated by the distribution of length bias across spatial distance (fig. 5*C*). Moreover, length biases within the whole metapopulation ranging from 34–77,923 (supplementary table S7, Supplementary Material online) are high in comparison to previously presented values for, for example, nonrecombinogenic *Salmonella* or *Corynebacterium* bacteria but also in relation to recombinogenic bacteria affiliated to *Salmonella*or *Vibrio* (compare with Figure 1C in Arevalo et al. 2019).

We used ClonalFrameML to quantify recombination imported from sources outside the studied metapopulation (supplementary fig.S5, Supplementary Material online). The inferred R/θ, δ, and ν were 0.442, 65.6, and 0.0920, respectively. Thus, the *r*/*m* ratio representing the effect of recombination relative to mutation was 2.67. This indicates that DNA import from external sources introduced 2.67 times more

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

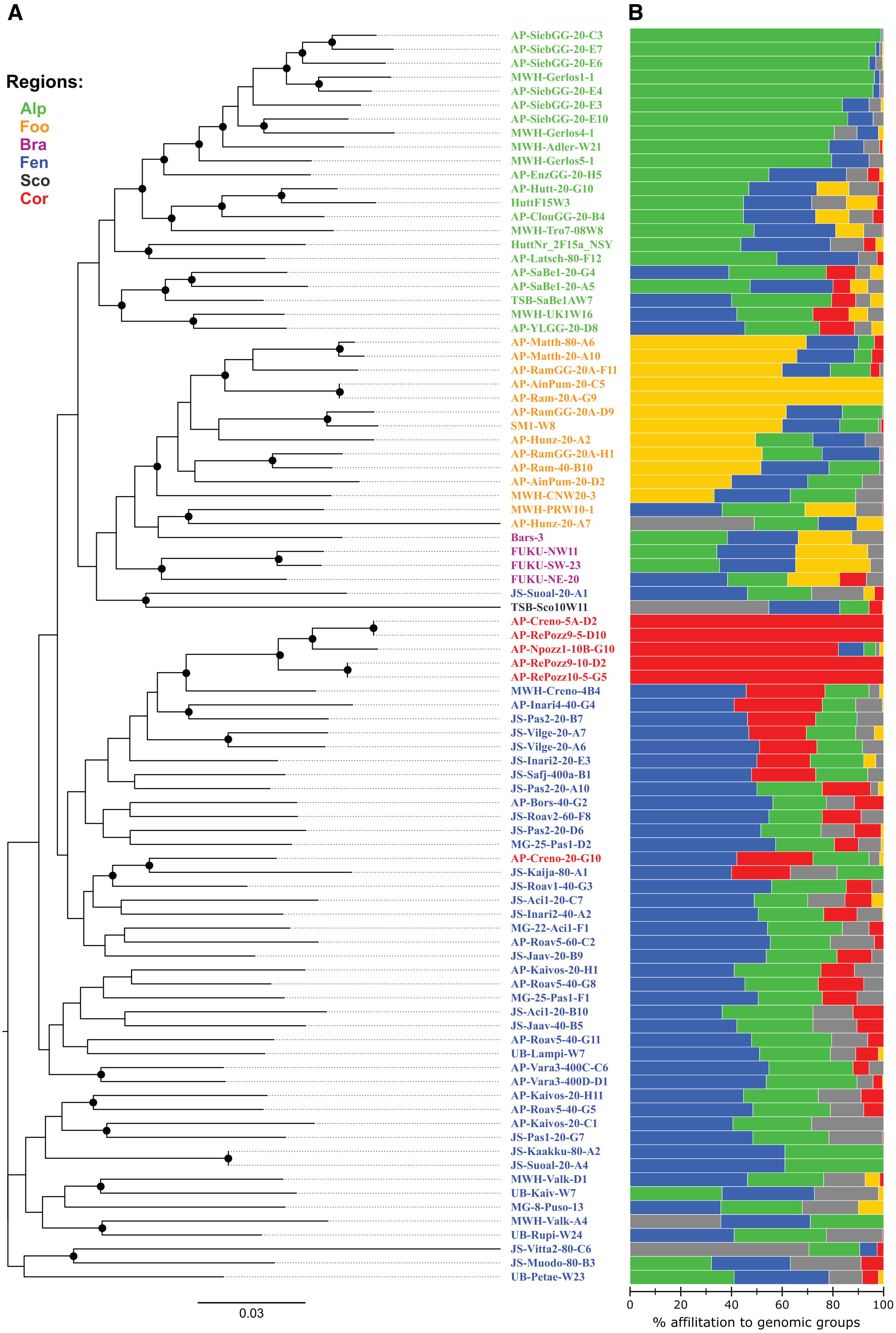


FIG. 4.—Blurry geographic population structure revealed by (*A*) phylogenetics and (*B*) analysis of nucleotide polymorphisms. (*A*) Midpoint-rooted RAxML tree based on the concatenated nucleotide alignment of 1,055 core genes. Nodes with bootstrap support >80% are indicated by black circles. Strain names are colored according to geographic origin. Highly similar strains (>99.9% gANI) were excluded, except if they stem from different samples (e.g., JS-Kaakku-80-A2 and JS-Suoal-20-A4). (*B*) Genomic population structure inferred by the Structure software under the admixture model using the same alignment than for the phylogenetic tree. The five inferred genomic groups are colored according to the geographic regions dominating them. The gray group could not be assigned to a specific region. Geographic origin and genomic grouping are not in flawless accordance, which reflects the admixture among bacteria from different regions.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

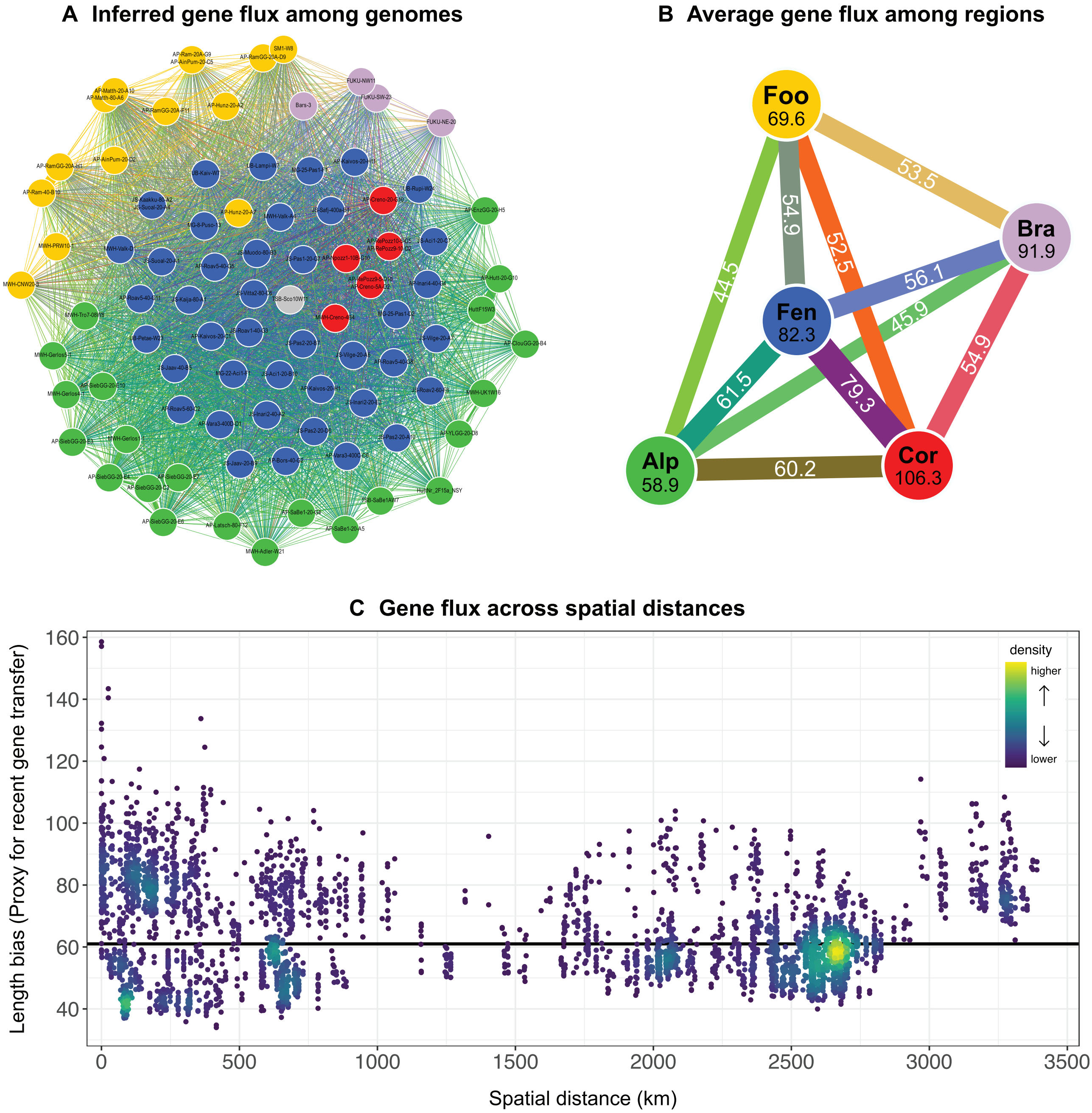


Fig. 5.—Recent gene transfer inferred by PopCOGenT. (*A*) Network of gene transfer among all 90 genomes. Strains (nodes) are colored according to geographic origin and have been arranged using the ForceAtlas2 algorithm in Gephi. Here, attraction between nodes was proportional to the respective length bias (proxy for gene transfer), which is also represented by edge thickness. Note that all genomes are connected with each other, illustrating that gene flow ranges over the whole metapopulation (compare with Figure 2 in Chase et al. 2019 or Figure 3 in Arevalo et al. 2019). Strains from Fennoscandia presumably representing the largest population are located in the center of the network, which may indicate that excessive gene flow spreads from this region to satellite populations further south. (*B*) Gene flow between and within populations. The network has been arranged similarly to the one in (*A*). Here, average length bias between populations determined attraction between nodes and edge thickness, respectively, and the respective numbers are given for each edge. Average length biases within populations are given for each node. (*C*) Gene transfer versus spatial distance. Each dot represents a genome pair. Dots are colored according to the density of nearby data points to account for indiscernible data due to overlapping dots. The black horizontal line indicates the median. Pairs with gANI >99% were excluded from this plot because length bias was disproportionately high (up to 77,923) for very closely related genomes (see supplementary fig. S13, Supplementary Material online), which might represent an artifact of the method rather than actual gene transfer. Extensive recent gene transfer was detected even among strains from the most distant habitats. Thus, dispersal limitation does not seem to isolate populations within this geographic range.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

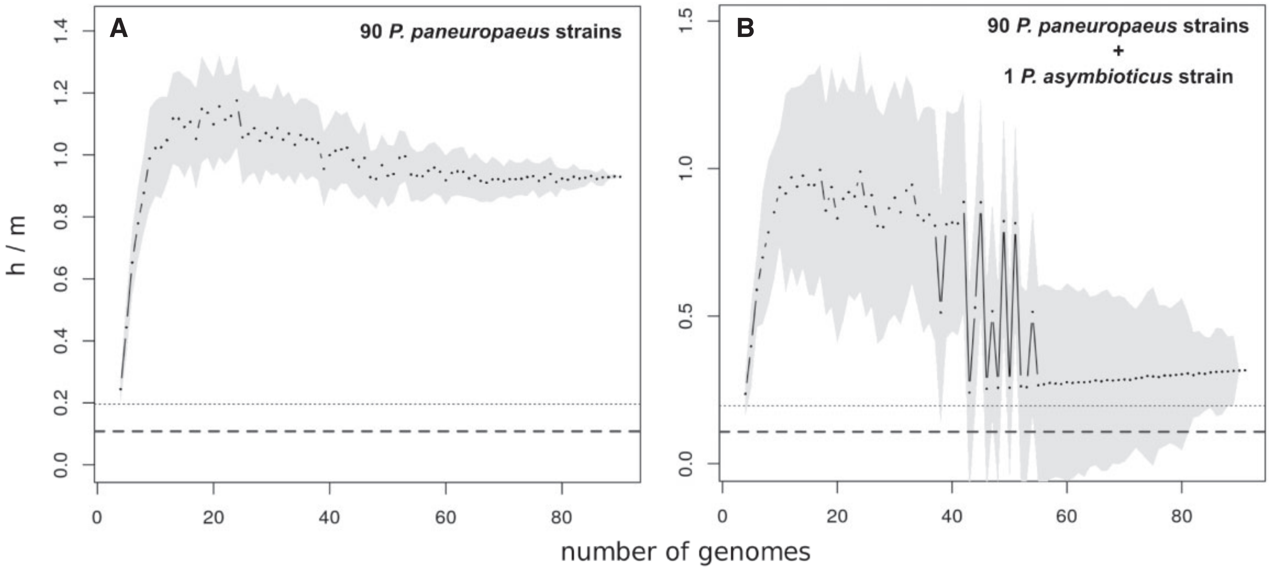


Fig. 6.—ConSpeciFix results showing that *P. paneuropaeus* represents a cohesive species due to high intraspecific homologous recombination rates. (*A*) Inferred numbers of homoplasic relative to nonhomoplasic polymorphisms (*h/m*) across randomly sampled subsets of the 90 *P. paneuropaeus* genomes. The dashed line at the bottom shows the expected *h/m* ratio and the dotted line above its upper 95% confidence bound in the absence of recombination, that is, when homoplasies are generated through convergent mutations only. The ratios for the *P. paneuropaeus* genomes well above the reference lines indicate extensive homologous recombination among the whole metapopulation. (*B*) The same analysis runs with the 90 *P. paneuropaeus* genomes and one additional *P. asymbioticus* genome. The decline of *h/m* ratios with higher numbers of genomes included in the sampled subsets reveals interspecies recombination barriers.

substitutions as did mutation, which corroborates a high propensity of *P. paneuropaeus* for recombination.

Gene transfer among individuals of a species effectively counteracts population divergence if the number of polymorphisms spread through such transfer is high in relation to the number of polymorphisms introduced through mutation. To estimate this relation and assess the effects of conspecific gene exchange on species coherence we used the ConSpeciFix software (Bobay et al. 2018). Here, gene transfer refers to homologous recombination among genes in the core genome, and is quantified based on the number of homoplasic alleles (*h*) and nonhomoplasic alleles (*m*). Based on the *h/m* ratio it can be estimated whether the tested genomes form a coherent species, that is, that intraspecific recombination accounts for genomic cohesion according to the Biological Species Concept. The program determined that all but one strain (AP-Hunz-20-A7) are members of the same species defined as such (fig. 6*A* and supplementary fig.S6*C*, Supplementary Material online). Including a *P. asymbioticus* strain in the analysis results in assignment of all *P. paneuropaeus* strains to the same species (fig. 6*B* and supplementary fig.S6*D*, Supplementary Material online). This indicates that recombination barriers to this strain belonging to a species frequently syntopic with *P. paneuropaeus* and sharing similar pH preferences have been yet much more pronounced than to AP-Hunz-20-A7. In line with the PopCOGenT results, the high *h/m* ratios for the whole metapopulation do not reveal strong gene flow barriers among geographic regions, although populations grouped by geographic origin approximate slightly higher ratios than the whole metapopulation (supplementary fig.S7, Supplementary Material online).

Biogeographic Patterns

To further investigate if geographic barriers to dispersal limit gene flow among distant habitats, we correlated core-genome differentiation between strains to air-line distance between their home habitats (fig. 7). Although we observed a weak (Spearman's rho=0.19) but significant ($P < 0.001$) positive correlation across the whole dataset (table 1), the Mantel correlogram revealed that this is mainly attributable to a strong correlation at short spatial distances, that is, strain pairs within the first distance class are genomically more similar than pairs among other distance classes (fig. 7*B*). Already the second distance class (170–340 km) lacks significant correlation, and the pattern is not uniform across longer spatial distances. Although the strain pairs from certain distance classes between 700 and 2,500 km show more nucleotide differences than expected by chance (significant negative correlation), a positive correlation is again observed for distances around 2,800 km. Thus, an isolation by distance pattern across the whole geographic range is not apparent.

To check if environmental differences explain variation in genomic dissimilarity, the association between nucleotide differences and pH, temperature, or conductivity was tested,

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

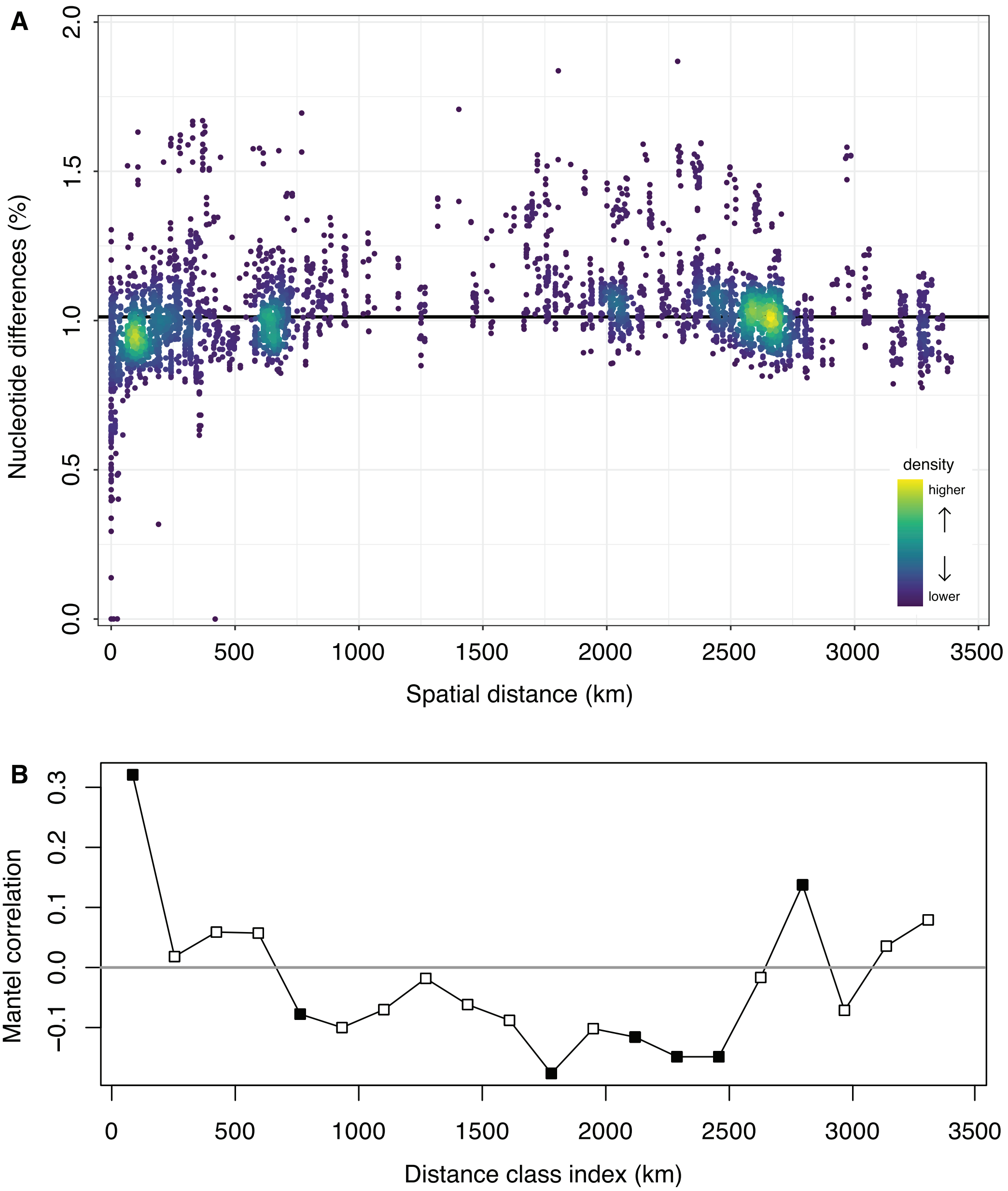


Fig. 7.—Isolation by distance pattern at short spatial scales diminishing along greater distances. (*A*) Nucleotide differences between strains are plotted against spatial distances of their home habitats. Each dot represents one strain pair. Nucleotide differences were determined from the core genome alignment. Dots are colored according to the density of nearby data points to account for indiscernible data due to overlapping dots. The black line indicates the median nucleotide difference of all pairs. Mantel test results for the correlation of nucleotide difference versus spatial distance are given in table 1. Note that highly similar genomes (>99.9% gANI) isolated from the same sample were excluded from these analyses to discount the effect of local clones. (*B*) The Mantel correlogram shows spatial correlation values (*y* axis) as a function of spatial distance classes (*x* axis). The correlogram was calculated based on the data shown in (*A*). The Mantel correlation values indicate the deviation of within-class genomic similarities from among-class similarities. Positive correlations indicate higher within-class similarities than expected by chance and vice versa. Significant correlations ($P < 0.05$) are depicted by filled squares. The correlation value decreases to negative already at relatively short spatial distances and is even positive for the distance class around 2,800 km. This indicates an isolation by distance pattern for short spatial distances that disappears for longer spatial ranges.

which all did not reveal significant correlation (table 1). Accordingly, controlling for pH and temperature differences in partial Mantel tests had minor effects on the correlation between genomic and spatial distance.

The isolated strains constitute only a small fraction of the *P. paneuropaeus* populations of the sampled habitats and it is unclear how well they represent the diversity present in different habitats. The *priB* amplicon dataset allowed us to test

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

Table 1
Mantel Test Results

Correlated matrices	Rho	*P*
nuc vs. spa	0.19**	0.001
nuc vs. pH	0.04	0.149
nuc vs. T	0.06	0.157
nuc vs. cond	0.11	0.058
F_{ST} vs. spa	0.12**	0.009
F_{ST} vs. pH	0.11	0.097
F_{ST} vs. T	0.07	0.174
F_{ST} vs. cond	0.14	0.079
nuc vs. spa (pH)	0.19**	0.001
nuc vs. spa (T)	0.18**	0.001
nuc vs. spa (cond)	0.18**	0.001
F_{ST} vs. spa (pH)	0.11*	0.021
F_{ST} vs. spa (T)	0.11*	0.020
F_{ST} vs. spa (cond)	0.11*	0.022

Note.—Spearman rank correlations were computed between distance matrices containing nucleotide differences between strain pairs (nuc) or F_{ST} values between pairs of amplicon samples (F_{ST}) and spatial distances (spa), pH differences (pH), temperature differences (T), and conductivity differences (cond) between the corresponding habitats/samples. Rows and columns of the nuc or respectively F_{ST} distance matrix were permuted to calculate the *P* values. The last six rows show the results of partial Mantel tests, where the correlation between nuc or respectively F_{ST} and spa was conditioned on a third matrix that is given in brackets. Significant correlations are marked by asterisks (*$P < 0.05$, **$P < 0.01$).

for signs of isolation by distance with larger samples sizes, albeit considering only a tiny fraction of the genomes (285 bp). Here, population differentiation between habitats was quantified by F_{ST} values (supplementary fig.S8, Supplementary Material online). A correlation between F_{ST} values and spatial distance was observed, similarly as in the analysis of individual genomes described above (table 1). The significance of the correlation decreased when the effect of environmental variables (pH, temperature, or conductivity) was removed in partial Mantel tests. Overall, the association between population differentiation and spatial distance appears to be weak (supplementary fig.S8, Supplementary Material online).

We further investigated if certain *priB* alleles were confined to certain geographic regions, which would indicate gene flow barriers. The same three alleles account for 80% of the 3.2×10^5 *P. paneuropaeus* reads obtained from the 114 habitats and for 81% of the sequences obtained from the 113 isolated strains, respectively. The relative abundances of these alleles in the amplicon dataset and among the isolated strains are similar (supplementary fig. S9, Supplementary Material online), which does not point to an amplicon sequencing nor to a cultivation bias within *P. paneuropaeus*. The three alleles were detected in all geographic regions of the amplicon dataset (supplementary fig. S9, Supplementary Material online). Rarer alleles were detected in only one geographic region each. However, the low abundances of these rare alleles prohibit conclusions on their geographic distribution, that is, potential detection upon further sampling cannot be excluded. In conclusion, the cross-regional distribution of the most prominent *priB* alleles supports the notion of continental-scale gene flow and an absence of strong dispersal barriers.

Discussion

We used both cultivation-independent and genomic data from isolated strains to assess different factors that determine gene flow among geographically separated *P. paneuropaeus* populations. Although admixture across the studied range apparently shaped its evolution and high recombination rates in the core-genome entailed species coherence, a certain degree of geographic population structure was revealed. It is worth noting that this population structure would have been missed in analyses based on single gene or multilocus sequencing, which is exemplified by phylogenetic trees in supplementary figure S10, Supplementary Material online. This highlights the importance of using genomic data to resolve biogeographic patterns on species level. The role of region-specific differences in populations densities, dispersal limitation, and historic colonization events in structuring the observed genomic variation will be discussed below, particularly with regard to the potential for allopatric speciation.

Geographic Center of *P. paneuropaeus* Supposably in the Boreal Zone

We used relative abundance data from cultivation-independent amplicon sequencing for gaining insights into the ecological preferences of *P. paneuropaeus*. The species was detected across the entire studied geographic range. A preference for slightly acidic freshwater systems is obvious (fig. 2*B*), yet, pH alone cannot explain the relative abundance differences observed among the different regions (fig. 3*A*). For instance, habitats in the region Foo showed rather small relative *P. paneuropaeus* abundances, even though many of those samples originate from the obviously preferred pH range (fig. 2*A*). Although we did not quantify the distribution of habitats within the sampled regions that are suitable for the species, it is apparent that the regions differ considerably in frequency and size of such habitats. The region Fen, mainly located in the boreal zone, is generally rich in freshwater habitats and comprises many lakes and ponds matching the habitat preferences of the species. Typical *P. paneuropaeus* habitats in this region are smaller lakes surrounded by boreal forest (supplementary fig.S1*A*, Supplementary Material online). Similar habitats also inhabited by this species can be found in the region Alp, however, such small lakes surrounded by coniferous forest with slightly acidic water are rather rare in this region. Typical habitats of *P. paneuropaeus* in the region Alp are shallow ponds located in sparsely forested sites or located above the timber line (supplementary fig.S1*B*, Supplementary Material online).

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

Compared with the boreal lakes inhabited by *P. paneuropaeus*, these ponds are less numerous and usually characterized by a much smaller surface size and volume (fig. 3*B*). The regions Alp and Fen have in common that all their habitats are located on crystalline bedrock, thus in limestone-poor environments. This results in rather low buffer capacity of their water, which favors pH values in the acidic range. By contrast, many freshwater habitats of the regions Foo and Cor possess a geological background resulting in pH values above 7, which is obviously not preferred by *P. paneuropaeus*. The few habitats of these regions with relative abundances of the species larger than 1% are either located on crystalline bedrock or are influenced by peat bogs, however, habitats assumed to be suitable for *P. paneuropaeus* are much rarer in Foo and Cor compared with Fen and Alp. Based on the higher abundances of *P. paneuropaeus* in samples from the region Fen (fig. 3*A*), the larger average size of habitats colonized by the species in this region (fig. 3*B*), and the higher number and density of habitats suitable for the species in Fen, we have to conclude on large differences in population sizes per area among regions. The entire *P. paneuropaeus* population in Fen is assumedly multiple times larger than the populations of the other regions. *Polynucleobacter paneuropaeus* appears to be an abundant boreal species with a metapopulation center in the boreal zone and satellite populations further south and probably also further north in tundra ponds. These differences in regional population sizes and region-specific differences in spatial distribution of habitats harboring larger *P. paneuropaeus* populations presumably shape the metapopulation structure of the species and the genomic diversity within regional populations.

Frequent Gene Flow Prevents Allopatric Divergence

Almost all lakes and ponds sampled in this study lack regular inflow and outflow of surface water, that is, dispersal among habitats by running waters is very limited. Previous studies indicate the potential for bacterial dispersal across long distances by wind. For instance, (DeLeon-Rodriguez et al. 2013) detected *Polynucleobacter* bacteria in the troposphere above the coast of California and the Gulf of Mexico. Besides passive transport by wind, bacteria may migrate as well within the feathering of water fowl. In any case, it is difficult to predict what fraction of dispersed bacteria is viable and has the potential to colonize a new habitat. It is generally conceivable that even bacteria that do not reproduce in a new habitat introduce genes into the local gene pool by recombination. Although such factors are very difficult to model or measure directly, genome analysis provides a way to observe the ultimate effects of dispersal and reconstruct the history of gene flow among populations.

To estimate dispersal limitation in our dataset, we looked for a signal of isolation by distance (fig. 7 and supplementary fig.S8, Supplementary Material online). Assuming uniform dispersal in all directions and across the whole geographic area, that is, neglecting specific factors like predominant wind direction or water fowl migration routes, the number of dispersed bacteria from a given source is inversely proportional to the square of the distance. Such an inverse-square law exists for several physical quantities (e.g., the intensity of electromagnetic radiation) and can be conceived as geometric dilution of the given quantity into 3D space. This effect applies to the direct migration of bacteria between habitats, and probably contributed to the higher similarity among strains from nearby habitats (fig. 7). Although several very closely related strains (<0.5% nucleotide differences in the core genome) were isolated from habitats less than 50 km apart, we found only two examples of close relatives isolated from more distant habitats. Strains SM1-W8 and AP-RamGG-20A-D9 were isolated from 191 km apart habitats and exhibit 0.32% nucleotide differences in the core genome, and the northern strains JS-Kaakku-80-A2 and JS-Suoal-20-A4 originate from 420 km apart habitats and revealed no difference in their core genomes. Over the whole genomes of the latter two strains, only 10 SNPs and no differences in gene content were identified. Apart from these remarkable exceptions, the great majority of strains originating from >50 km distant habitats exhibits >0.75% core genome nucleotide differences. Hence, the direct migration of bacteria between habitats may play a minor role for longer distances, but gene flow is potentially realized via intermediate habitats serving as stepping stones. This effect may be responsible for the low association of genomic and spatial distances over longer distances. The high density of lakes and ponds in Fen may particularly favor gene flow via stepping stones, which is evident by a less steep isolation by distance pattern in Fen as compared with the southern regions (supplementary fig.S11, Supplementary Material online). Overall, dispersal limitation across the studied distances does not seem to cause divergence among populations, and most genome pairs from habitats more than 3,000 km apart are even less differentiated than the median 1.01% (fig. 7*A*). This is mostly attributed to the high relatedness of the Corsican population to strains from northern Lapland (fig. 4 and supplementary table S4, Supplementary Material online). The sampled habitats in Corsica revealed the lowest abundances of *P. paneuropaeus* (fig. 3*A*), suggesting an overall small population size in this region. In accordance, Cor revealed the lowest average intrapopulation diversity (fig. 3*C*). Furthermore, genomes from Cor exhibit the smallest genome sizes (supplementary fig.S12, Supplementary Material online), a feature that is typically associated with small effective population sizes and high genetic drift (Bobay and Ochman 2017). As discussed above, lakes and ponds in Fen exhibited exceptionally high abundances of *P. paneuropaeus*, habitats apparently similar to the investigated ones are plenty in this region, and Fen most likely constitutes the largest population size among the studied regions.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

A high average nucleotide diversity and a large number of segregating sites among Fen genomes (fig. 3*C*) further supports this conclusion. Dispersal of bacteria from the boreal region might dominate the flux of *P. paneuropaeus* across whole Europe. The arrangement of the gene flow network, where the Fen population possesses a rather central position (fig. 5*A*) and accounts for highest average interpopulation gene flux (fig. 5*B*), corroborates this notion. Excessive flux from Fen and elevated importance of genetic drift in Cor due to the large and small population sizes, respectively, seems a plausible explanation for the unanticipated relatedness of the spatially distant populations. As an alternative to genetic drift, local adaptations to similar environmental conditions could explain higher genomic similarities among certain populations. An influence of pH or temperature differences on genomic differences is not suggested by the computed correlations, yet, further analysis would be necessary to assess the role of local adaptation in population differentiation. For instance, acquisition of accessory genes that provide additional functions for certain strains might lead to ecological divergence, similarly as has been suggested, for example, for closely related bacteria thriving in adjacent soil layers (VanInsberghe et al. 2015). Although most analyses in this study focused on the core genome, which represents the genome regions that promote species coherence (Hoetzinger and Hahn 2017), accessory genomic islands that are transferred even across species boundaries were shown to drive intraspecific diversification in another *Polynucleobacter* species (Hoetzinger et al. 2017). Clarifying the possibility of ecological differentiation through the accessory genome is beyond the scope of this study, yet warrants further research.

In any case, geographic separation is suggested to play a minor role in driving divergence within the studied range due to extensive gene flow even between the spatially furthermost populations. The lack of an isolation by distance signal across long spatial distances is in stark contrast to findings about *Sulfolobus islandicus* populations (Whitaker et al. 2003). The connectivity of potential habitats for these thermophilic archaea is apparently too low to enable gene flow high enough for population admixture, as is the case in *P. paneuropaeus*.

Blurry but Detectable Population Structure—A Remnant of Colonization History?

Most regions sampled in this study were covered by ice during the Last Glacial Maximum (26,500–19,000 years ago). During that time, the Fennoscandian ice sheet covered Northern Europe, passing through current days Germany and Poland at its southern boundary (Stroeven et al. 2016). South of this ice sheet, Europe was dominated by permafrost down to Southern Hungary. The Alps were covered by an ice sheet advancing into the Alpine Foreland, and also the Corsican mountains were extensively glaciated (Kuhlemann et al. 2005). Most lakes and ponds sampled here were most likely formed after the abrupt warming and rapid melting of remaining ice sheets at the beginning of the Holocene (11,700 years ago). It seems legitimate to assume that *P. paneuropaeus* is older, although reliable molecular clocks for free-living bacteria are lacking. Proposed substitution rates for host-associated bacteria vary by several orders of magnitude, ranging from approximately 10^{-5} to 10^{-8} nucleotide substitutions per site per year (Duchêne et al. 2016). Applying the mean substitution rate estimated for *Bordetella pertussis* (1.74×10^{-7} substitutions per site per year), affiliated to the same order (*Burkholderiales*) than *Polynucleobacter*, the maximum core genome sequence dissimilarity among the *P. paneuropaeus* isolates (0.0187 substitutions per site, supplementary table S4, Supplementary Material online) would correspond to 107,000 years of divergence. This suggests that most diversity within the studied metapopulation has not been generated after colonization of the newly formed habitats during the Holocene, but that substantial diversity has been introduced into these regions through migration. Possibly, initial colonization led to population bottlenecks and genetic drift in the different regions (founder effects). The blurry geographic population structure observed today may be a remnant of such founder effects, that is, initial population differentiation through genetic drift has subsequently been mitigated by interpopulation gene flow. The high *h*/*m* ratios obtained from the ConSpeciFix analysis indicate that *P. paneuropaeus* evolved similarly to a sexual species, that is, homologous recombination had a homogenizing effect on the studied metapopulation. This admixture likely characterizes the recent history of the species, given the extensive gene flow revealed by PopCOGenT, which was specifically designed to differentiate contemporary from historic events. The signal of gene transfer identified by this software decays within the time it takes for 0.001 mutations to accumulate per site according to simulations (Arevalo et al. 2019). Applying again the substitution rate calculated for *B. pertussis* estimates that the detected events have happened within the last 6,000 years. This suggests that the detected gene transfer most likely represents gene flow among already geographically separated populations, and has effectively counteracted their divergence. An evolutionary model stating that population differentiation increases over time and might ultimately lead to allopatric speciation is less parsimonious. Postulating that the populations were initially less differentiated than they are today would imply the assumption that a homogeneous but already diversified founder population colonized all regions. This in turn requires the effective dispersal of large populations across whole Europe, and would thus hardly allow for subsequent allopatric population differentiation.

Conclusions

Our results suggest that gene flow among *P. paneuropaeus* populations spanning a range of more than 3,000 km

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

effectively counteracts allopatric divergence. This contrasts population differentiation that has been observed in prokaryotes inhabiting geothermal hot springs. Although distances between potential habitats of *P. paneuropaeus* are certainly much smaller and more continuously distributed (stepping stone model) than those of thermophilic organisms (island model), the studied lakes and ponds are hardly connected by running waters, suggesting a high dispersal potential through airway. The species furthermore constitutes at least in some regions exceptionally high local abundances, large population sizes, and in general high recombination rates, which presumably all promote gene flow. We mainly investigated longitudinal (north–south) gene flow of a boreal bacterium between the region of its metapopulation center and southern satellite populations. It will be exciting to see if latitudinal gene flow across the boreal zone circling the northern hemisphere and especially intercontinental gene flow is as high as observed here in longitudinal direction.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) (project number 27160-B22). We thank Johanna Schmidt and Ulrike Koll for isolation of strains, DNA isolation, and processing of some samples for *priB* amplicon sequencing.

Data Availability

The data underlying this article are available in the NCBI database at https://www.ncbi.nlm.nih.gov (last accessed February 15, 2021), and can be accessed with the following accession numbers.

Genomes

CP030085–CP030088; CP049617–CP049647; CP049679–CP049684; JAANET000000000–JAANHG000000000; JAANHT000000000–JAANHW000000000; QMCG00000000; QMCH00000000.

BioProjects

PRJNA278737 and PRJNA295639.

BioSamples

SAMN02724733; SAMN03430691; SAMN03430798; SAMN04080026; SAMN04086652; SAMN04086667–SAMN04086669; SAMN06014615; SAMN07200920; SAMN08383909; SAMN08383917–SAMN08383921; SAMN14212605–SAMN14212701.

Literature Cited

Anderson RE, Kouris A, Seward CH, Campbell KM, Whitaker RJ.2017. Structured populations of *Sulfolobus acidocaldarius* with susceptibility to mobile genetic elements. Genome Biol Evol.9(6):1699–1710.

Antony-Babu S, et al. 2017.Multiple *Streptomyces* species with distinct secondary metabolomes have identical 16S rRNA gene sequences. Sci Rep. 7(1):8.

Arevalo P, VanInsberghe D, Elsherbini J, Gore J, Polz MF.2019. A reverse ecology approach based on a biological definition of microbial populations. Cell178(4):820–834.e14.

Baas BeckingLGM. 1934. Geobiologie of inleiding tot de milieukunde.Den Haag(Netherlands): W.P. Van Stockum & Zoon.

Bankevich A, et al. 2012.SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 19(5):455–477.

Bastian M, Heymann S, Jacomy M. 2009. Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media. 2. [Database]

Bobay L-M, Ellis BS-H, Ochman H. 2018.ConSpeciFix: classifying prokaryotic species based on gene flow. Bioinformatics34(21):3738–3740.

Bobay L-M, Ochman H. 2017.Biological species are universal across life's domains. Genome Biol Evol.9(3):491–501.

Bolyen E, et al. 2019.Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol. 37(8):852–857.

Borcard D, Legendre P. 2012. Is the Mantel correlogram powerful enough to be useful in ecological analysis? A simulation study. Ecology93(6):1473–1481.

Chase AB, et al. 2019. Maintenance of sympatric and allopatric populations in free-living terrestrial bacteria. mBio 10(5):e02361–19.

Cohan FM. 2001.Bacterial species and speciation.Syst Biol. 50(4):513–524.

Cui Y, et al. 2015.Epidemic clones, oceanic gene pools, and eco-LD in the free living marine pathogen *Vibrio parahaemolyticus*. Mol Biol Evol.32(6):1396–1410.

Darling AE, Mau B, Perna NT. 2010.progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One5(6):e11147.

DeLeon-Rodriguez N, et al. 2013.Microbiome of the upper troposphere: species composition and prevalence, effects of tropical storms, and atmospheric implications. Proc Natl Acad Sci USA. 110(7):2575–2580.

Didelot X, Wilson DJ.2015. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 11(2):e1004041.

Doolittle WF, Zhaxybayeva O. 2009.On the origin of prokaryotic species.Genome Res. 19(5):744–756.

Duchêne S, et al. 2016.Genome-scale rates of evolutionary change in bacteria. Microbial Genomics.2(11):e000094.

Excoffier L, Lischer HEL.2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. Mol Ecol Resourc.10(3):564–567.

Falush D, et al. 2006.Mismatch induced speciation in *Salmonella*: model and data. Philos Trans R Soc B.361(1475):2045–2053.

Fraser C, Hanage WP, Spratt BG. 2007. Recombination and the nature of bacterial speciation.Science315(5811):476–480.

Hahn MW.2003. Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones. Appl Environ Microbiol. 69(9):5248–5254.

Hahn MW, et al. 2012.The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-living *Polynucleobacter* population. PLoS One7(3):e32772.

Hahn MW, et al. 2017.*Polynucleobacter wuianus* sp. nov., a free-living freshwater bacterium affiliated with the cryptic species complex PnecC. Int J SystEvol Microbiol. 67(2):379–385.

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021

Hahn MW, Jezberová J, Koll U, Saueressig-Beck T, Schmidt J. 2016.Complete ecological isolation and cryptic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. ISME J. 10(7):1642–1655.

Hahn MW, Koll U, Jezberová J, Camacho A. 2015.Global phylogeography of pelagic *Polynucleobacter* bacteria: restricted geographic distribution of subgroups, isolation by distance and influence of climate. Environ Microbiol. 17(3):829–840.

Hahn MW, Pöckl M. 2005.Ecotypes of planktonic Actinobacteria with identical 16S rRNA genes adapted to thermal niches in temperate, subtropical, and tropical freshwater habitats. Appl Environ Microbiol. 71(2):766–773.

Hahn MW, Schmidt J, Pitt A, Taipale SJ, Lang E. 2016.Reclassification of four *Polynucleobacter necessarius* strains as representatives of *Polynucleobacter asymbioticus* comb. nov., *Polynucleobacter duraquae* sp. nov., *Polynucleobacter yangtzensis* sp. nov. and*Polynucleobacter sinensis* sp. nov., and emended description of *Polynucleobacter necessarius*. Int J Syst Evol Microbiol. 66(8):2883–2892.

Hahn MW, Stadler P, Wu QL, Pöckl M. 2004.The filtration–acclimatization method for isolation of an important fraction of the not readily cultivable bacteria.J Microbiol Methods.57(3):379–390.

Hellweger FL, Sebille E, van, Fredrick ND.2014. Biogeographic patterns in ocean microbes emerge in a neutral agent-based model. Science345(6202):1346–1349.

Hoetzinger M, et al. 2019.*Polynucleobacter paneuropaeus* sp. nov., characterized by six strains isolated from freshwater lakes located along a 3000 km north–south cross-section across Europe. Int J Syst Evol Microbiol. 69(1):203–213.

Hoetzinger M, Hahn MW.2017. Genomic divergence and cohesion in a species of pelagic freshwater bacteria. BMC Genomics18(1):794.

Hoetzinger M, Schmidt J, Jezberová J, Koll U, Hahn MW.2017. Microdiversification of a pelagic *Polynucleobacter* species is mainly driven by acquisition of genomic islands from a partially interspecific gene pool. Appl Environ Microbiol. 83: e02266–e02316.

Jezbera J, et al. 2012.Contrasting trends in distribution of four major planktonic betaproteobacterial groups along a pH gradient of epilimnia of 72 freshwater habitats.FEMS Microbiol Ecol.81(2):467–479.

Katoh K, Standley DM. 2013.MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol.30(4):772–780.

Koeppel AF, et al. 2013.Speedy speciation in a bacterial microcosm: new species can arise as frequently as adaptations within a species. ISME J. 7(6):1080–1091.

Konstantinidis KT, Ramette A, Tiedje JM.2006. The bacterial species definition in the genomic era. Philos Trans R Soc B.361(1475):1929–1940.

Kuhlemann J, et al. 2005.Würmian maximum glaciation in Corsica.Austrian J Earth Sci.97:68–81.

Kuhner MK, Felsenstein J. 1994.A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates.Mol Biol Evol.11:459–468.

Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol.35(6):1547–1549.

Markowitz VM, et al. 2012.IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 40(D1):D115–D122.

Mayr E. 1942. Systematics and the origin of species from the viewpoint of a zoologist.New York:Columbia University Press.

Müller AL, et al. 2014.Endospores of thermophilic bacteria as tracers of microbial dispersal by ocean currents.ISME J. 8(6):1153–1165.

Oden NL, Sokal RR.1986. Directional autocorrelation: an extension of spatial correlograms to two dimensions. Syst Zool. 35(4):608–617.

Oksanen J, et al. 2020. vegan: community ecology package. R package version 2.5-7. Available from: https://CRAN.R-project.org/package=vegan. Accessed February 15, 2021.

Page AJ, et al. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics31(22):3691–3693.

Page AJ, et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics.2(4):e000056.

Papke RT, Ramsing NB, Bateson MM, Ward DM. 2003. Geographical isolation in hot spring cyanobacteria.Environ Microbiol. 5(8):650–659.

Paradis E, Schliep K. 2019.ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics35(3):526–528.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 25(7):1043–1055.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data.Genetics155(2):945–959.

R Core Team.2019. R: a language and environment for statistical computing. Vienna (Austria):R Foundation for Statistical Computing.

Ramasamy RK, Ramasamy S, Bindroo BB, Naik VG.2014. STRUCTURE PLOT: a program for drawing elegant STRUCTURE bar plots in user friendly interface. SpringerPlus3(1):431.

Richter M, Rosselló-Móra R. 2009.Shifting the genomic gold standard for the prokaryotic species definition.Proc Natl Acad Sci U S A.106(45):19126–19131.

Schliep KP.2011. phangorn: phylogenetic analysis in R. Bioinformatics27(4):592–593.

Shapiro BJ, Polz MF.2015. Microbial speciation. Cold Spring Harb Perspect Biol. 7(10):a018143.

Sikorski J, Nevo E. 2005.Adaptation and incipient sympatric speciation of *Bacillus simplex* under microclimatic contrast at "Evolution Canyons" I and II, Israel.Proc Natl Acad Sci U S A. 102(44):15924–15929.

Stamatakis A. 2014.RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics30(9):1312–1313.

Stroeven AP, et al. 2016.Deglaciation of Fennoscandia.Quaternary Sci Rev.147:91–121.

VanInsberghe D, et al. 2015.Non-symbiotic *Bradyrhizobium* ecotypes dominate North American forest soils. ISME J. 9(11):2435–2441.

Whitaker RJ.2006. Allopatric origins of microbial species. Philos Trans R Soc B.361(1475):1975–1984.

Whitaker RJ, Grogan DW, Taylor JW.2003. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science301(5635):976–978.

Whitaker RJ, Grogan DW, Taylor JW.2005. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. Mol Biol Evol.22(12):2354–2361.

Whittaker KA,Rynearson TA.2017. Evidence for environmental and ecological selection in a microbe with no geographic limits to gene flow. Proc Natl Acad Sci U S A. 114(10):2651–2656.

Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y.2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol.8(1):28–36.

Zwart G, et al. 1998.Nearly identical 16S rRNA sequences recovered from lakes in North America and Europe indicate the existence of clades of globally distributed freshwater bacteria. Syst Appl Microbiol. 21(4):546–556.

Associate editor: Howard Ochman

Downloaded from https://academic.oup.com/gbe/article/13/3/evab019/6126423 by guest on 05 May 2021