Hoetzinger and Hahn *BMC Genomics* (2017) 18:794
DOI 10.1186/s12864-017-4199-z

BMC Genomics

RESEARCH ARTICLE Open Access

Genomic divergence and cohesion in a species of pelagic freshwater bacteria

Matthias Hoetzinger* and Martin W. Hahn

Abstract

Background: In many prokaryotic genera a clustered phylogeny is observed, akin to the occurrence of species in sexually reproducing organisms. For some taxa, homologous recombination has been invoked as the underlying mechanism providing genomic cohesion among conspecific individuals. Whether this mechanism is applicable to prokaryotes in freshwaters with low habitat connectivity – i.e. elevated geographic barriers to gene flow – is unclear. To investigate further we studied genomic trends within the globally abundant PnecC cluster (genus *Polynucleobacter*, *Betaproteobacteria*) and analyzed homologous recombination within the affiliated species *P. asymbioticus*.

Results: Comparisons among 20 PnecC genomes revealed a clearly discontinuous distribution of nucleotide sequence similarities. Among the nine conspecific individuals (*P. asymbioticus*) all average nucleotide identity (ANI) values were greater than 97%, whereas all other comparisons exhibited ANI values lower than 85%. The reconstruction of recombination and mutation events for the *P. asymbioticus* core genomes yielded an r/m ratio of 7.4, which is clearly above estimated thresholds for recombination to act as a cohesive force. Hotspots of recombination were found to be located in the flanking regions of genomic islands. Even between geographically separated habitats a high flux of recombination was evident. While a biogeographic population structure was suggested from MLST data targeting rather conserved loci, such a structure was barely visible when whole genome data was considered. However, both MLST and whole genome data showed evidence of differentiation between two lineages of *P. asymbioticus*. The ratios of non-synonymous to synonymous substitution rates as well as growth rates in transplantation experiments suggested that this divergence was not selectively neutral.

Conclusions: The high extent of homologous recombination among *P. asymbioticus* bacteria can act as a cohesive force that effectively counteracts genetic divergence. At least on a regional scale, homologous recombination can act across geographically separated ecosystems and therefore plays an important role in the evolution and consistency of bacterial freshwater species. A species model akin to the biological species concept may be applicable for *P. asymbioticus*. Nonetheless, two genetically distinct lineages have emerged and further research may clarify if their divergence has been initiated by reinforced geographical barriers or has been evolving in sympatry.

Keywords: *Polynucleobacter*, Freshwater bacteria, Homologous recombination, Gene flow, Population structure, Genomic cohesion

* Correspondence: matthias.hoetzinger@uibk.ac.at
Research Institute for Limnology, University of Innsbruck, Mondseestrasse 9, A-5310 Mondsee, Austria

Background

Diversity within many prokaryotic genera is not continuous but clustered. This is evidenced by sequencing of genomes [1–3], metagenomic DNA [4–7] and single genes [8–10]. Considering average nucleotide identity (ANI), there is a noticeable sparsity of values between 82 and 96% [11]. For many taxa this "ANI-gap" has been proven useful for species delineation on a genomic basis; that is, the upper bound of the gap coincides with suggested thresholds of 95–96% ANI [11–13]. It can be assumed that there exists an underlying biological process that is maintaining genomic coherence. Two important mechanisms have been proposed that can eventually promote genomic similarity among individuals: periodic selection [14, 15] and homologous recombination (HR) [16, 17]. In the periodic selection model the evolving diversity within populations is periodically purged by selective sweeps [18, 19]. In selective sweeps, possibly certain loci only but not entire genomes hitchhike to high frequencies along with selectively beneficial mutations, which may be the consequence of recombination unlinking different regions of the genome [7, 20, 21]. However, even without the immediate action of selection, HR can reduce intralineage divergence [17, 22, 23]. This mechanism of cohesion is consulted in the biological species concept [24]. It is still a matter of debate if or rather to what extent this concept is applicable to microbes [25–28]. Generally, recombination can provide cohesion for a species only if it spreads genetic variability faster than it is accumulated by mutation. Theoretical models based on neutral Fisher-Wright populations, which consider an exponential decrease of HR rate with sequence divergence according to experimental data [29–31], suggest a threshold for recombination rate relative to mutation (r/m), above which population divergence is prevented [17, 22, 23]. This threshold is passed at r/m ratios between 0.25 and 2 according to [17], i.e. populations with lower ratios are free to diverge clonally, whereas above the threshold divergence is constrained by the cohesive effects of HR.

Prokaryotic HR has been widely studied in pathogenic bacteria, where it is often linked to antibiotic resistance [32, 33]. Data from multilocus sequence typing (MLST) have proved helpful in assessing the degree of HR in various populations of bacteria [34, 35]. The estimated extent of recombination varies drastically between different prokaryotic taxa [36–38]. The estimated r/m ratio of 63.1 for marine SAR11 isolates is one of the highest recorded for bacteria [36]. Various studies invoke HR within lineages as a cohesive force, which may result in speciation of ecologically differentiated lineages [20, 39–41]. In contrast, there are examples of monomorphic pathogens for which highly clonal population structures are suggested [42–45]. Exceptionally low recombination frequencies have also been reported for the freshwater clade of SAR11, estimated to be more than two orders of magnitudes lower than in the marine sister group [36, 46, 47]. The low recombination rates detected for freshwater SAR11 are supposedly the result of purged diversity following the transition from marine to freshwater systems [46]. Beyond that, the geographic separation of inland waters may have an important effect on HR rates and speciation. While many freshwater habitats are connected by stretches of running water, others can be largely isolated from other water bodies [48]. In the latter case, prokaryotic gene flow across distant habitats is dependent on dispersal by air or animals. The availability of stepping stones can increase the connectivity across distant habitats [49]. If stepping stones are missing, populations may evolve in isolation, as has been shown for thermoacidophilic archaea [50, 51]. However, biogeographical isolation has been observed even in microbial species with less restrictive habitat requirements and may enforce allopatric speciation [52, 53]. The role of HR in giving coherence to prokaryotic freshwater species is widely unclear. Here we studied this issue in *Polynucleobacter* bacteria affiliated with the PnecC cluster.

PnecC bacteria are abundant in the pelagic zone of diverse freshwater habitats [54–57] and show cosmopolitan distribution [58]. PnecC represents a cryptic species complex, including a few described and an unknown number of undescribed species [59–61]. *Polynucleobacter asymbioticus*, formerly termed the F10 lineage [48, 56], has so far been detected only in central Europe. The species is particularly abundant in dystrophic ponds in the Austrian Alps, where it accounted for up to 46% of total bacterioplankton and was shown to maintain persistent populations [48, 62]. Isolates, which cannot be obtained by standard cultivation methods [63], are available from only three sites (Loibersbacher Höhe (Loi), Rauriser Urwald (Rau), Trög (Tro)) within a maximum distance of 76 km separating the three sites [64]. At each site there are several dystrophic ponds not directly connected to other water bodies by running water. While numerous ecologically similar ponds are present throughout the Alps, such ponds are rare in the surrounding lowlands. The Loi site, at the foothills of the Alps, therefore represents the periphery of the *P. asymbioticus* range. The habitat range of *P. asymbioticus* and other PnecC species/lineages has been shown to be restricted, as strains isolated from acidic water were unable to grow in alkaline water and vice versa [59]. This may strengthen geographic barriers for PnecC bacteria, which calls into question whether species coherence can be maintained among subpopulations from distinct sites. For the flexible gene pool of *P. asymbioticus* it has been shown that gene flow across geographically separated sites is realized. This gene flow is mainly attributed to

illegitimate recombination of genomic islands, which can even be exchanged across species boundaries and may often be mediated by phages [64]. The exchange of genomic islands is a major driver for diversification regarding the flexible genome of the species. In contrast, the extent of HR in the core genome potentially mediating species coherence is unclear and was investigated in this study.

Results

Genome clustering in PnecC

Twenty strains affiliated with the PnecC cluster, nine of which representing *P. asymbioticus*, have been whole genome sequenced in previous studies [60, 64–70] (Additional file 1). These genomes were analyzed for the distribution of pairwise ANI values (Fig. 1, Additional file 2). A phylogenetic tree based on partial 16S rRNA gene sequences of the used and related strains is shown in Additional file 3. No ANI values in the range of 85% - 97% are found. The values above 97% ANI represent intraspecies comparisons corresponding to the nine *P. asymbioticus* genomes and two genomes affiliated with a *Polynucleobacter* species which has not yet been described. A separation into two lineages, simplex and amplus [64], is visible for *P. asymbioticus*. Each other genome analyzed here represents a putatively distinct species, several of which have been described [59–61]. It is evident that different species within the cryptic species complex PnecC are genomically well differentiated. Potential mechanisms responsible for the coherence within the species *P. asymbioticus* are discussed in the following.

HR and population differentiation inferred from MLST and genomic data

The 37 *P. asymbioticus* strains used in this study have been isolated over eleven years [64]. Removal of identical MLST sequence types obtained from single samplings resulted in a remaining data set of 19 strains. Subpopulations clustered according to the three different sites of origin are all significantly differentiated in MLST data (Fig. 2). The fixation index (F_{ST}) is highest between Loi and Rau (0.74), medium between Tro and Rau (0.56), and lowest between Loi and Tro (0.28). The r/m ratio estimated by ClonalFrame [71] is 1.34 (Table 1), i.e. the estimated number of polymorphisms arisen from HR is higher than the one from mutation. If only strains affiliated with the simplex lineage are included in the analysis, the r/m ratio increases to 4.08.

The r/m values estimated for nine strains from whole genome data are 7.4, i.e. higher than those estimated from MLST data (Table 1). In order to see whether the high HR rates inferred would be reflected in incongruent phylogenies of different genome segments, the core genome alignment was exemplarily split into nine sections of equal length. The nine respective phylogenetic trees were all incongruent to each other (Fig. 3), which was assessed with the Shimodaira-Hasegawa test [72]. The test rejects the null hypothesis for all combinations of trees and alignment data with *p*-values lower than 0.001,

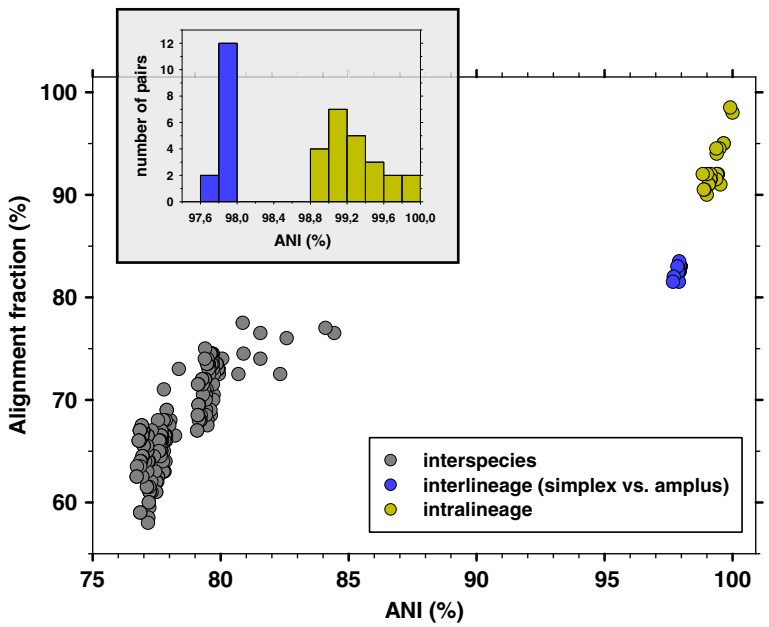

Fig. 1 Average nucleotide identities in PnecC. Alignment fraction plotted against ANI for pairwise comparisons of 20 PnecC genomes. The insert at the top left shows the number of genome pairs at 0.2% ANI intervals for the intraspecific comparisons

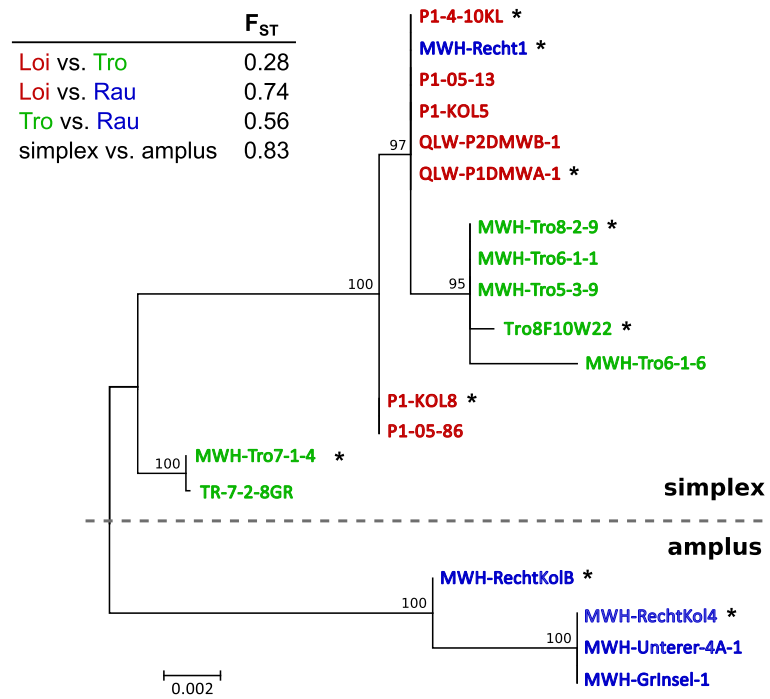

Fig. 2 Unrooted maximum likelihood tree from MLST data of *P. asymbioticus* isolates. From each sampling only one representative of each sequence type was included, to avoid the possible effects of transient blooming of certain genotypes. The F_{ST} data refers to the same set of strains shown in the tree. Genome sequenced strains are indicated by stars

which certainly confirms incongruence among all trees [73]. Differences in log-likelihood scores for different tree topologies with respect to certain sequence data are given in Table 2. Figure 3 shows that the remote position of the two amplus genomes is rather conserved in all nine trees. This suggests that HR between the lineages may be less frequent than within the simplex lineage. This is also evident from the higher r/m ratio inferred from the MLST data when strains of the amplus lineage are excluded from the analysis. In any case, the r/m and μ/ρ ratios, respectively, estimated from MLST as well as genomic data imply that HR can act as a cohesive force within *P. asymbioticus* according to theoretical models [17, 23]. The strains used in the analysis were isolated from distant sites. Consequently it is suggested that the cohesive effect of HR can act across distant sites, counteracting divergence of the species at least on this geographic scale (76 km max. Distance). Nonetheless, population differentiation between sites is evident to some extent in genomic data (Fig. 3) and to a greater extent in the MLST data (Fig. 2). The simplex and amplus lineages are well differentiated with respect to all sequencing data.

Lack of mismatch repair system

Mismatch repair systems have been shown to prevent HR in various taxa [74–77]. All 20 PnecC genomes were analyzed for mismatch repair systems, and strikingly, *mutS* and *mutL* are missing in all PnecC genomes. In order to check whether this would be a rather unique characteristic of PnecC, or if related taxa would share the absence of *mutS* and *mutL*, the IMG/ER public database was screened for genomes of the genera most closely related to *Polynucleobacter*, i.e. *Cupriavidus* and

Table 1 Results of the ClonalFrame analysis

Data	STs / genomes	# Sites	δ	r / m	π	μ / ρ
MLST	7 simplex, 2 amplus	7259	662 [249–1636]	1.3 [0.5–2.7]	0.0126	0.0094 [0.0047–0.0245]
MLST	7 simplex	7259	609 [230–1498]	4.1 [1.5–9.1]	0.0073	0.0018 [0.0008–0.0048]
genomes	7 simplex, 2 amplus	1,814,913	319[a] [305–333]	7.4 [6.9–7.8]	0.0111	0.0015 [0.0016–0.0014]

STs sequence types, *δ* average tract length of recombination events, *r/m* ratio of probabilities that a given site has been altered through recombination and mutation, *π* average nucleotide diversity per site, *μ/ρ*, ratio of mutation to recombination calculated as $\pi\ (r/m)^{-1}$ according to Fig. 4 in [23]. 95% confidence intervals are given in brackets

[a]As a result of the ClonalOrigin analysis a median δ of 2340 was obtained

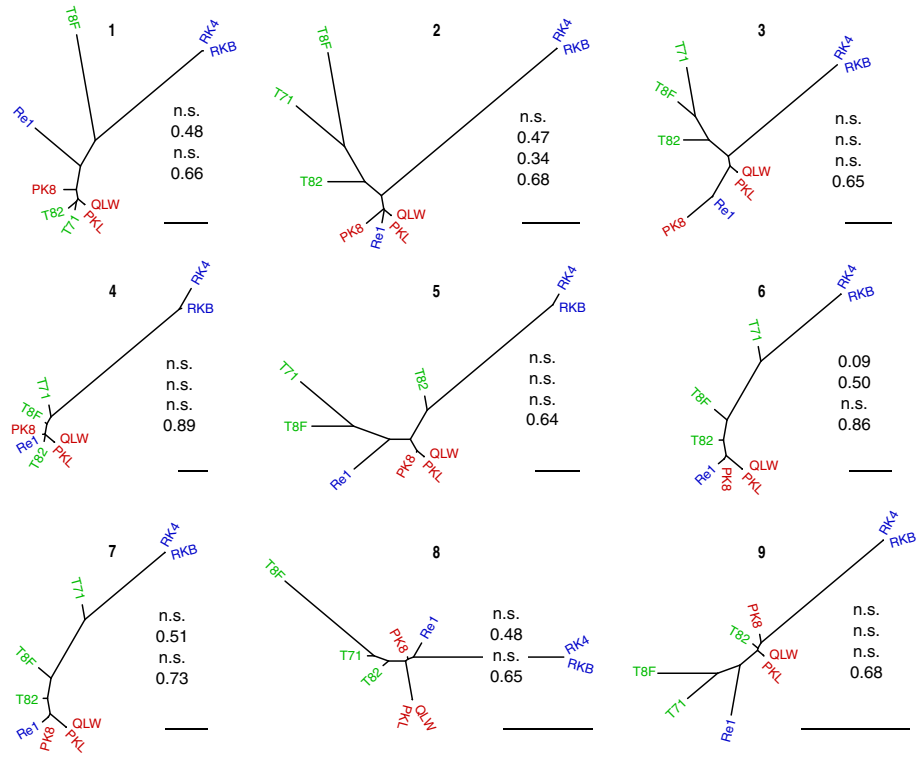

Fig. 3 Unrooted RAxML trees for different segments of the core genome alignment. The scale bars indicate nucleotide divergence of 0.005. F_{ST} values in the same order as given in Fig. 2 are shown for each tree. n.s., not significant ($p > 0.1$); QLW, QLW-P1DMWA-1; PKL, P1–4-10KL; PK8, P1-Kol8; Re1, MWH-Recht1; RKB, MWH-RechtKolB; RK4, MWH-RechtKol4; T71, MWH-Tro7–1-4; T82, MWH-Tro8–2-9; T8F, Tro8F10W22

Ralstonia [78]. After removal of single-cell genomes, 29 *Cupriavidus* and 65 *Ralstonia* genomes were obtained. Both *mutS* and *mutL* were found in all 65 *Ralstonia* and in 28 of the 29 *Cupriavidus* genomes. As well, *mutS* and *mutL* were found in *Polynucleobacter* genomes not affiliated with PnecC, i.e. *P. rarus* [79] and *P. cosmopolitanus* [80], but both genes are missing in *P. acidiphobus* [81] (unpublished data).

Flanking regions of replacement genomic islands as hotspots of HR

The estimated frequency of HR events, inferred from the core genome alignment of the nine *P. asymbioticus* strains was plotted against the genome position, exemplarily on QLW-P1DMWA-1 (Fig. 4). As the synteny in the core genome of *P. asymbioticus* is well conserved, the figure looks similar if any other of the nine strains is

Table 2 Shimodaira-Hasegawa test for incongruence among tree topologies for different segments of the core genome alignment

Segment	1	2	3	4	5	6	7	8	9
1	–	10,378	17,700	26,494	11,041	12,955	15,541	8745	12,441
2	7100	–	1908	11,511	4635	4823	5961	5295	8089
3	13,765	5472	–	13,087	10,540	15,234	10,690	10,714	10,508
4	14,493	11,159	7432	–	6411	9415	8898	12,331	13,047
5	13,057	8894	12,004	16,317	–	14,431	10,825	12,648	11,782
6	8979	7055	8418	5340	8223	–	2023	9517	8523
7	15,661	10,081	10,433	6309	11,983	4014	–	13,848	15,287
8	3819	3428	4378	8328	5980	3416	3095	–	3069
9	6474	4660	6312	9181	3951	4865	4985	2889	–

For the sequence data of each segment (rows), the differences in log-likelihood scores between the original tree topology and each other tree topology (columns) are given.
P-values for the differences are all lower than 0.001

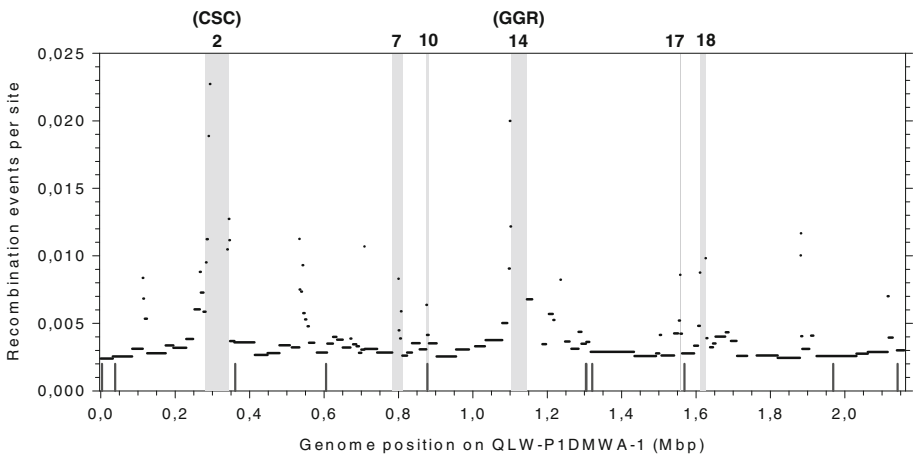

Fig. 4 Frequency of recombination with respect to genome position. The y-axis indicates the frequency of recombination events as inferred by ClonalOrigin for each block in the core genome alignment of the nine *P. asymbioticus* strains. The x-axis shows the respective genome position, exemplarily for strain QLW-P1DMWA-1. Genomic islands which have been assigned to the replacement type in [64], i.e. presumably resulted from HR between the flanking sequences of the islands, are highlighted by grey areas and numbered according to their genomic position defined earlier (Fig. 3C in [64]). CSC (cell surface composition) and GGR (giant gene region), designated as in [64], are typical examples of replacement genomic islands (cf. [83]). The position of the loci used for MLST is indicated by short lines at the bottom of the graph

used as reference for genome position. Merely, the x-axis shifts due to differing numbers and sizes of genomic islands among the strains (cf. [64]). Several regions with clearly elevated HR frequencies are apparent. Some of these regions correspond to the flanking regions of six genomic islands which were presumed to be the result of HR earlier [64]. The numbering of these genomic islands in Fig. 4 corresponds to their genomic position assigned in [64]. In particular, the borders of two genomic islands are clearly represented as HR hotspots here. It has been suggested in [64] that the island designated as CSC (cell surface composition) may be involved in the synthesis of exposed structures of the cell surface. Similar genomic islands are found in various other prokaryotes [82]. The other island, designated as GGR (giant gene region), has been shown to contain a giant gene with sequence lengths up to 42 kbp. CSC and GGR are typical examples of replacement genomic islands according to the definition in [83]. In replacement genomic islands relatively big DNA fragments are assumed to be transferred by single recombination events, where homology is required only for the ends of these fragments [82]. It is worth noting that, although the hotspots show clearly elevated per site HR rates, their contribution to the total number of inferred HR events is minor, as they are allocated to relatively short blocks/genomic regions only (Additional file 4).

The loci regarding the MLST data are also indicated in Fig. 4. It is evident that these loci mostly correspond to regions with relatively low recombination frequencies, except for the locus at about 0.88 Mbp genome position, which represents a core gene located within a genomic island.

HR among different branches of the genealogy

The numbers of observed HR events among all branches of the genealogy relative to the expected numbers under the prior model were inferred by ClonalOrigin [84]. The results are shown in Fig. 5 and the underlying numbers are given an Additional file 5. It is conspicuous that the flux of HR from near ancestors of Loi genomes towards the branches leading to the Tro genomes was clearly enhanced in comparison to the expectations. On the other hand, the flux in the reverse direction was approximately as expected. Another distinctive feature is that recombinational export from the branch leading to the amplus genomes was found less frequently than under the prior model, whereas observed import to that branch was slightly higher than expected. The terminal branches of the two amplus genomes exhibited reduced import from the ancestral branch of the simplex genomes.

Purifying selection in *P. asymbioticus*

K_a/K_s ratios regarding the 1820 core genes of the nine *P. asymbioticus* genomes show that these genomes diverged predominantly under purifying selection (Fig. 6, Additional file 6). Only the two almost clonal strains QLW-P1DMWA-1 and P1–4-10KL exhibit a markedly higher K_a/K_s value of 6.6, which indicates positive selection. It should be noted that relatively high K_a/K_s values have frequently been reported for very closely related strains, which can often be explained by divergence times that have been too short for purifying selection to act on detrimental non-synonymous mutations [85, 86]. However, between the genomes of QLW-P1DMWA-1 and P1–4-10KL there are 19 polymorphisms in coding

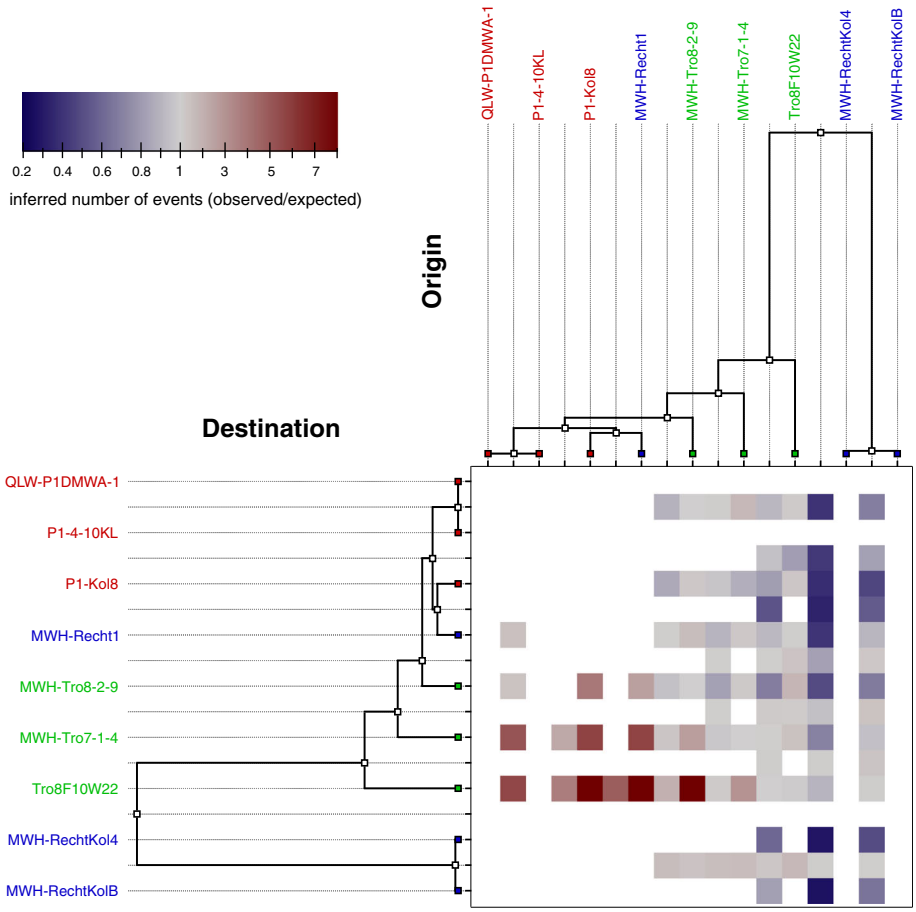

Fig. 5 Flux of recombination. Heat map showing the number of recombination events inferred by ClonalOrigin relative to its expectation under the prior model for each donor/recipient pair of branches. Cells for which both the number of observed and expected events are less than three are left white. The underlying numbers are given in Additional file 5

regions, only two of which are non-synonymous. This ratio is very unlikely to arise in the absence of selection. Disregarding this exception, the median K_a/K_s values for pairwise comparisons within the simplex lineage are significantly lower than the values between the simplex and amplus lineages. Generally, in the presence of purifying selection only, K_a/K_s values tend to decrease with divergence time [85, 86], which would rather favor an opposite trend. It is likely therefore that the divergent evolution of the two lineages has been accompanied by positive selection to some extent. No statistically significant difference in the K_a/K_s ratios is found if they are tested in view of local adaptation, i.e. when ratios within sites are compared to ratios between sites.

Transplantation experiment

The growth potential of four *P. asymbioticus* strains in their home and in two foreign environments was tested by reciprocal transplantation experiments. The pH of the water from Rechteckteich, Trog-7 and Trog-8 was 5.0, 5.6 and 4.7, respectively, in the first experiment (August 2014) and 5.3, 5.6 and 4.9, respectively, in the second experiment (September 2014). All strains were able to grow in all habitats at both times of the year. The bacterial numbers multiplied on average by a factor of 42 and 14 during the 48 h of the first and second experiment, respectively. Growth curves are shown in Additional file 7. Box plots of all determined growth rates for the exponential growth phase are shown in Fig. 7. No statistical differences were found in comparisons of growth rates between habitats or between dates. Comparisons of "home" and "away" likewise yielded no significant differences in growth rates and therefore no evidence for local adaptation. However, regarding the lineages, the evidence reveals that the strain affiliated with the amplus lineage exhibits significantly lower growth rates in the tested habitats compared to each of the three lineage simplex strains.

Discussion

Stepping stones

The K_a/K_s calculations and transplantation experiments did not show signs of local adaptation. All tested strains

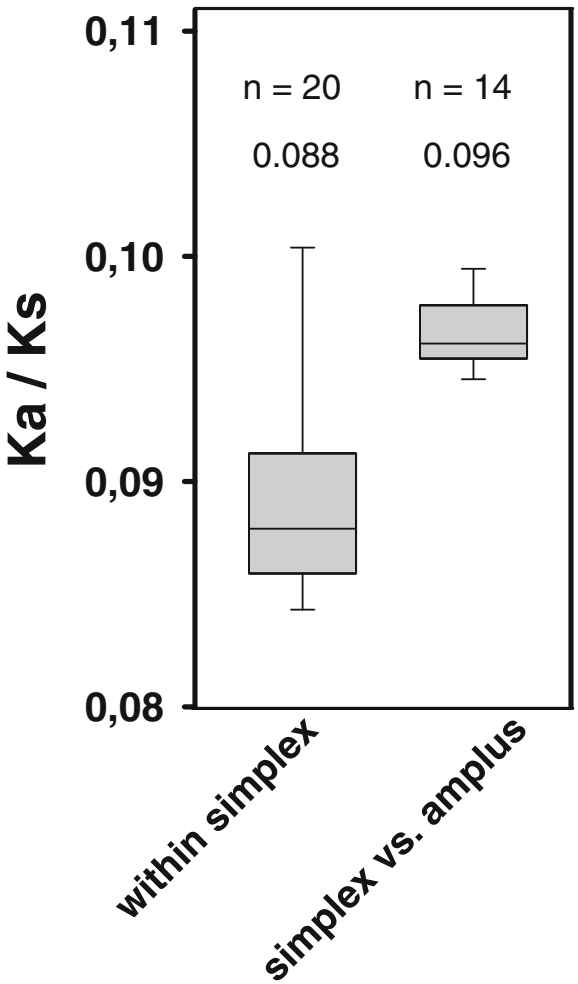

Fig. 6 Boxplot of K_a/K_s values. The middle lines indicate median values, the outer lines of the boxes represent the 25th and 75th, and the whiskers the 5th and 95th percentiles. Sample sizes (n) and median values are given above the plots. There is a statistically significant difference between the median values of the two groups shown according to the Mann-Whitney rank test ($p \leq 0.001$). The K_a/K_s between the two almost identical simplex genomes (QLW-P1DMWA-1 and P1–4-10KL) is 6.6 and was excluded from these analyses. The K_a/K_s between the two amplus genomes is 0.092 (Additional file 6)

were able to grow in water from different sites. Numerous habitats within the Alps that are ecologically similar to the tested ones could potentially serve as stepping stones for gene flow among local *P. asymbioticus* populations across distant sites. Moreover, even if dispersed bacteria go extinct in foreign habitats, their genes may still be introduced into the resident populations by recombination.

Sexual species and mismatch repair

The estimated r/m of 7.4 for the investigated *P. asymbioticus* core genome clearly exceeds estimated thresholds for the transition from clonal to sexual species [17, 23], which is passed at r/m ratios between 0.25 and 2 according to [17]. The respective strains were isolated from habitats within a maximum distance from each other of 76 km. Gene flow concerning the core genome of *P. asymbioticus* is obviously sufficient across these habitats to provide genomic coherence of the species. Dispersal limitation may be effectively reduced by interjacent habitats that potentially serve as stepping stones in the dispersal of these bacteria. In addition, the high abundance of *P. asymbioticus* in certain habitats [48, 54] may increase the chance for an effective dispersal. Furthermore it is conceivable that the success rate of single recombination events is enhanced within this species due to the lack of the *mutS*/*mutL* mismatch repair system [74–77]. Interestingly, according to genome data this mismatch repair system seems to be intact in other species of the *Polynucloebacter* genus not affiliated with the PnecC cluster and in the closest related genera, i.e. *Cupriavidus* and *Ralstonia*, as well. An extraordinarily high diversification has been reported for the PnecC cluster, coinciding with a disproportionally high overall genome diversity in relation to 16S rRNA gene diversity [59]. It can be speculated that radiation in the PnecC cluster has been promoted in consequence of loss of the *mutS*/*mutL* system. Possibly, HR acts as the cohesive force also within other species of the PnecC cluster and consequently is responsible for the clear genomic demarcation between *Polynucleobacter* species within PnecC. It should be noted that flexible genes can blur species boundaries, as genomic islands can be transferred between different species [64]. Hence, the biological species concept might be useful for PnecC bacteria but in a less strict manner than for sexually reproducing organisms.

HR and the evolution of prokaryotes in inland waters

Studies of HR in prokaryotes isolated from inland waters are rare. Examples are found for freshwater bacteria of the SAR11 clade, for which notably low recombination rates have been reported in contrast to the marine sister clade [46]. The freshwater SAR11 population may be affected by a putative bottleneck after the transition from marine to freshwater systems; that is, the low effective population size may be the reason for low rates of detectable recombination [46]. Similar reasons may underlie the observed clonality of many pathogens, which could be a result of their physically constrained population structure and population bottlenecks after shifts to new hosts rather than a mechanistic handicap of recombination (compare [87] and [39]).

For archaea affiliated with the genus *Sulfolobus* it has been shown that geographic gene flow barriers shape these populations. While pronounced HR has been evidenced in syntopic *Sulfolobus islandicus* subpopulations [40, 88], recombination across distant sites is effectively prevented [50, 51]. For these thermoacidophilic archaea,

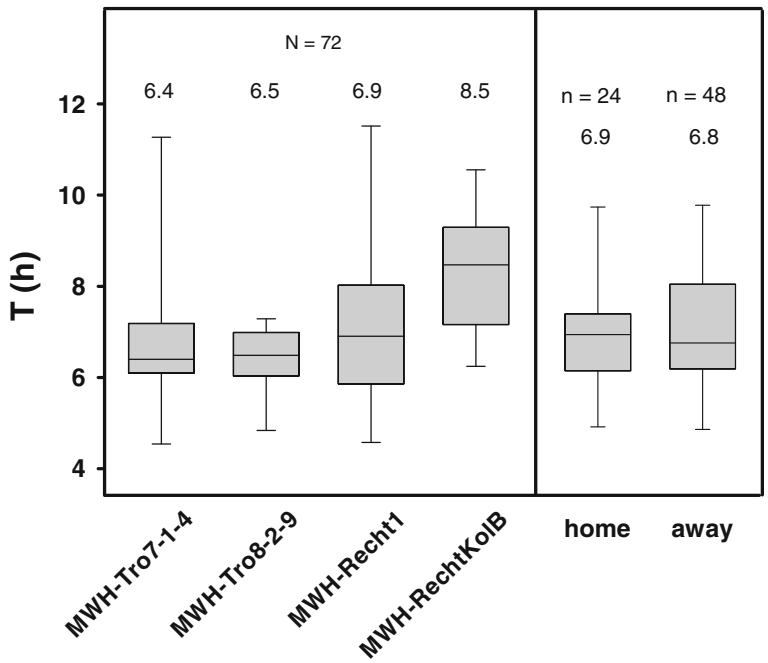

Fig. 7 Boxplot of doubling times from the transplantation experiments. The middle lines indicate median values, the outer lines of the boxes represent the 25th and 75th, and the whiskers the 5th and 95th percentiles. Sample sizes (n) and median values are given above the plots. The home group includes doubling times of strains grown in water collected from the same habitat that they were isolated from. The away group includes doubling times of strains grown in water collected from different habitats than those they were isolated from. There is a statistically significant difference between the median doubling times of strain MWH-RechtKolB and all other strains according to the Mann-Whitney rank test ($p < 0.05$). The differences between median doubling times of the other strains as well as between the groups home and away are not significantly different ($p > 0.05$). The underlying growth curves are shown in Additional file 7

regions providing potential habitats are generally far apart, and hence, stepping stones for dispersal may not be available [50, 51].

Pronounced HR among geographically separated *P. asymbioticus* subpopulations as suggested in this study contrasts the specific cases of freshwater SAR11 and *Sulfolobus islandicus*. Further investigations concerning recombination in other taxa would be necessary to clarify whether this corresponds to another extreme or rather to the prevalent situation in prokaryotes of inland waters.

Differences between MLST and genome-wide data

There is evidence that HR rates can vary greatly between different genome regions [41, 82, 84, 89], which suggests that rates determined by use of MLST data may often not represent genome-wide rates. Furthermore, as the detection of HR is usually limited to events that introduced a relatively high number of polymorphisms rather than events between very similar sequences, rates may often be underestimated (e.g. compare [90] and [91]). As shown in Fig. 4, the inferred HR frequency in *P. asymbioticus* varies considerably along the genome. In some regions positive selection pressure may tend to introduce variation. This is assumed for the genomic island related to cell surface composition (CSC in [64] and Fig. 4), where phage predation may impose selection pressure towards variation [64, 92]. Other genome regions are rather conserved, and in these regions purifying selection may act against HR. This may be particularly the case for genes which are required for basic cellular functions, which are typically sequenced for MLST [36]. Therefore, MLST data may generally be prone to underestimate genome-wide recombination rates, which is confirmed by this study. The lower HR rates are the probable cause of the higher F_{ST} values for MLST data in comparison to genome-wide data of this study. As total panmixis of the *P. asymbioticus* population across sites is precluded by dispersal limitations, genetic drift can cause differentiation among the geographically separated subpopulations. This differentiation is more likely in genomic regions which are less impacted by recombination or the MLST data, respectively. The signals of subpopulation differentiation may be transient however, and would possibly fade if strains isolated over a longer time period were analyzed.

Flux of HR

The ClonalOrigin analysis indicated an increased flux of HR from Loi strains towards Tro strains but not vice

versa. Loi is located at the foothills of the Austrian Alps, at the periphery of the *P. asymbioticus* range ([48, 57] and unpublished data), and the diversity found at this site was comparably low and stable (see [48, 62] and Fig. 2). It can be speculated that the resident *P. asymbioticus* population is less impacted by gene flow due to its peripheral location. In contrast, Tro is located more centrally, which could be the reason for a larger effective population size, higher diversity and a greater impact of recombination at this site.

Rau is the only site where isolates affiliated with the amplus lineage have been obtained. The lineage has never been detected at Loi and Tro, but it has been detected at a few other sites in the Austrian Alps; always, however, in small numbers compared to the simplex lineage (unpublished data). It is conceivable that the sampled habitats in the Alps only represent an overlap of two distinct taxon ranges of the two lineages. Thus, the amplus lineage may be more abundant in regions that have not yet been sampled, which could be different areas of the Alps or even a more distant mountain range. Recombinational export from the branch leading to the amplus genomes was found less frequently than expected. This may be explained by the low abundance of the amplus lineage compared to the simplex lineage in the sampled area. Moreover, the reduced import from the ancestral branch of the simplex genomes towards the terminal branches of the two amplus genomes could indicate a HR barrier between the two lineages. Such a barrier is also suggested by the distant phylogenetic position of the two amplus genomes from all simplex genomes in all trees in Fig. 3, and the higher r/m ratio inferred from MLST data when amplus strains are excluded from the analysis. A possible scenario could be that the two lineages inhabited different refugia during the last glacial period (ca. 110,000–12,000 years ago). In such a scenario, the lineages could now be in a process of convergence after sympatry may have been restored. Recombination has been suspected of causing convergence in other taxa, after recombination barriers between formerly separated lineages had disappeared [93, 94]. Alternatively, amplus and simplex strains may have diverged too much for subsequent cohesion by HR, and the divergence between the lineages may thus be ongoing. This would give rise to speciation between the two lineages. Incipient speciation is also conceivable in a scenario where the HR barrier between the two lineages was not a result of geographic but of ecological separation. Moreover, geographic and ecological separation may go hand in hand. Higher K_a/K_s ratios in comparisons between the lineages relative to comparisons within simplex point to a selective divergence, which may be expected in scenarios of both geographical and ecological separation.

Conclusions

Different *Polynucleobacter* species affiliated with the globally abundant PnecC cluster can be clearly delineated by whole genome comparisons. High HR rates in *P. asymbioticus* demonstrate that recombination can play an important role in the evolution of bacterial species inhabiting geographically separated freshwater habitats, at least on a regional scale. It is suggested that HR acts as the cohesive force that effectively counteracts genetic divergence between *P. asymbioticus* subpopulations. A species model resembling the biological species concept may therefore be applicable to these bacteria. Divergence between two lineages of *P. asymbioticus* coincides with decreased HR between these lineages. To learn more about potential recombination barriers and possible speciation between the two lineages, further isolation of respective strains from different regions would be necessary.

Methods

P. asymbioticus strains

The 37 *P. asymbioticus* strains were isolated over eleven years from nine habitats at three different sites (Loi, Rau and Tro) in the Austrian Alps, as has been described elsewhere [64]. Data on the strains is given in Additional file 1 and data on the sites of origin in Additional file 8. All 37 strains have been characterized earlier by multilocus sequence typing (MLST) of eleven loci [48]. The genomes of nine strains, three from each aforementioned site, have been sequenced, assembled and annotated in a previous study [64]. Further characterization of the strains and a detailed analysis on their flexible genome can be found in [64].

PnecC genomes analyses

Twenty strains affiliated with the PnecC cluster have been whole genome sequenced in previous studies [60, 64–70]. Data regarding these strains is given in Additional file 1. ANI values for all pairwise comparisons of the 20 PnecC genomes were calculated using the IMG/ER comparative analysis system [78] (Fig. 1 and Additional file 2). The IMG/ER system was also used to identify homologous genes among the genomes (Additional file 9).

Screening for *mutL* and *mutS*

Genomes affiliated with PnecC and its closest related taxa were screened for the presence/absence of the genes *mutL* and *mutS* using the IMG/ER analysis system [78]. The dataset included 22 *Polynucleobacter* genomes, i.e. 20 PnecC, one PnecA and one PnecD genome. Furthermore, 29 *Cupriavidus* and 65 *Ralstonia* genomes were obtained from the IMG/ER public database after excluding single-cell genomes.

Phylogenetics

For the unrooted tree in Fig. 2 the nucleotide sequences of the eleven MLST loci were concatenated (7259 bp) and aligned in MEGA7 [95] by Muscle [96]. A DNA substitution model analysis was conducted in MEGA7, and the model with the lowest BIC score was selected for Maximum Likelihood tree calculation, i.e. the Tamura 3-parameter model [97] and a discrete Gamma distribution (5 categories), to model evolutionary rate differences among sites. Bootstrap values were calculated from 100 bootstrap replications.

The trees in Fig. 3 are based on a multiple whole genome alignment of the nine *P. asymbioticus* genomes performed by progressiveMauve [98, 99]. The alignment was refined with Gblocks v0.91b [100] using default parameters to omit poorly aligned positions. The resulting alignment of 1.81 Mbp was split into nine consecutive segments of equal length. For each segment an unrooted maximum-likelihood based RAxML [101] tree was inferred using the Cipres Science Gateway V.3.3 [102]. The GTRCAT model with 25 distinct rate categories was used.

Incongruences among the trees in Fig. 3 were determined using the Shimodaira-Hasegawa test [72]. The test was executed for each of the nine segments. In each test the log-likelihood for each of the nine tree topologies was calculated given the sequence alignment data of the segment. The tests were conducted in R [103] using the package phangorn 1.99–13 [104].

Population differentiation

Similarly as in [64], the strains were grouped according to geographic origin Loi, Rau and Tro and according to lineage (simplex and amplus), respectively. F_{ST} values were calculated based on pairwise differences (pi) in Arlequin 3.5.2.2 [105]. Significances of F_{ST} values were determined through comparisons to a permutation test, in which 10,000 permutations and $p < 0.1$ were considered significant.

Inference of HR

HR regarding MLST data was inferred using ClonalFrame [71] similarly as described in [36]. ClonalFrame estimates the clonal genealogy of a sample of strains based on a coalescent model [106]. By attempting to reconstruct the mutation and recombination events that took place on the branches of this genealogy, several evolutionary parameters are estimated, including the r/m ratio. The analysis was performed for two datasets, one including both simplex and amplus strains and the other including only simplex strains. Only one representative of each sequence type was included to avoid underestimating r/m ratios, as recombination between identical sequences cannot be detected. Thereby, also possible effects from transient blooming of certain genotypes are precluded. For both datasets, two runs of the ClonalFrame Monte Carlo Markov chain (MCMC) were performed, each consisting of 2,000,000 iterations. The first half of the chains was discarded and the second half was sampled every hundred iterations. Convergence of the MCMC for the two runs of each dataset was confirmed by the Gelman-Rubin convergence test [107]. For each dataset only the results of the run with the higher average log-likelihood are given in Table 1. The μ/ρ ratio as it is used in [23] was calculated as $\pi\ (r/m)^{-1}$. The nucleotide diversity π for each dataset was calculated using DnaSP 5.10.1 [108].

For inference of HR in the core genome of *P. asymbioticus*, the process as described in [84] was followed. At first, the nine genomes were aligned by progressiveMauve [98, 99]. The stripSubsetLCBs script was used to leave only alignment blocks longer than 500 bp. A total of 111 such blocks was obtained, ranging in size from 607 bp to 116,266 bp with a concatenated length of 1,814,913 bp (Additional file 4). These 1.8 Mbp represent the core genome of the nine *P. asymbioticus* strains. Three runs of the ClonalFrame MCMC were performed on this core genome alignment, each consisting of 1,010,000 iterations. The first 10,000 iterations were discarded and the subsequent 1,000,000 iterations were sampled every ten iterations. Again, convergence of the runs was confirmed by the Gelman-Rubin convergence test [107]. By comparing the output trees it was verified that the three runs produced a consistent genealogy. Based on this genealogy, further properties of HR were inferred using ClonalOrigin [84], which performs approximate inference under the coalescent model with gene conversion [109]. ClonalOrigin infers the origin and destination of recombination events on the genealogy, as well as the three parameters δ (average length of recombination events), θ_s (scaled mutation rate) and ρ_s (scaled recombination rate). The ClonalOrigin MCMC was run independently on each of the 111 blocks of the core genome alignment specified above. The values of δ, θ_s and ρ_s for each block were calculated by the computeMedians.pl script [84] (Additional files 4 and 10). The same script was used to calculate the weighted median values of these parameters across all blocks, at which the following values were obtained: δ = 2340 bp, θ_s = 0.0109, and ρ_s = 0.0056. ClonalOrigin was then rerun for each block with the three parameters set equal to these estimates. Both ClonalOrigin runs were performed with 2,000,000 iterations of the MCMC chain, the first half of which discarded and the second half sampled every 10,000 iterations. The data used in Fig. 4 and Fig. 5 was extracted from the output of the second run using the ClonalOrigin GUI.

K_a/K_s calculations

The core genes of the nine *P. asymbioticus* genomes were identified as follows. A blast was performed using

the IMG/ER comparative analysis system [78], including all protein coding genes of one genome as reference (QLW-P1DMWA-1 was chosen arbitrarily) and each other genome as query. Best blast hits for each genome were retained, applying a threshold >70% amino acid sequence identity. For query genes present multiple times in the obtained dataset, only the hit with the highest bit score was kept. The resulting dataset comprised 1820 reference genes with hits in all query genomes, which were defined as core genes (Additional file 9). Stop codons were removed from the core genes of all nine strains. The gene sequences of each strain were concatenated such that homologous genes were in the same order for all strains. From these concatenated nucleotide sequences the respective amino acid sequences were generated using EMBOSS Transeq [110]. The amino acid sequences of all pairwise combinations of strains were aligned with MAFFT version 7 [111]. The pal2nal.pl script [112] was used to generate the codon alignments from the amino acid alignments and the respective nucleotide sequences. Axt files were generated using the parseFastaIntoAXT.pl script [113] and the K_a/K_s ratios were calculated from these files for pairwise comparisons of strains using the K_a/K_s calculator 2.0 [114] and the γ-MYN method [115]. The strains QLW-P1DMWA-1 and P1–4-10KL differ by only 15 SNPs within the core genes and were therefore excluded from the analysis described in the following. The pairwise K_a/K_s ratios within the simplex strains as well as the ratios between simplex and amplus strains were grouped together and tested for normal distribution by the Shapiro-Wilk test. The test failed for both groups on a significance level $p = 0.05$. The Mann-Whitney rank sum test was performed to test for significant differences of K_a/K_s ratios between the two groups.

Transplantation experiments

Reciprocal transplantation of strains was conducted to study growth performance of strains in sterile water of their home habitats ("home") and in sterile water of the home habitats of the other strains included ("away"). The experiments were performed with four pure culture strains, two of which were isolated from Rechteckteich in Rau (MWH-Recht1 and MWH-RechtKolB), one from Trog-7 in Tro (MWH-Tro7–1-4) and one from Trog-8, also in Tro (MWH-Tro8–2-9). Two experiments were conducted, the first in August 2014 and the second in September/October 2014. Water from Rechteckteich, Trog-7 and Trog-8 [64] was collected on 19th August and 25th September 2014 and sterilized by 0.1 μm filtration. The experiment was conducted in 50 ml cultures in agitated Erlenmeyer flasks at 15 °C in the dark. The tested strains were pre-cultured for eight days in diluted (1 g l^{-1}) NSY medium [63], with an initial pH of 5.8. Subsequently, they were pre-cultured in the water in which the experiment was conducted for four days. The strains were inoculated into fresh medium/water several times during pre-cultivation. The experiment was initiated by transferring the pre-cultivated strains during the exponential growth phase into fresh water to yield a starting concentration of about 10^5 cells/ml. All four strains were tested in parallel in the three waters (triplicates each). Growth was monitored for 139 and 48 h, in the first and second experiment, respectively. Bacterial numbers were determined by flow cytometry. To ensure that growth rates were determined within the exponential growth phase, only the first 15 h of the experiment were consulted for growth rate calculations. For this purpose, three samples taken at approximately seven-hour intervals and four samples taken at approximately four-hour intervals were analyzed in the first and second experiment, respectively. The growth rate for each culture was estimated as the slope of the logarithm of cell concentration versus time, determined by linear regression.

All doubling times (Log(2) / growth rate) regarding a certain strain were grouped together and tested for normal distribution by the Shapiro-Wilk test. Furthermore the doubling times for "home" as well as those for "away" were grouped together and tested for normality. The test failed for the strains as well as for "home" and "away" on a significance level $p = 0.05$. The Mann-Whitney rank sum test was performed to test for significant differences of doubling times between the strains or between home and away, respectively.

Additional files

Additional file 1: List of all PnecC strains used in this study. (XLSX 17 kb)

Additional file 2: ANI values for all PnecC genome pairs. (XLSX 26 kb)

Additional file 3: 16S rRNA tree of *Polynucleobacter*. (PDF 61 kb)

Additional file 4: Core alignment blocks. The 111 blocks obtained from the alignment of the nine *P. asymbioticus* genomes are listed. Respective parameters obtained from the two ClonalOrigin runs are given. The last column gives the segment number of the split alignment (Fig. 3 and Table 2) in which the respective blocks are contained. (XLSX 24 kb)

Additional file 5: Inferred numbers of events between branches of the genealogy. The number of events as inferred from the second ClonalOrigin run and used in Fig. 5 are given. Observed numbers, expected numbers and their ratios are given in the first, second and third sheet, respectively. (XLSX 20 kb)

Additional file 6: Ka/Ks values for the alignment of the 1820 core genes for all *P. asymbioticus* genome pairs. (XLSX 16 kb)

Additional file 7: Growth curves of the two transplantation experiments performed with four *P. asymbioticus* strains from three different habitats. (PDF 62 kb)

Additional file 8: Characterization of the nine habitats from which the 37 *P. asymbioticus* strains have been isolated. (XLSX 11 kb)

Additional file 9: Homologous genes among the 20 PnecC genomes. The genome with the highest number of genes (MWH-RechtKol4) was used as a reference. For each of the 2362 protein coding genes of MWH-RechtKol4, the presence/absence of a homologous gene in each of the 19 other genomes was determined by applying an amino acid identity threshold of 70%. The number of reference genes for which a homolog is found in the respective query genome is given in the second row. It should be noted that these numbers can be slightly higher than the actual number of homologous genes in the query, as certain query genes may be counted twice when compared to paralogs of the reference. When considering each gene only once, 1820 *P. asymbioticus* core genes have been identified, which are highlighted in the first column. (XLSX 245 kb)

Additional file 10: Parameters inferred from the first ClonalOrigin run. The parameters δ, θ_s and ρ_s are plotted along the genome. The computed median values, which have been used in the second ClonalOrigin run, are indicated by red, dashed lines. (PDF 28 kb)

Abbreviations

ANI: Average nucleotide identity; bp: Base pairs; F_{ST}: Fixation index; HR: Homologous recombination; K_a/K_s: Ratio of the number of non-synonymous substitutions per non-synonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s); Loi: Loibersbacher Höhe site; MCMC: Monte Carlo Markov chain; MLST: Multilocus sequence typing; *P. asymbioticus*: *Polynucleobacter asymbioticus*; PnecC: *Polynucleobacter* subcluster C; r/m: Ratio of substitutions introduced by recombination relative to mutation; Rau: Rauriser Urwald site; Tro: Trög site

Acknowledgements

The computational results presented have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck. We are grateful to Michael Fink for considerate support regarding usage of the HPC cluster. We thank Xavier Didelot for answering various questions regarding the ClonalFrame and ClonalOrigin software.

Funding

This study was supported by the Austrian Science Fund (FWF) project I482-B09 (Ecological diversification in *Polynucleobacter*), the European Science Foundation (ESF) project FREDI, and the Austrian Academy of Sciences project 23791 (DOC fellowship).

Availability of data and materials

Genome sequences and MLST data are deposited in GenBank/EMBL under the accession numbers given in Additional file 1. Phylogenetic data are deposited in TreeBASE under the study ID 21594.

Authors' contributions

MH and MWH conceived and executed the study. MH wrote the manuscript. MWH edited and proofed the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Received: 24 January 2017 Accepted: 8 October 2017
Published online: 16 October 2017

References

1. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, Ding H, Marttinen P, Malmstrom RR, Stocker R, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344(6182):416–20.
2. Marcelletti S, Scortichini M. Definition of plant-pathogenic *Pseudomonas* Genomospecies of the *Pseudomonas syringae* complex through multiple comparative approaches. Phytopathology. 2014;104(12):1274–82.
3. Lopez-Perez M, Rodriguez-Valera F. Pangenome evolution in the marine bacterium *Alteromonas*. Genome Biol Evol. 2016;8(5):1556–70.
4. Konstantinidis KT, DeLong EF. Genomic patterns of recombination, clonal divergence and environment in marine microbial populations. ISME J. 2008; 2(10):1052–65.
5. Caro-Quintero A, Konstantinidis KT. Bacterial species may exist, metagenomics reveal. Environ Microbiol. 2012;14(2):347–55.
6. Brown MV, Lauro FM, DeMaere MZ, Muir L, Wilkins D, Thomas T, Riddle MJ, Fuhrman JA, Andrews-Pfannkoch C, Hoffman JM, et al. Global biogeography of SAR11 marine bacteria. Mol Syst Biol. 2012;8:595.
7. Bendall ML, Stevens SLR, Chan L-K, Malfatti S, Schwientek P, Tremblay J, Schackwitz W, Martin J, Pati A, Bushnell B, et al. Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. ISME J. 2016;10(7):1589–601.
8. Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. Fine-scale phylogenetic architecture of a complex bacterial community. Nature. 2004;430(6999):551–4.
9. Tanabe Y, Kasai F, Watanabe MM. Multilocus sequence typing (MLST) reveals high genetic diversity and clonal population structure of the toxic cyanobacterium *Microcystis aeruginosa*. Microbiology. 2007;153(Pt 11):3695–703.
10. Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science. 2008;320(5879):1081–5.
11. Kim M, Oh HS, Park SC, Chun J. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. Int J Syst Evol Microbiol. 2014;64(Pt 2): 346–51.
12. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci. 2009;106(45): 19126–31.
13. Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, Pati A. Microbial species delineation using whole genome sequences. Nucleic Acids Res. 2015;43(14):6761-71.
14. Atwood KC, Schneider LK, Ryan FJ. Periodic selection in *Escherichia coli*. Proc Natl Acad Sci U S A. 1951;37(3):146–55.
15. Cohan FM, Perry EB. A systematics for discovering the fundamental units of bacterial diversity. Curr Biol. 2007;17(10):R373–86.
16. Shen P, Huang HV. Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics. 1986;112(3): 441–57.
17. Fraser C, Hanage W, Spratt B. Recombination and the nature of bacterial speciation. Science. 2007;315(5811):476-80.
18. Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23(1):23–35.
19. Guttman DS, Dykhuizen DE. Detecting selective sweeps in naturally occurring *Escherichia coli*. Genetics. 1994;138(4):993–1003.
20. Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G, Polz MF, Alm EJ. Population genomics of early events in the ecological differentiation of bacteria. Science. 2012;336(6077):48–51.
21. Takeuchi N, Cordero OX, Koonin EV, Kaneko K. Gene-specific selective sweeps in bacteria and archaea caused by negative frequency-dependent selection. BMC Biol. 2015;13:20.
22. Hanage WP, Spratt BG, Turner KM, Fraser C. Modelling bacterial speciation. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361(1475):2039–44.
23. Doroghazi JR, Buckley DH. A model for the effect of homologous recombination on microbial diversification. Genome Biol Evol. 2011;3:1349–56.
24. Mayr E. Systematics and the origin of species, from the viewpoint of a zoologist. Cambridge: Harvard University Press; 1942.
25. Dykhuizen DE, Green L. Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol. 1991;173(22):7257–68.
26. Cohan FM. Bacterial species and speciation. Syst Biol. 2001;50(4):513–24.
27. Achtman M, Wagner M. Microbial diversity and the genetic nature of microbial species. Nat Rev Micro. 2008;6(6):431–40.
28. Doolittle WF, Zhaxybayeva O. On the origin of prokaryotic species. Genome Res. 2009;19(5):744–56.

29. Zawadzki P, Roberts MS, Cohan FM. The log-linear relationship between sexual isolation and sequence divergence in bacillus transformation is robust. Genetics. 1995;140(3):917–32.
30. Vulić M, Dionisio F, Taddei F, Radman M. Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci. 1997;94(18):9763–7.
31. Majewski J, Zawadzki P, Pickerill P, Cohan FM, Dowson CG. Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J Bacteriol. 2000;182(4):1016–23.
32. Spratt BG, Bowler LD, Zhang QY, Zhou J, Smith JM. Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal *Neisseria* species. J Mol Evol. 1992;34(2): 115–25.
33. Freel KC, Millán-Aguiñaga N, Jensen PR. Multilocus sequence typing reveals evidence of homologous recombination linked to antibiotic resistance in the genus *Salinispora*. Appl Environ Microbiol. 2013;79(19):5997–6005.
34. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE, et al. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc Natl Acad Sci U S A. 2001;98(1):182–7.
35. Maiden MC. Multilocus sequence typing of bacteria. Annu Rev Microbiol. 2006;60:561–88.
36. Vos M, Didelot X: A comparison of homologous recombination rates in bacteria and archaea. ISME J 2009, 3.
37. Fraser C, Hanage WP, Spratt BG: Neutral microepidemic evolution of bacterial pathogens. Proc Natl Acad Sci U S A 2005, 102.
38. Perez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol J Mol Epidemiol Evol Genet Infect Dis. 2006;6(2):97–112.
39. Didelot X, Bowden R, Street T, Golubchik T, Spencer C, McVean G, Sangal V, Anjum MF, Achtman M, Falush D *et al*: Recombination and population structure in *Salmonella enterica*. PLoS Genet 2011, 7.
40. Cadillo-Quiroz H, Didelot X, Held NL, Herrera A, Darling A, Reno ML, Krause DJ, Whitaker RJ: Patterns of gene flow define species of thermophilic Archaea. PLoS Biol 2012, 10.
41. Didelot X, Méric G, Falush D, Darling AE. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. BMC Genomics. 2012;13(1):256.
42. Leopold SR, Magrini V, Holt NJ, Shaikh N, Mardis ER, Cagno J, Ogura Y, Iguchi A, Hayashi T, Mellmann A, et al. A precise reconstruction of the emergence and constrained radiations of *Escherichia coli* O157 portrayed by backbone concatenomic analysis. Proc Natl Acad Sci U S A. 2009;106(21): 8713–8.
43. Monot M, Honore N, Garnier T, Zidane N, Sherafi D, Paniz-Mondolfi A, Matsuoka M, Taylor GM, Donoghue HD, Bouwman A, et al. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat Genet. 2009;41(12):1282–9.
44. Morelli G, Song Y, Mazzoni CJ, Eppinger M, Roumagnac P, Wagner DM, Feldkamp M, Kusecek B, Vogler AJ, Li Y, et al. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. Nat Genet. 2010;42(12):1140–3.
45. Achtman M. Insights from genomic comparisons of genetically monomorphic bacterial pathogens. Philos Trans R Soc Lond Ser B Biol Sci. 2012;367(1590):860–7.
46. Zaremba-Niedzwiedzka K, Viklund J, Zhao W, Ast J, Sczyrba A, Woyke T, McMahon K, Bertilsson S, Stepanauskas R, Andersson SGE. Single-cell genomics reveal low recombination frequencies in freshwater bacteria of the SAR11 clade. Genome Biol. 2013;14(11):R130.
47. Vergin KL, Tripp HJ, Wilhelm LJ, Denver DR, Rappe MS, Giovannoni SJ. High intraspecific recombination rate in a native population of *Candidatus* Pelagibacter ubique (SAR11). Environ Microbiol. 2007;9(10): 2430–40.
48. Hahn MW, Scheuerl T, Jezberova J, Koll U, Jezbera J, Simek K, Vannini C, Petroni G, Wu QL. The passive yet successful way of planktonic life: genomic and experimental analysis of the ecology of a free-living *Polynucleobacter* population. PLoS One. 2012;7(3):e32772.
49. van Gremberghe I, Leliaert F, Mergeay J, Vanormelingen P, Van der Gucht K, Debeer A-E, Lacerot G, De Meester L, Vyverman W. Lack of Phylogeographic structure in the freshwater Cyanobacterium *Microcystis aeruginosa* suggests global dispersal. PLoS One. 2011;6(5):e19561.
50. Whitaker RJ, Grogan DW, Taylor JW. Geographic barriers isolate endemic populations of hyperthermophilic archaea. Science. 2003;301(5635):976–8.
51. Reno ML, Held NL, Fields CJ, Burke PV, Whitaker RJ. Biogeography of the *Sulfolobus islandicus* pan-genome. Proc Natl Acad Sci U S A. 2009;106(21):8605–10.
52. Whitaker RJ. Allopatric origins of microbial species. Philos Trans R Soc Lond Ser B Biol Sci. 2006;361(1475):1975–84.
53. Ramette A, Tiedje JM. Biogeography: an emerging cornerstone for understanding prokaryotic diversity, ecology, and evolution. Microb Ecol. 2007;53(2):197–207.
54. Hahn MW. Isolation of strains belonging to the cosmopolitan *Polynucleobacter necessarius* cluster from freshwater habitats located in three climatic zones. Appl Environ Microbiol. 2003;69(9):5248–54.
55. Jezberova J, Jezbera J, Brandt U, Lindstrom ES, Langenheder S, Hahn MW. Ubiquity of *Polynucleobacter necessarius* ssp. *asymbioticus* in lentic freshwater habitats of a heterogeneous 2000 km area. Environ Microbiol. 2010;12(3):658–69.
56. Jezbera J, Jezberova J, Brandt U, Hahn MW. Ubiquity of *Polynucleobacter necessarius* subspecies *asymbioticus* results from ecological diversification. Environ Microbiol. 2011;13(4):922–31.
57. Jezbera J, Jezberova J, Koll U, Hornak K, Simek K, Hahn MW. Contrasting trends in distribution of four major planktonic betaproteobacterial groups along a pH gradient of epilimnia of 72 freshwater habitats. FEMS Microbiol Ecol. 2012;81(2):467–79.
58. Hahn MW, Koll U, Jezberova J, Camacho A. Global phylogeography of pelagic *Polynucleobacter* bacteria: restricted geographic distribution of subgroups, isolation by distance and influence of climate. Environ Microbiol. 2015;17(3):829–40.
59. Hahn MW, Jezberova J, Koll U, Saueressig-Beck T, Schmidt J. Complete ecological isolation and cryptic diversity in *Polynucleobacter* bacteria not resolved by 16S rRNA gene sequences. ISME J. 2016;10(7):1642–55.
60. Hahn MW, Schmidt J, Pitt A, Taipale SJ, Lang E. 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. 2016;66(8):2883–92.
61. Hahn MW, Huymann LR, Koll U, Schmidt J, Lang E, Hoetzinger M: *Polynucleobacter wuianus* sp. nov., a free-living freshwater bacterium affiliated with the cryptic species complex PnecC. Int J Syst Evol Microbiol 2016.
62. Hahn MW, Pockl M, Wu QL. Low intraspecific diversity in a *Polynucleobacter* subcluster population numerically dominating bacterioplankton of a freshwater pond. Appl Environ Microbiol. 2005;71(8):4539–47.
63. Hahn MW, Stadler P, Wu QL, Pockl M. The filtration-acclimatization method for isolation of an important fraction of the not readily cultivable bacteria. J Microbiol Methods. 2004;57(3):379–90.
64. Hoetzinger M, Schmidt J, Jezberova J, Koll U, Hahn MW: Microdiversification of a pelagic *Polynucleobacter* species is mainly driven by acquisition of genomic islands from a partially interspecific gene pool. Appl Environ Microbiol 2017, 83(3).
65. Meincke L, Copeland A, Lapidus A, Lucas S, Berry KW, Del Rio TG, Hammon N, Dalin E, Tice H, Pitluck S, et al. Complete genome sequence of *Polynucleobacter necessarius* subsp. *asymbioticus* type strain (QLW-P1DMWA-1(T)). Stand Genomic Sci. 2012;6(1):74–83.
66. Hahn MW, Huymann LR, Koll U, Schmidt J, Lang E, Hoetzinger M. *Polynucleobacter wuianus* sp. nov., a free-living freshwater bacterium affiliated with the cryptic species complex PnecC. Int J Syst Evol Microbiol. 2017;67(2):379–85.
67. Hahn MW, Karbon G, Koll U, Schmidt J, Lang E. *Polynucleobacter sphagniphilus* sp. nov. a planktonic freshwater bacterium isolated from an acidic and humic freshwater habitat. Int J Syst Evol Microbiol. 2017;67(9): 3261-67.
68. Vannini C, Pockl M, Petroni G, Wu QL, Lang E, Stackebrandt E, Schrallhammer M, Richardson PM, Hahn MW. Endosymbiosis in statu nascendi: close phylogenetic relationship between obligately endosymbiotic and obligately free-living *Polynucleobacter* strains (*Betaproteobacteria*). Environ Microbiol. 2007;9(2):347–59.
69. Hao Z, Li L, Liu J, Ren Y, Wang L, Bartlam M, Egli T, Wang Y. Genome sequence of a freshwater low-nucleic-acid-content bacterium, betaproteobacterium strain CB. Genome Announcements. 2013;1(2):e00135–13.
70. Garcia SL, McMahon KD, Grossart H-P, Warnecke F. Successful enrichment of the ubiquitous freshwater acI Actinobacteria. Environ Microbiol Rep. 2014; 6(1):21–7.

71. Didelot X, Falush D: Inference of bacterial microevolution using multilocus sequence data. Genetics 2007, 175.
72. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to Phylogenetic inference. Mol Biol Evol. 1999;16(8):1114.
73. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002;51(3):492–508.
74. Worth L, Clark S, Radman M, Modrich P. Mismatch repair proteins MutS and MutL inhibit RecA-catalyzed strand transfer between diverged DNAs. Proc Natl Acad Sci. 1994;91(8):3238–41.
75. Evans E, Alani E. Roles for mismatch repair factors in regulating genetic recombination. Mol Cell Biol. 2000;20(21):7839–44.
76. Li G-M. Mechanisms and functions of DNA mismatch repair. Cell Res. 2008; 18(1):85–98.
77. Tham K-C, Hermans N, Winterwerp Herrie HK, Cox Michael M, Wyman C, Kanaar R, Lebbink Joyce HG. Mismatch repair inhibits Homeologous recombination via coordinated directional unwinding of trapped DNA structures. Mol Cell. 2013;51(3):326–37.
78. Markowitz VM, Chen IM, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res. 2012; 40(Database issue):D115–22.
79. Hahn MW, Lang E, Tarao M, Brandt U. *Polynucleobacter rarus* sp. nov., a free-living planktonic bacterium isolated from an acidic lake. Int J Syst Evol Microbiol. 2011;61(Pt 4):781–7.
80. Hahn MW, Lang E, Brandt U, Lunsdorf H, Wu QL, Stackebrandt E. *Polynucleobacter cosmopolitanus* sp. nov., free-living planktonic bacteria inhabiting freshwater lakes and rivers. Int J Syst Evol Microbiol. 2010;60(Pt 1):166–73.
81. Hahn MW, Lang E, Brandt U, Spröer C. *Polynucleobacter acidiphobus* sp. nov., a representative of an abundant group of planktonic freshwater bacteria. Int J Syst Evol Microbiol. 2011;61(0 4):788–94.
82. Lopez-Perez M, Martin-Cuadrado AB, Rodriguez-Valera F. Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes. Front Genet. 2014;5:147.
83. Lopez-Perez M, Gonzaga A, Rodriguez-Valera F. Genomic diversity of "deep ecotype" *Alteromonas macleodii* isolates: evidence for pan-Mediterranean clonal frames. Genome Biol Evol. 2013;5(6):1220–32.
84. Didelot X, Lawson D, Darling A, Falush D: Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 2010, 186.
85. Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH, Feil EJ. Comparisons of dN/dS are time dependent for closely related bacterial genomes. J Theor Biol. 2006;239(2):226–35.
86. Mugal CF, Wolf JB, Kaj I. Why time matters: codon evolution and the temporal dynamics of dN/dS. Mol Biol Evol. 2014;31(1):212–31.
87. Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I, Rance R, Baker S, Maskell D, Wain J *et al*: High-throughput sequencing provides insights into genome variation and evolution in *Salmonella Typhi*. Nat Gen 2008, 40.
88. Whitaker RJ, Grogan DW, Taylor JW. Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*. Mol Biol Evol. 2005;22(12):2354–61.
89. Yahara K, Kawai M, Furuta Y, Takahashi N, Handa N, Tsuru T, Oshima K, Yoshida M, Azuma T, Hattori M, et al. Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. Genome Biol Evol. 2012;4(5):628–40.
90. Feil EJ, Cooper JE, Grundmann H, Robinson DA, Enright MC, Berendt T, Peacock SJ, Smith JM, Murphy M, Spratt BG, et al. How clonal is *Staphylococcus aureus*? J Bacteriol. 2003;185(11):3307–16.
91. Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H. Population genomics in bacteria: a case study of *Staphylococcus aureus*. Mol Biol Evol. 2012;29(2):797–809.
92. Rodriguez-Valera F, Martin-Cuadrado AB, Rodriguez-Brito B, Pasic L, Thingstad TF, Rohwer F, Mira A. Explaining microbial population genomics through phage predation. Nat Rev Microbiol. 2009;7(11):828–36.
93. Didelot X, Achtman M, Parkhill J, Thomson N, Falush D: A bimodal pattern of relatedness between the *Salmonella* Paratyphi A and Typhi genomes: Convergence or divergence by homologous recombination? Genome Res 2007, 17.
94. Sheppard S, McCarthy N, Falush D, Maiden M. Convergence of *Campylobacter* species: implications for bacterial evolution. Science 2008, 320.
95. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33(7):1870–4.
96. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
97. Tamura K. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol Biol Evol. 1992;9(4):678–87.
98. Darling AE, Mau B, Perna NT. ProgressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010;5(6):e11147.
99. Darling AC, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 2004; 14(7):1394–403.
100. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
101. Stamatakis A, et al. Bioinformatics. 2006;22(21):2688–90.
102. Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science gateway for inference of large phylogenetic trees. In: Proceedings of the gateway computing environments workshop (GCE). New Orleans, LA: IEEE. 2010;1–8.
103. R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2013.
104. Schliep KP. Phangorn: phylogenetic analysis in R. Bioinformatics. 2011;27(4): 592–3.
105. Excoffier L, Lischer HE. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and windows. Mol Ecol Resour. 2010;10(3):564–7.
106. Kingman JFC. The coalescent. Stoch Process Appl. 1982;13(3):235–48.
107. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Statist Sci. 1992;7(4):457-72.
108. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 2009;25(11):1451–2.
109. Wiuf C, Hein J. The coalescent with gene conversion. Genetics. 2000;155(1): 451–62.
110. Rice P, Longden I, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16(6):276–7.
111. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80.
112. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 2006;34(Web Server issue):W609–12.
113. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4(4):259–63.
114. Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics. 2010;8(1):77–80.
115. Wang D-P, Wan H-L, Zhang S, Yu J. γ-MYN: a new algorithm for estimating ka and Ks with consideration of variable substitution rates. Biol Direct. 2009;4:20.