REPINs are facultative genomic symbionts of bacterial genomes

Frederic Bertels[1] and Paul B. Rainey[1,2]

[1]Department of Microbial Population Biology, Max Planck Institute for Evolutionary Biology, Plön 24306, Germany

[2]Laboratory of Biophysics and Evolution, CBI, ESPCI Paris, Université PSL, CNRS, Paris, France

Corresponding author: Frederic Bertels, August-Thienemann-Straße 2, 24306 Plön, Germany, bertels@evolbio.mpg.de.

Running title: REPINs are genomic symbionts

Keywords: Mobile genetic elements – Symbiosis – Interactions – Levels of selection

Abstract

Relationships among organisms, in which one lives inside of another, with benefits accruing to both partners, are referred to as endosymbiotic. Such relationships are common in the biological world, yet descriptions are confined to organismal life. Here we argue that short sequence repeats known as REPINs – whose replication is dependent on a non-jumping RAYT transposase – form a similar facultative symbiotic relationship with the bacterial chromosome. Evidence stems from distribution patterns across the eubacteria: persistence times of many millions of years, exceedingly rare duplication rates, vertical transmission, and population biology typical of living organisms, including population size fluctuations that correlate with available genome space. Additional analysis of patterns of REPIN evolution conform with theoretical predictions of conflict (and resolution) arising from the effects of selection operating simultaneously on REPINs and host cells. A search for similar kinds of genomic symbionts suggests that the REPIN-RAYT system is not unique.

Introduction

The merging of once independently replicating entities is central to the emergence of life's complexity. For example, chromosomes arose from interactions among independently replicating genes. The eukaryotic cell evolved from once independently replicating archaebacteria- and eubacteria-like microbes. Multicellular organisms evolved from single-celled ancestral types, and eusocial insect colonies from solitary types. Events of this kind are given special significance and are marked as "major evolutionary transitions" [1].

Similar kinds of hierarchically structured life forms, in which two independently replicating entities become nested within a single "higher level" autonomously replicating type, are common. For example, many organisms form intimate associations (endosymbioses) with microbes [2]. In some instances these are obligate with the microbial component being confined to specific cell types and transmitted strictly vertically [3]. In other instances, interactions are facultative with the microbial partner capable of independent existence and horizontal movement between hosts [3-6].

While endosymbiosis involving two or more organisms is a widespread and well-studied feature of biological life, there is no reason to suppose that such endosymbiotic associations are confined to cellular organisms. Indeed, early steps during life's evolution may have involved independent chemical replicators that formed networks of self-replicating chemical reactions [7]. Self-replicating DNA molecules, for instance selfish genetic elements such as a transposons or insertion sequences, are nested within chromosomes [8].

Examples of the latter are commonplace, yet the fate of selfish genetic elements is extinction, at least in bacterial genomes [9-14]. In order to persist long-term, selfish genetic elements must constantly infect new hosts. In this regard there is no permanent alignment, or replicative nesting, of selfish genetic elements within chromosomes.

Here we wish to outline the case for the existence of an interaction among a particular class of self-replicating genetic element and the chromosome of eubacteria that has all the

hallmarks of an enduring mutualism. The element is a bipartite system involving a **R**EP-**a**ssociated **t**yrosine **t**ransposase (RAYT) and a family of short, palindromic, non-autonomous elements termed REPINs (**REP** doublets forming hair**pin**s). RAYTs are incapable of mediating their own replication, but facilitate the replication and persistence of REPINs [15-17]. While previous work has argued that RAYTs have been co-opted for some cellular function [16,18,19], recognition that they, along with their cognate REPINs might constitute semi-autonomous, self-replicating entities, nested within bacterial genomes, has thus far not been considered.

We begin with a brief description of key features and attributes of the REPIN-RAYT system, including likely origin, distribution, persistence, duplication rates and mode of transmission. We then report analyses that show that REPINs form populations typical of living organisms, including evidence of population size fluctuations that correlate with available resources (genome space). These observations lead to the suggestion that REPINs are analogous to facultative, self-replicating, endosymbiotic bacteria found within certain eukaryotes. Consistent with such a prediction is the expectation of conflict between resident REPIN populations and host cells, evidence of which we present from an analysis of *Pseudomonas* and *Neisseria* genomes. Together our data indicate that REPINs are genomic symbionts that define a new class of replicative agent.

Results and Discussion

The REPIN-RAYT system

REPINs are short (~100 bp long), repetitive (typically 100s of copies per genome), extragenic, palindromic sequences (**Figure 1**) found in approximately 10% of eubacteria, but have not been detected in archaea. REPINs consist of two inverted repeats (REP sequences [20]) each defined by a short (~14bp) palindrome (green arrows in **Figure 1**) separated by a spacer region [15]. The spacer region is highly variable, while the flanking REP sequences are conserved.

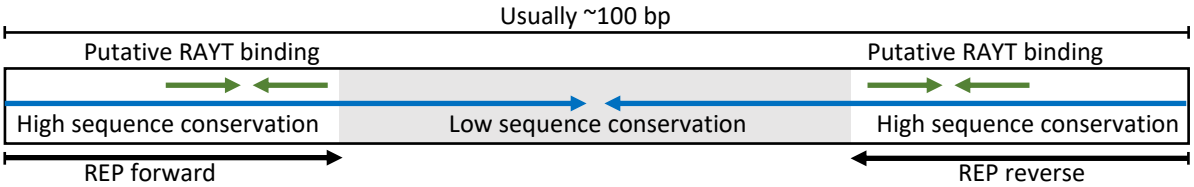


Figure 1. Schematic representation of a REPIN. A typical REPIN consists of two REP sequences in inverted orientation, separated by a short and less conserved DNA sequence (grey highlighting). Blue and green arrows indicate inverted repeats. REPIN length can differ but is usually in the range of about 100bp. REPINs reside almost exclusively in extragenic space.

REPINs are duplicated by a single copy transposase termed RAYT (**R**EP **A**ssociated t**Y**rosine **T**ransposase) [15,17,19]. Although RAYTs show sequence similarity to transposases, there is no evidence that RAYTs mediate their own transposition. Within bacterial species, RAYTs are typically found in the same extragenic space, are absent from plasmids and other horizontally transferred elements and their phylogeny is congruent with that of the host cell. RAYTs are thus vertically inherited and rarely move between bacterial hosts [16]. This is in marked contrast with selfish genetic elements, such as transposons and insertion sequences, whose persistence depends on horizontal transfer to new hosts [16,21].

RAYTs originated at least 150 million years ago (based on likely presence in the common ancestor of enterobacteria [16]), most likely from an IS*200*-like precursor. Extant RAYTs and IS*200* sequences share identical catalytic domains (HUH and a catalytic tyrosine) and both are associated with short palindromic sequences. Whereas IS*200* sequences form a single cohesive cluster, RAYTs form five sequence groups. All RAYTs are found as single copy

genes in bacterial chromosomes, but only Group 2 and Group 3 RAYTs are associated with REPINs [16].

REPINs and RAYTs evolved from an IS*200*-like ancestor

The association between both transposases (IS*200* and RAYTs) and short palindromic sequences suggests a common evolutionary origin. IS*200* transposases are flanked by short palindromic repeats that are essential for transposition function [22]. While extant IS*200* elements retain a classical transposition-based life history, the lineage that gave rise to RAYTs took a markedly different route: here, palindromic sequences flanking an IS element appear to have fused to form a REPIN capable of exploiting the transposase function for duplication [18]. The typical evolutionary fate of such non-autonomous elements is extinction, marked first by loss of the full-length transposase genes, followed by degradation of the non-autonomous elements (**Figure 2**, [23]). Curiously, RAYT transposases have not been lost. But inability to replicate independently of the chromosome, means that maintenance of RAYTs depends on beneficial fitness contributions to the host bacterium. In actual fact, the benefit must arise from a combination of both RAYT and REPIN activity. If this were not so, RAYTs would lose ability to duplicate REPINs, and in turn, REPINs would go extinct.

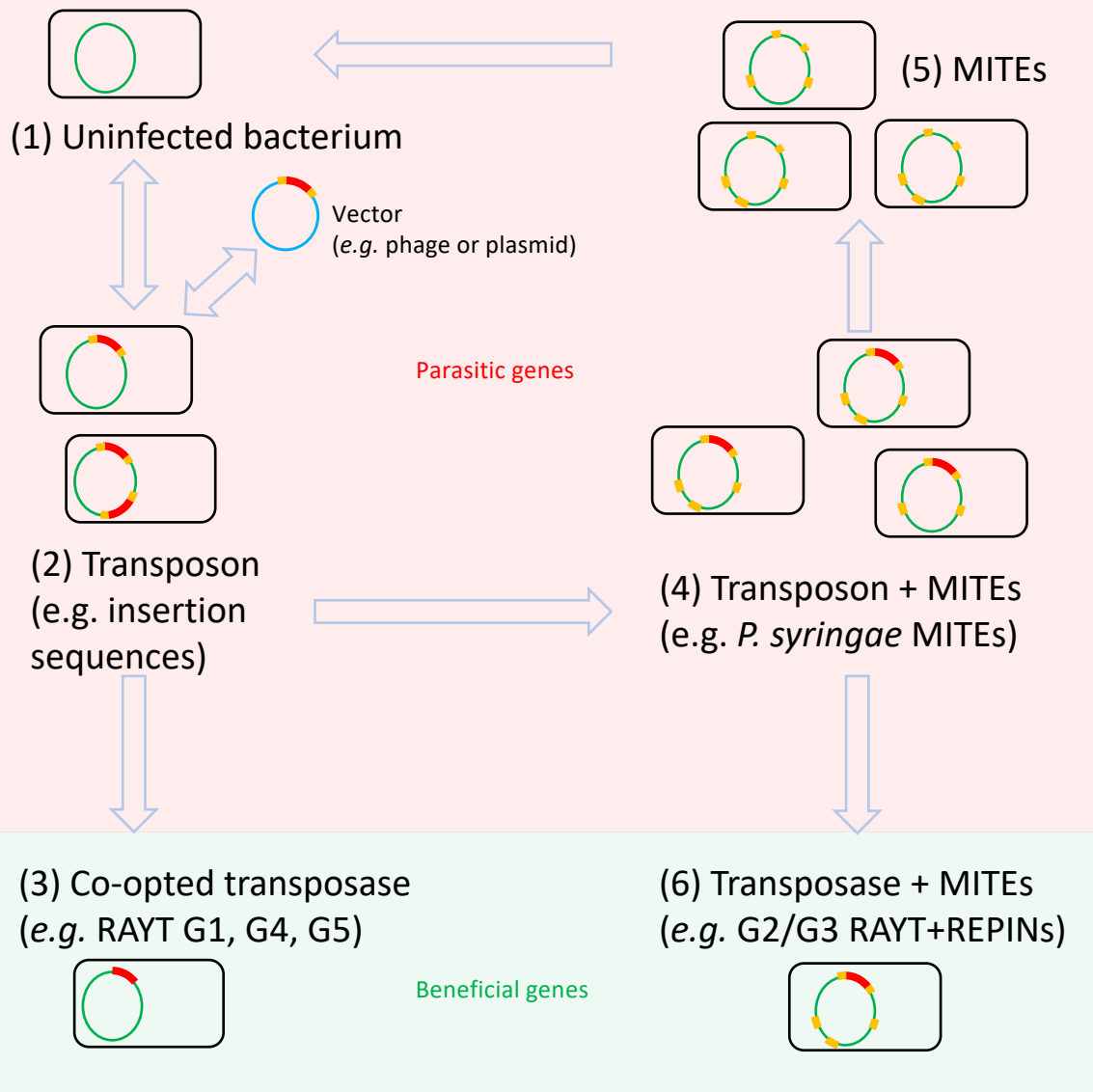


Figure 2. Stages of REPIN-RAYT evolution. (1) An uninfected bacterium becomes infected by a vector, such as a plasmid or phage (blue circle), which contains a transposon (e.g., an insertion sequence) with flanking sequences (orange) and transposase gene (red). (2) The vector infects the bacterium and the transposon inserts into the genome. Once in the genome it may duplicate, go extinct (1), or (3), on rare occasions the transposase may be co-opted by the host to perform some host-beneficial function. The latter results in the transposase becoming a single copy gene, and unable to duplicate as observed for Group 1, 4 and 5 RAYTs [16]. (4) Insertion sequences may also duplicate inside the bacterial genome and leave multiple copies. Sometimes the transposase gene in one of the copies is deactivated. Yet, flanking genes persist and duplicate by parasitizing transposase function encoded elsewhere in the genome. Flanking sequences lacking transposase function are called miniature inverted repeat transposable elements (MITEs) [24]. (5) MITEs can parasitize full-length transposons and may even cause loss of transposon movement. Once the last functional transposase is eradicated from the genome MITEs face extinct [9,23]. (6) Alternatively, both transposases and associated MITEs are co-opted by the host bacterium. This is the case for Group 2 and Group 3 RAYTs and their associated REPINs. REPINs differ markedly from MITEs. While REPINs duplicate slowly (~10^{-8} per host

replication) and persist for long periods of time [25] MITEs duplicate as fast as their parental elements (~10^{-5} [26]) and persist only briefly before extinction.

By their nature, REPINs share many similarities with **m**iniature-**i**nverted repeat **t**ransposable **e**lements (MITEs) [24,27], which like REPINs, have also evolved from repeat sequences flanking transposons. REPINs, however, are distinct from MITEs in several important regards. Firstly, the evolution of MITEs occurs repeatedly – and from diverse transposon types – with short persistence times that are tied to the fate of the cognate transposon. In contrast, REPIN-RAYT systems have evolved rarely –possibly only once – and have persisted for millions of years (see below and [16]).

Secondly, and a primary distinguishing feature, is the nature of the interaction between MITEs and their cognate transposases, and REPINs and their cognate RAYTs. MITEs directly parasitise transposon-encoded transposase function – with presumably detrimental effects on transposon fitness – whereas parasitism is not possible in the case of REPINs and their cognate RAYTs. This is because RAYT-encoded function does not mediate RAYT duplication. While the relationship between REPIN and RAYT most likely began as exploitative it appears to have evolved toward a mutualistic association (see below) – both in terms of the interaction between RAYT and host cell [15], and between RAYT and REPIN [14].

Thirdly, and a likely consequence of the second point, is duplication rate. MITEs duplicate about once every 100 thousand host-cell generations [26], whereas REPINs duplicate just once every 100 million host-cell generations (see below and [25]). A faster rate of duplication is expected in the case of an exploitative interaction, with selection likely to favour ever increasing rates of transposition until the point of either host-cell extinction, or extinction of the cognate transposon [14,28].

REPIN sequence diversity is evidence of long-term persistence

REPINs persist over millions of years and display patterns of molecular evolution distinct from selfish genetic elements such as transposons and insertion sequences. Evidence comes from analyses of REPIN diversity. Central to such analyses is the relationship between the rate of duplication and the number of mutations per sequence replication (per site mutation

rate multiplied by sequence length). Elements whose duplication rate is much higher than the mutation rate show little within-element diversity, whereas diversity accumulates (through mutational decay) in elements that duplicate at rates similar to the mutation rate. These mutation dynamics are described by Quasispecies theory (also known as mutation-selection model [29]). According to Quasispecies theory, the most common REPIN sequence in the population defines the "master sequence" (mutation class 0 in **Figure 3**). Other sequence categories are defined by the number of nucleotides different from the master sequence. For a given mutation rate the model predicts that REPIN duplication rates are in the order of 10^{-8} duplications per host generation [25]. This means that REPINs are duplicated at a rate that is at least three magnitudes slower than duplication rates of insertion sequences [10].

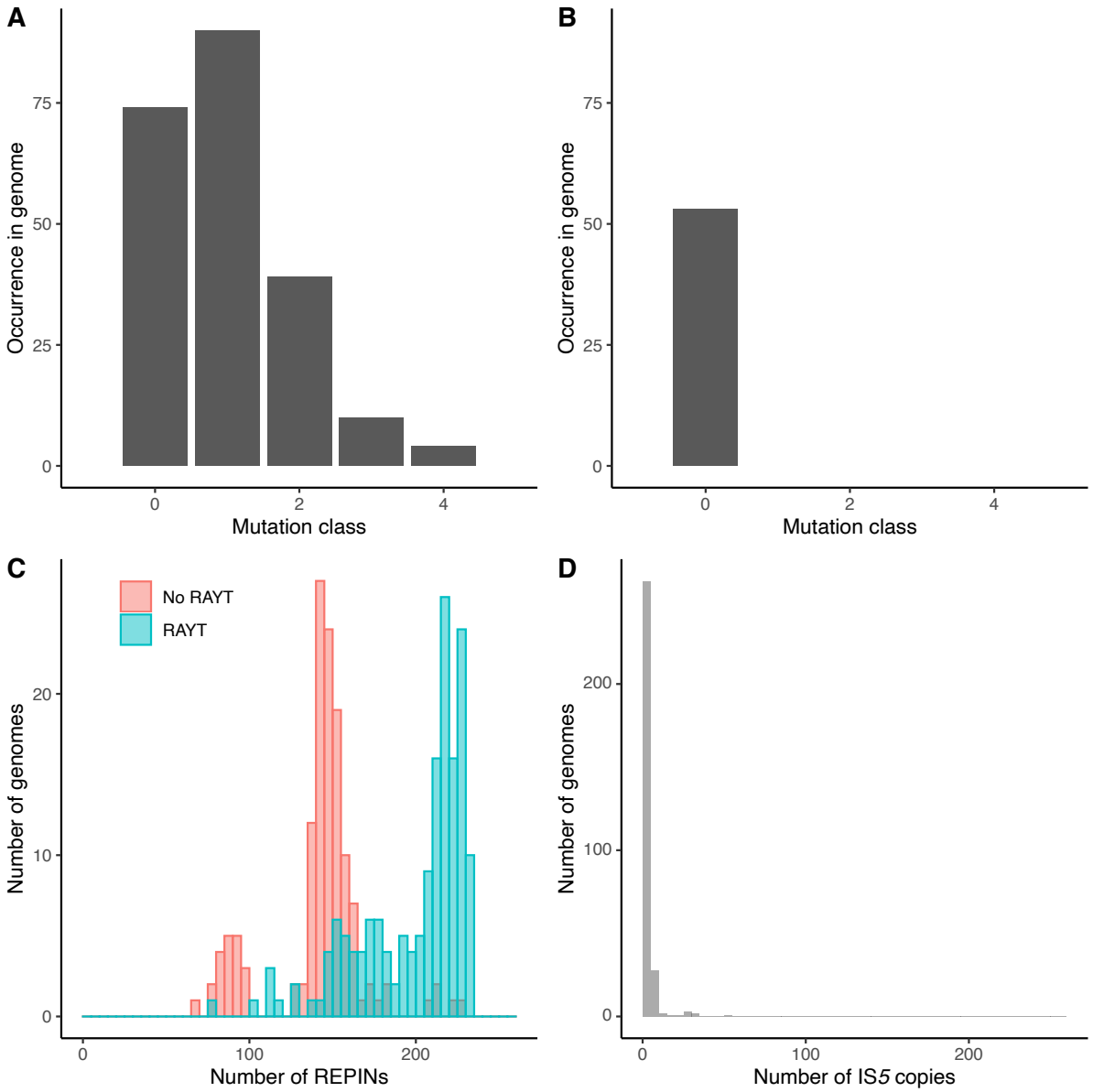


Figure 3. REPIN sequence diversity within genomes and population size distribution across genomes are indicative long persistence times. (**A**) Sequence distribution of the most frequent 21 bp long sequence (part of a REPIN) as well as all relatives from *E. coli* L103-2. Mutation classes are a Quasispecies concept. Mutation class 0 is the master sequence, the most common sequence in a population. Mutation class 1 contains all sequences that differ from the master sequence by exactly one nucleotide. Mutation class 2 are all sequences in the population that differ by exactly two nucleotides and so on. (**B**) The same data as in (A) for the most common 21bp sequence found as part of IS*5* in *E. coli* L103-2. This is also the *E. coli* strain that contains the most IS*5* copies across all 300 strains analyzed. (**C**) REPIN population size across 300 *E. coli* strains (duplicate genomes are not included). Note that *E. coli* strains not containing RAYT genes contain significantly fewer REPINs. (**D**) IS*5* copy numbers in the same 300 *E. coli* genomes. IS*5* copy numbers are skewed to the left of the graph, whereas REPIN population sizes are more reminiscent of a normal distribution with a mean of about 143 (mode of 142) in the case where no RAYT genes are present in the genome (red) or about 200 (mode of 217) where RAYTs are present in the genome (turquoise). (C and D) adapted from Park et al. [14].

Low REPIN duplication rates are consistent with long REPIN persistence times. A REPIN must be present within a genome for approximately 100 million host-cell generations in order to be duplicated. During those 100 million generations REPINs accumulate on average about 0.5 mutations. Hence, for a REPIN to acquire four mutations it must be maintained inside the host genome for about 800 million generations. Accumulation of diversity within REPINs thus requires very long REPIN persistence times.

High levels of REPIN sequence diversity is evident in *E. coli* strains [25]. For example, in *E. coli* L103-2, less than half of all REPINs match the master sequence category. The remainder show a degree of divergence from the master sequence due to acquisition of mutations over extended time periods (**Figure 3A**). Yet, across the entire *E. coli* species the master sequence has not changed: it is identical in every strain.

By way of contrast, the exact same analysis for the insertion sequence IS*5* in *E. coli* L103-2, shows no diversity across IS*5* copies within a single genome (**Figure 3B**). Absence of IS sequence diversity has been observed previously [21], and demonstrates that insertion sequences do not persist long enough to diversify within genomes.

Short persistence times of insertion sequences and long persistence times of REPINs can also be observed in comparisons made among genomes. Across 300 *E. coli* genomes that encompass the currently known sequence diversity of *E. coli*, there is not a single genome from which REPINs are absent (**Figure 3C**). This means that REPINs have been maintained within *E. coli* for at least the last 15 million years, ever since divergence from the most recent common *E. coli* ancestor [25]. Interestingly, some *E. coli* strains have lost the RAYT gene and the REPIN population size in such RAYT-less genomes is significantly reduced, consistent with the role that RAYTs play in maintenance of REPIN populations. In contrast to REPINs, IS*5* is present in less than half of all *E. coli* genomes (133 out of 300, **Figure 3D**). The patchy presence of insertion sequences has been previously reported and indicates, together with the lack of genetic variation within genomes, that insertion sequences are frequently purged from genomes, with constant reinfection of new hosts being necessary for persistence [10].

REPIN populations respond to available niche space

REPINs vary one to another, they replicate and leave offspring copies that resemble parental types. They thus appear to be Darwinian entities in their own right, replete with their own population biology that can be read from patterns of diversity and abundance within bacterial genomes. As with any population of entities, population size depends on available resources. For example, 1 mL of rich broth medium allows *E. coil* populations to reach ~10^9 individuals; 1 L allows the population to expand to 10^{12} cells. In both volumes the density of cells at stationary phase remains the same. REPINs are expected to behave similarly – and drawing a parallel between genome size and resource availability – thus a REPIN population of 100 individuals in a genome of 1 Mb, should increase 10-fold with a 10-fold expansion in genome size. REPIN density should remain constant.

As shown in **Figure 4**, across 2828 representative bacterial chromosomes, the density of 16-mer repeats (a proxy for REPINs abundance in genomes that contain a RAYT gene) remains constant for genomes that contain Group 2 or Group 3 RAYT genes (only Group 2 and Group 3 RAYTs are associated with REPINs [16]). Repeat density decreases with increasing genome length in the absence of RAYT genes. Even in genomes that contain IS*200* genes, the most closely related IS transposase, repeat density decreases with increasing genome length. These data indicate that REPIN population size does indeed increase with available niche space, exactly as expected for populations of any biological entity. In contrast, genomes that do not contain RAYT genes do not contain replicating sequence populations.

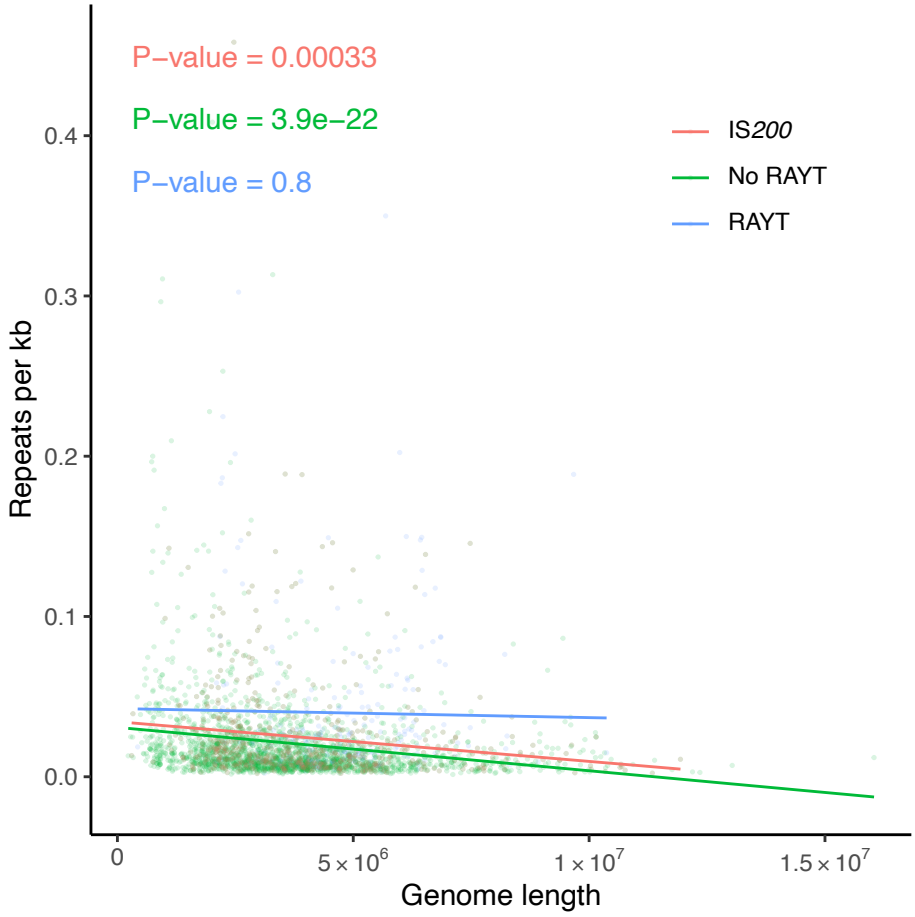


Figure 4. Repeat density remains constant in genomes containing RAYT genes. Repeat density (the frequency of the most abundant 16 bp sequence divided by genome length) significantly decreases in chromosomes that do not contain RAYT genes (2567 chromosomes, green) as well as chromosomes that do not contain RAYT genes, but do contain IS*200* genes (596 chromosomes, orange). Repeat density does not significantly decrease in genomes containing either Group 2 or Group 3 RAYTs (232 chromosomes, blue). We analyzed 2828 representative bacterial chromosomes downloaded from NCBI on the 16.11.2020 https://www.ncbi.nlm.nih.gov/genome/browse/.

REPINs resemble endosymbiotic bacteria

Conceptually, REPINs resemble endosymbiotic bacteria. REPINs are non-infectious, semi-autonomous entities that replicate within bacterial chromosomes. They are distinct from other replicative sequences, such as transposons, insertion sequences, plasmids and integrative / conjugative elements in that they are transmitted vertically, they are not parasites, they form stable, enduring populations, and have persisted across all members of entire species for millions of years (**Figure 5**).

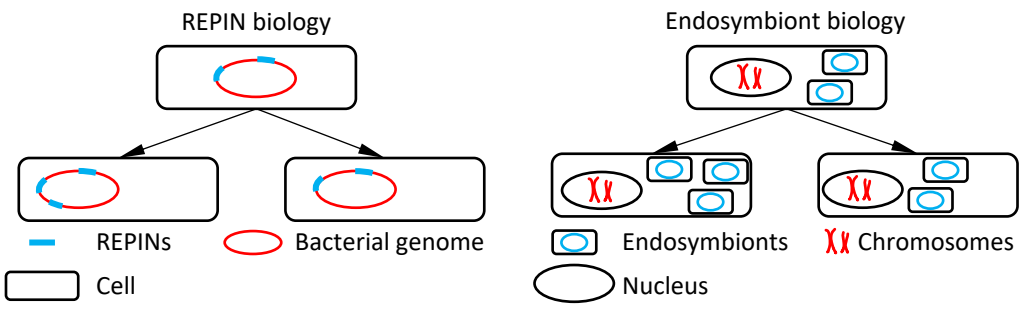


Figure 5. REPIN biology resembles endosymbiont biology. The left panel depicts REPINs that reside in genomes, that duplicate and are passed on vertically to bacterial offspring. REPIN duplication inside the chromosome is mediated through the RAYT transposase at extremely low rates. The right panel shows endosymbionts residing in the cells of an organism as for example, bacteria, such as *Rickettsia* inside arthropod cells, or mitochondria inside the eukaryotic cell. Here endosymbionts duplicate and are passed on vertically to offspring. Chromosomes of endosymbionts are replicated independent of the host's chromosomes, while REPINs are replicated as part of the host chromosome.

The association between REPINs and host genomes is facultative – at least on part of the host – thus bringing the resemblance of REPINs closer to facultative endosymbionts, such as *Hamiltonella* and whiteflies, than to obligate endosymbionts, such as *Buchnera* and aphids, or mitochondria and the eukaryotic cell [3]. RAYTs can be deleted from bacterial genomes, thus depriving REPINs of replicative capacity and eliminating presumed benefits to host, without noticeable effects on cell growth [30]. Moreover, through evolutionary time, RAYTs have been lost from host genomes accompanied by gradual erosion and loss of REPIN populations (**Figure 6A**). Despite loss, descendant lineages remain viable.

However, unlike facultative endosymbiotic bacteria that are transmitted horizontally, that can switch hosts [31,32] and live independently of hosts [3], REPINs are almost exclusively passed to offspring genomes by vertical transmission. While horizontal transfer is possible, establishment of a REPIN in a new host requires that there be a resident RAYT with which the transferred REPIN is compatible. Given the specificity between RAYTs and REPINs within individual lineages – a reflection of persistent co-evolution between REPINs and their cognate RAYT ([15] and **Figure 3**) – the chances of transfer leading to establishment of a new REPIN population is vanishingly small. A related issue is the complete absence of RAYTs from plasmids and other horizontally mobile agents suggesting that carriage of RAYTS on such elements is costly [16].

An additional departure of REPINs from the life history of facultative endosymbionts stems from semiconservative replication of the bacterial genome. Whereas facultative endosymbionts often recolonise hosts from limiting numbers of cells, that is, they frequently passage through population bottlenecks, chromosome replication during cell division means each daughter cell receives equal numbers of REPINs. In this regard, REPINs have much less autonomy compared to even obligate, vertically transmitted, endosymbionts: each generation of an endosymbiont host requires numerous divisions of the endosymbiont, whereas REPINs replicate just once in about 100 million host generations.

A related issue concerns the consequences of replication. Replication of endosymbionts is essential for re-establishment of functional (beneficial) interactions with hosts. While excessive replication could potentially cause harm, for the most part this is not possible, because endosymbionts are often constrained to particular organismal structures, such as, in the vibrio-squid symbiosis, the crypts of the bobtail squid [33]. In the case of REPINs, duplication, unless into a permissive intergenic region, is invariably lethal.

While REPINs are replicative entities nested within a higher order self-replicating structure, just as endosymbiotic bacteria replicate within the higher order structure of, for example, the insect cell (**Figure 5**), REPINs do not share all features typical of endosymbionts. They nonetheless have characteristics typical of symbionts and thus define a new class of semi-autonomous replicator that we suggest might be usefully referred to as "genomic symbionts".

Signatures of intragenomic conflict

Interactions among independently replicating Darwinian populations, such as those that exist between endosymbiotic bacteria and their hosts, or mitochondria and the eukaryotic nucleus, establish conditions that render conflict likely. In instances where one Darwinian population is nested within another, conflict is assured [34]. For example, selection at the level of REPINs is expected to favour variants with enhanced replicative capacity, even though heightened REPIN activity stands to harm host cell lineages. In turn, cell-level

selection is expected to reduce REPIN replication to a level that minimizes harm, most potently *via* effects on RAYTs. While, over evolutionary time, selection will favour those cell lineages containing REPINs whose replicative capacity most closely aligns with the long-term fate of cells, co-existence is likely to be punctuated by periods of conflict.

Ability to capture evidence of conflict – and possibly also conflict resolution – depends on when in the past conflict occurred, the time scale and mode of resolution, and whether signatures of conflict are sufficiently conserved in extant lineages. The inset to **Figure 6B** shows a plausible cyclical dynamic. The blue trajectory on the left depicts the zone of REPIN persistence. Increases in REPIN population size are driven by REPIN-level selection, but counterbalanced by cell-level selection. While persistence within the blue zone is expected to benefit both host and symbiont, REPIN-level selection may nonetheless drive REPIN numbers beyond a critical threshold to the point where cell-level costs become unsustainable. Such instances will lead to extinction of affected cell lineages, or elimination of RAYT function. The latter will result in REPIN decay (both in number and match to the master sequence) and is indicated by the red trajectory. Decay is expected to be a gradual process unfolding over the course of millions of years. There remains the possibility of recovery if RAYT function can be restored.

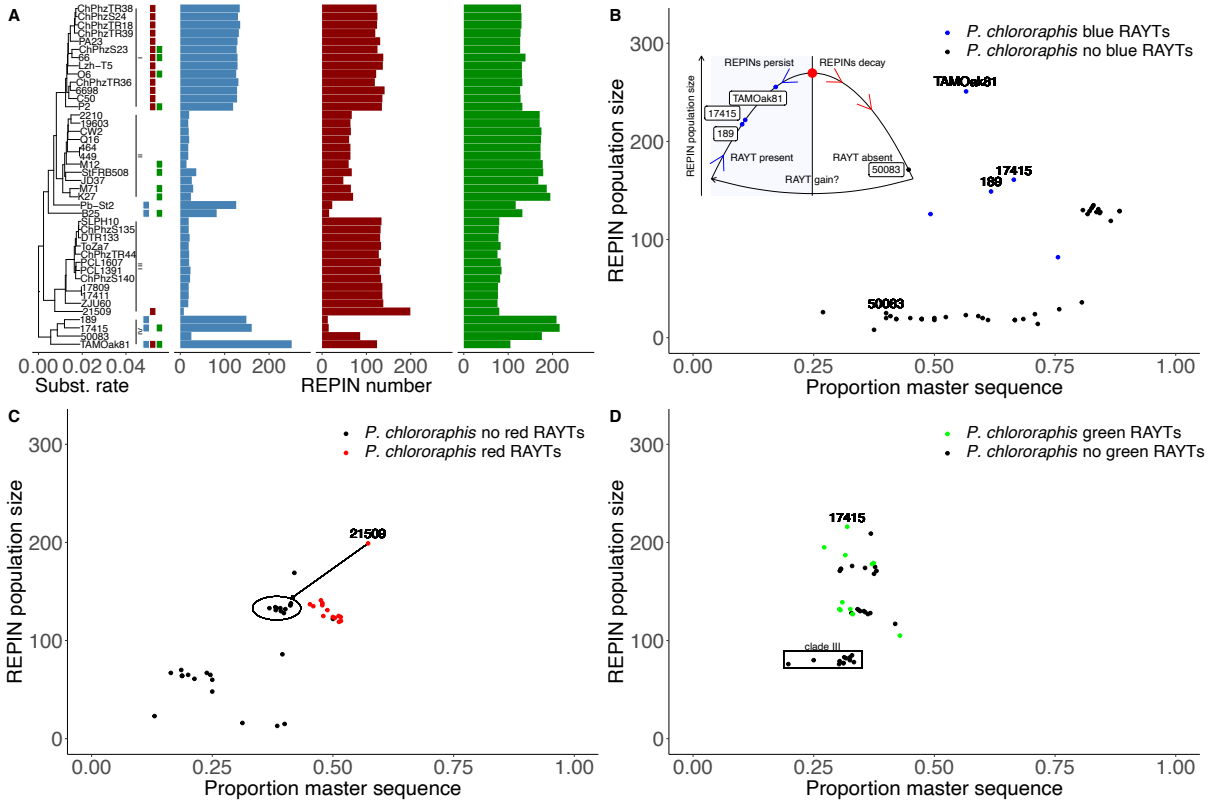


Figure 6. RAYT presence and absence determines REPIN dynamics in *Pseudomonas chlororaphis*. (**A**) The tree on the left side of the figure shows a phylogeny of different *P. chlororaphis* strains. The next column shows the presence and absence of three different RAYT transposases (blue, red and green). The tree was built applying neighbor joining [35] to a distance matrix generated with the program "andi" applied to whole genomes [36]. (**B-D**) The proportion of master sequences (indicates sequence conservation) in a REPIN population and the REPIN population size. According to Quasispecies theory or mutation-selection balance, the higher proportion of master sequences (the most common sequence in the population) correlates with higher duplication rates of the sequence population. The inset in (B) shows a plausible cyclical dynamic of REPINs arising from the tension between REPIN and cell-level selection. See text for details.

We turn to two sets of closely related genomes in which unusual patterns of REPIN-RAYT evolution are suggestive of long-term evolutionary conflicts. The first is a set of 42 strains of the plant associated bacterium *Pseudomonas chlororaphis*; the second is a set of 130 strains of *Neisseria* including *N. gonorrhea* and *N. meningitidis*. Both lineages display similar levels of sequence divergence (about 0.04 substitutions per site), similar also to the divergence observed in *E. coli* [25,37]. If we assume that evolutionary rates are comparable between the two lineages and *E. coli*, then both, *P. chlororaphis* and the two *Neisseria* species, diverged approximately 15 million years ago.

Figure 6A shows the phylogenetic relationship among *P. chlororaphis* strains. The coloured (blue, red and green) boxes indicate the presence / absence of three different RAYT transposases. Each RAYT type is found at the same position in the *P. chlororaphis* genome and forms a monophyletic group in the RAYT phylogeny (**Supplementary Figure 1**). It is likely that the ancestral genotype contained all three RAYTs. Also depicted in **Figure 6A** is the REPIN population size corresponding to each RAYT type.

Patterns of REPIN evolution most consistent with conflict and resolution can be seen in clades III and IV. Focusing firstly on clade IV: the ancestral type of this set of four strains contained a blue RAYT. Strains 189, 17415 and TAMOak81 retain the blue RAYT: each harbours between 150 and 250 REPINs (blue labelled dots in **Figure 6B**, also indicated on the inset) and more than 60% of these REPINs are an exact match to the master sequence. This is in striking contrast to 50083 (black labelled dot in **Figure 6B** and also inset) from which the blue RAYT has been eliminated. Strain 50083 contains few REPINs with few matching the master sequence.

Clade III comprises 12 strains that, with the exception of strain 21509, have lost all three RAYTs. Based on patterns of REPIN abundance and proportion of master sequence it is likely that blue RAYTs were the first to be eliminated, followed by green and lastly red. However, 21509 has retained the red RAYT, with this strain containing more than 200 associated REPINs and 60% matching the master sequence (**Figure 6C**). In the 11 strains lacking the red RAYT the number of REPINs has declined along with matches to the master sequence (black circled dots in **Figure 6C**).

Turning to the green RAYT family, patterns of conflict are less evident than for blue and red, with green RAYTs being evident in three of the four marked clades. This suggests that loss of green may be overall more recent and thus signatures of conflict less prominent. In clade III, where green RAYTs are absent, decay of REPINs is apparent (black rectangle in **Figure 6D**). It is notable that 17415 (Clade IV), which contains a green RAYT, harbours the largest number of REPINs. The sister taxon (strain 189) has lost the green RAYT, possibly recently, and contains the second largest REPIN population.

An additional observation warranting comment concerns clade I. This clade harbours large numbers of all three REPIN families. The red RAYT is present in all strains, with the green RAYT evident in just four strains, while blue RAYTs are absent. Lack of blue RAYTs begs an explanation for the abundance of blue REPINs. A distinct possibility is that red RAYTs have evolved the capacity to aid persistence of both red and blue REPINs. Such a possibility is consistent with the fact that the master sequence of blue and red REPINs differ by just four nucleotides over the 42 nucleotides that define the master sequence. This raises the intriguing possibility of competition between RAYTs for associated REPINs.

Figure 7A shows the phylogenetic relationship among strains belonging to two different *Neisseria* species, the number of RAYTs, and corresponding REPIN population size. Both species contain the same RAYT type and all REPINs share the same identical master sequence. Strains of the *N. gonorrhoeae* clade all share a non-functional RAYT caused by a frameshift mutation (red bars in **Figure 7A**), and all show evidence of REPIN decay (blue dots in **Figure 7B**). In contrast, most strains of *N. meningitidis* harbour functional RAYTs and show no sign of REPIN decay (green dots in **Figure 7B**).

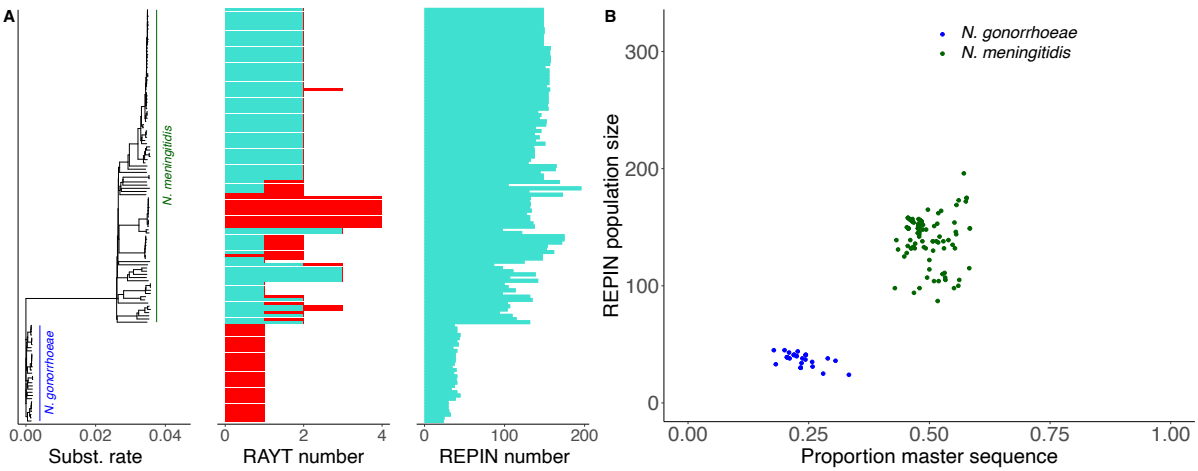


Figure 7. REPIN population size is regulated in *Neisseria*. (A) The tree on the left side of the figure shows a phylogeny of different *N. meningitidis* and *N. gonorrhoeae* strains. The first barplot shows the number of RAYT transposases found in each strain. The second barplot shows the number of REPINs present in the largest sequence cluster. The tree was built by applying neighbor joining [35] to a distance matrix generated with "andi" from complete genomes [36]. **(B)** The proportion of master sequences (proxy for REPIN duplication rate and sequence conservation) for REPIN populations found in *N. gonorrhoeae* (blue) and *N. meningitidis* (green) plotted in relation to population size. Once the RAYT is lost REPIN sequences decay and populations shrink (moving to the bottom left of the graph), as observed for populations from *N. gonorrhoeae.*

A curious feature of *N. meningitidis* is the presence of multiple RAYT copies. This is likely caused by *in trans* transposition, because all RAYT genes are linked to an IS*1106* fragment. Transposition is likely mediated by a full length IS*1106* copy that is present in multiple copies in *N. meningitidis* but absent from *N. gonorrhoeae*. While some RAYTs are present in multiple copies, not all of these are functional. Intriguingly, all non-functional RAYTs in *N. meningitidis* – as in *N. gonorrhoeae* – are caused by frameshift mutations (red bars in **Figure 7A**). In most instances the frameshifts occur in short homopolymeric tracts (**Supplementary Figure 2**) raising the possibility that gain and loss of RAYT function *via* replication slippage may be an adaptive mechanism allowing tension between interests of REPINs and cells to be readily resolved [38].

While presenting the *Neisseria* data set in the context of conflict between the interests of REPINs and host cells, an alternate possibility is that the change in niche preference associated with divergence of *N. gonorrhoeae* from *N. meningitidis* led to selection against functional REPIN-RAYT systems in *N. gonorrhoeae.* Distinguishing between these competing hypotheses is not possible on the basis of current data, but the apparent maladaptive nature of REPIN-RAYT systems in *N. gonorrhoeae* may warrant future investigation. With reference to the inset of **Figure 6B**, this would be the equivalent of changes in the environment leading to alteration in the balance of effects arising from selection on REPINs versus selection at the level of cells – in effect, a "squashing" of the triangle.

Evidence that REPINs and RAYTs are not unique genomic symbionts

It is conceivable that genomic symbionts evolved on more than one occasion. A starting point for such an inquiry is a search for highly abundant 16mers followed by identification of RAYTs. There is no reason to restrict the search to repetitive 16mers, but it was the discovery of over-represented 16mers that lead to the discovery of REPINs [15].

Table 1 shows ten genomes with the highest 16mer frequency arising from interrogation of 2828 bacterial chromosomes. Eight of these genomes contain at least one RAYT gene, hence these 16mers define REPIN populations. However, there are two exceptions: *Dokdonia donghaensis* DSW-1 and *Simonsiella muelleri* ATCC 29453. In both strains, 16mers are

distributed across the entire genome and occur mostly as inverted repeats and thus show features typical of REPINs.

Table 1. Bacterial genomes with the most repetitive 16mers.

Bacterium	Genome length	16mer freq.	RAYT present?
Nostoc sphaeroides	6531820	906	Yes
Pseudomonas corrugata	6124363	918	Yes
Limnospira fusiformis SAG 85.79	6423694	949	Yes
Pseudomonas trivialis	6452803	964	Yes
Dokdonia donghaensis DSW-1	3293944	1032	No
Oscillatoria nigro-viridis PCC 7112	7479014	1089	Yes
Simonsiella muelleri ATCC 29453	2469862	1132	No
Pseudomonas orientalis	5986236	1211	Yes
Moorea producens PAL-8-15-08-1	9673108	1825	Yes
Microcystis panniformis FACHB-1757	5686839	1990	Yes

Three additional fully sequenced *Dokdonia* strains are currently available: *D.* sp. 4h-3-7-5, *D.* sp. Dokd-P16 and *D.* sp. MED134. Each contains similar repeats to *D. donghaensis* DSW-1, with identical master sequences (except for *D.* sp. Dokd-P16, where the master sequence differs by one nucleotide) [39]. The inferred duplication rate is in the range of 10^{-8} and REPIN-like population sizes are between 188 to 289 individuals, which is typical of REPIN populations. None of these strains contains a RAYT gene. Similarly, no RAYT is present in the additionally sequenced genomes of the *Simonsiella* genus.

It is possible that the RAYT genes in these organisms have been recently lost as is evident in certain of the *P. chlororaphis* strains described above (**Figure 6**). If this is true, then additional sequencing of related strains will reveal their presence. However, it is entirely possible that the repeats present in *Simonsiella* and *Dokdonia* define a new intragenomic sequence population that has evolved from a different and currently unknown transposase.

Conclusion

REPINs and their cognate RAYTs are the first members of a newly recognized, yet widespread class of biological entity that we suggest are genomic symbionts. These Darwinian entities have persisted for hundreds of millions of years, have population biology typical of biological organisms, and as expected, given their nesting within higher order structures, show evidence of periodic tension between the replicative interests of REPINs and host cells.

An obviously missing part of the symbiont narrative is knowledge of the contribution that the REPIN-RAYT system makes to host cell fitness. There are many possibilities, ranging from localised effects on mutation and recombination rate [40], to facilitation of genome amplification [37,38], and modulation of tertiary genome structure [41-43]. REPINs may act as preferential targets for insertion sequences [44,45], or have specific effects on gene expression, either through transcription termination [46], or effects on the stability of mRNA [47-49]. An additional possibility are roles in transcription activation and or repression [18]. Studies investigating such possibilities are underway, with mounting evidence suggesting a role in modulation of mRNA stability under conditions of environmental change.

Materials and Methods

Graphs were generated in R using the packages ape, ggplot2, ggtree and cowplot [50-53].

Identifying REPINs in bacterial genomes

REP and REPIN sequences were identified within bacterial species using a specific program that is available through a webserver (http://rarefan.evolbio.mpg.de).

REP sequence groups were identified using a previously described procedure [15]. First, we extracted all 21bp long sequences that occurred more than 55 times in the genome (*P. chlororaphis* TAMOak81, *Dokdonia* sp. 4h-3-7-5 and *Neisseria meningitidis* 2594). We then grouped all sequences that occurred in the genome at least once at a distance of less than 30bp.

For *Neisseria* this procedure identified six REP sequence groups [54]. Group 1 and Group 6 yielded identical REPINs (most common 21bp long sequence: ATTCCCGCGCAGGCGGGAATC occurs 241 times in *N. meningitidis* 2594). None of the remaining four sequence groups yielded REPINs (i.e., they did not occur in opposite orientation at a distance of less than 130bp). Group 2 (CTTGTCCTGATTTTTGTTAAT occurs 207 times) overlaps with the CORREIA sequence TATAGTGGATTAACAAAAATCAGGACAA [55,56].

There are six REP sequence groups in *P. chlororaphis* [57]. Group 1 and Group 6 generated the same REPINs, so did Group 2 and Group 4 as well as Group 3 and Group 5. We analyzed REPINs formed by REP Groups 1 to 3, which we assigned the colours blue, red and green respectively. We did not include Group 6 in our analysis, because the number of REPINs contained was low (<40) and the REPINs were not associated with a RAYT transposase. Each REP sequence group was uniquely defined by the most common 21bp long sequence in that group. The most common 21bp long sequences in the first three sequence groups were: TAGGAGCGAGCTTGCTCGCGA (blue, occurs 466 times in TAMOak81); TCGCGGGCAAGCCTCGCTCCT (red, occurs 238 times); CGCAGCCTGCGGCAGCGGCTA (green, occurs 226 times).

There was only a single REP sequence group in *Dokdonia* (TTACGCTTTCGCGAAAGCGTA occurs 154 times) that forms REPINs [39].

From these REP sequence groups, we determined REPINs (two REP sequences in inverted orientation) across all strains in this study as described previously [25]. First, from the REP seed sequence we generate all single point mutations. We then determined whether any of the generated mutant 21 bp long sequences are present in the genome. If so, we recursively generated all possible point mutants of the sequences present in the genome until no more sequences were found (files ending in ".ss.REP" [39,54,57]). For all identified REP sequences genome positions were determined. Any two sequences that occur at a distance of less than 130bp, with the two sequences occurring on opposite strands (in opposite orientations), are defined as forming a REPIN. Although an entire REPIN consists of three parts (conserved 3' and 5' sequence as well as the variable spacer region), here we only analyzed evolution of the 3' and 5' conserved regions of the REPIN (found in files ending in ".ss" [39,54,57]). The files contain fasta formatted sequences where the sequence name contains all the genomic positions, at which the REPIN was identified. The sequence is a concatenation of the 5' and 3' 21bp sequence of the REPIN. REP singlets are all sequences, for which a partner sequence could not be identified. In this case 21As are concatenated to the end of the sequence.

Files are categorized into subfolders starting with the strain name and ending in "_[number]", where the number corresponds to the REPIN type. There is only a single REPIN type in *Neisseria* (i.e. ending in "_0"). In *P. chlororaphis* there are three REPIN types: blue ending in "_0", red ending in "_1" and green ending in "_2".

Once all REPINs in a single genome for a single REP sequence group were identified, we compared the conserved parts of each REPIN (total length of 42bp). Because these data also contain sequences that are likely not to be mobile anymore we defined a REPIN population as all REPINs that differ by no more than three nucleotides from the master sequence. Clusters consisting solely of REP sequences were ignored.

Identify RAYTs in the genome

To identify RAYT genes in *Neisseria* we ran TBLASTN with the RAYT protein (NMAA_0235) from *N. meningitidis* WUE 2594 on 130 *N. meningitidis* and *N. gonorrhoeae* genomes with an e-value cut-off of 1e-80 (only RAYTs mobilizing REPINs are identified, other RAYTs are ignored). For *P. chlororaphis* we used the RAYT protein (PFLU4255) from *P. fluorescens* SBW25 on 42 *P. chlororaphis* genome, with an e-value cut-off of 1e-20 (three divergent RAYT sequences associated with each of the REPIN populations can be identified). The the datasets can be downloaded under https://zenodo.org/record/4043615 (*P. chlororaphis*) [57], https://doi.org/10.5281/zenodo.4049437 (*Neisseria*) [54] and https://doi.org/10.5281/zenodo.4117577 (*Dokdonia*) [39].

RAYT alignments were performed using MUSCLE with default parameters [58]. Phylogenies were calculated in PHYML [59].

Whether RAYTs belong to a specific REPIN population (colour) in *P. chlororaphis* was determined in three different ways. First, RAYTs that are associated with a specific REPIN population are also flanked by a REP sequence / REPIN from this population. Second, RAYTs that are associated with the same REPIN population are found in the same extragenic space in the genome (between the same genes). Third, a phylogenetic analysis shows that all RAYTs of the same group form monophyletic clades in the phylogeny (**Supplementary Figure 1**).

Analyzing 16mer repeats across bacterial genomes

We downloaded complete, representative bacteria from NCBI (http://www.ncbi.nlm.nih.gov/genome/browse/) on the 16.11.2020. This resulted in a list of 2667 genomes (2828 chromosomes). For each of the chromosomes we determined the frequency of the most abundant 16mer.

We also determined whether a genome contained a RAYT gene via TBLASTN [60] using PFLU4255 from SBW25 (e-Value threshold set to 0.01). The amino acid sequences from all identified RAYTs were then compared to each other using the Needleman–Wunsch algorithm. To the distance matrix we applied MCL with standard settings to identify sequence clusters as described previously (see [16] for details). IS*200* genes were identified

via TBLASTN using an IS*200* gene from *E. coli* RHB29-C15 gene QMG24152.1 (e-Value threshold set to 0.001). If a match had an e-Value of above 0.001, then the genome was designated an IS*200* containing genome.

Acknowledgements

We thank Jenna Gallie for helpful comments on the manuscript, Eric Hugoson for bioinformatics support, and generous core funding from the Max Planck Society.

References

1. Szathmáry E, Smith JM. The major evolutionary transitions. Nature. 1995;374: 227–232. doi:10.1038/374227a0

2. Margulis L. Symbiosis in cell evolution: Life and its environment on the early earth. San Francisco: W. H. Freeman and Company; 1981.

3. Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 2008;42: 165–190. doi:10.1146/annurev.genet.41.110306.130119

4. Sachs JL, Essenberg CJ, Turcotte MM. New paradigms for the evolution of beneficial infections. Trends Ecol Evol. 2011;26: 202–209. doi:10.1016/j.tree.2011.01.010

5. Pochon X, Stat M, Takabayashi M, Chasqui L, Chauka LJ, Logan DDK, et al. Comparison of Endosymbiotic and Free-Living Symbiodinium (Dinophyceae) Diversity in a Hawaiian Reef Environment. J Phycol. 2010;46: 53–65. doi:10.1111/j.1529-8817.2009.00797.x

6. Nyholm SV, McFall-Ngai M. The winnowing: establishing the squid– *vibrio* symbiosis. Nat Rev Micro. 2004;2: 632–642. doi:10.1038/nrmicro957

7. Vaidya N, Manapat ML, Chen IA, Xulvi-Brunet R, Hayden EJ, Lehman N. Spontaneous network formation among cooperative RNA replicators. Nature. 2012;491: 72–77. doi:10.1038/nature11549

8. Mclintock B. The origin and behavior of mutable loci in maize. Proc Natl Acad Sci U S A. 1950;36: 344–355.

9. Startek M, Le Rouzic A, Capy P, Grzebelus D, Gambin A. Genomic parasites or symbionts? Modeling the effects of environmental pressure on transposition activity in asexual populations. Theor Popul Biol. 2013;90: 145–151. doi:10.1016/j.tpb.2013.07.004

10. Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, et al. Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*. Genetics. 1987;115: 51–63.

11. Bichsel M, Barbour AD, Wagner A. Estimating the fitness effect of an insertion sequence. J Math Biol. 2013;66: 95–114. doi:10.1007/s00285-012-0504-2

12. Sawyer S, Hartl D. Distribution of transposable elements in prokaryotes. Theor Popul Biol. 1986;30: 1–16. doi:10.1016/0040-5809(86)90021-3

13. Dolgin ES, Charlesworth B. The fate of transposable elements in asexual populations. Genetics. 2006;174: 817–827. doi:10.1534/genetics.106.060434

14. Park HJ, Gokhale CS, Bertels F. How sequence populations persist inside bacterial genomes. Genetics. 2021;217. doi:10.1093/genetics/iyab027

15. Bertels F, Rainey PB. Within-Genome Evolution of REPINs: a New Family of Miniature Mobile DNA in Bacteria. PLoS Genet. 2011;7: e1002132. doi:10.1371/journal.pgen.1002132

16. Bertels F, Gallie J, Rainey PB. Identification and Characterization of Domesticated Bacterial Transposases. Genome Biol Evol. 2017;9: 2110–2121. doi:10.1093/gbe/evx146

17. Nunvar J, Huckova T, Licha I. Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. BMC Genom. 2010;11: 44. doi:10.1186/1471-2164-11-44

18. Bertels F, Rainey PB. Curiosities of REPINs and RAYTs. Mob Genet Elements. 2011;1: 262–268. doi:10.4161/mge.18610

19. Ton-Hoang B, Siguier P, Quentin Y, Onillon S, Marty B, Fichant G, et al. Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. Nucleic Acids Res. 2012;40: 3596–3609. doi:10.1093/nar/gkr1198

20. Higgins CF, Ames GF, Barnes WM, Clement JM, Hofnung M. A novel intercistronic regulatory element of prokaryotic operons. Nature. 1982;298: 760–762.

21. Lawrence JG, Ochman H, Hartl DL. The evolution of insertion sequences within enteric bacteria. Genetics. 1992;131: 9–20.

22. Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, et al. Single-stranded DNA transposition is coupled to host replication. Cell. 2010;142: 398–408. doi:10.1016/j.cell.2010.06.034

23. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41: 331–368. doi:10.1146/annurev.genet.40.110405.090448

24. Wessler SR, Bureau TE, White SE. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev. 1995;5: 814–821.

25. Bertels F, Gokhale CS, Traulsen A. Discovering Complete Quasispecies in Bacterial Genomes. Genetics. 2017;206: 2149–2157. doi:10.1534/genetics.117.201160

26. Sousa A, Bourgard C, Wahl LM, Gordo I. Rates of transposition in *Escherichia coli*. Biol Lett. 2013;9: 20130838. doi:10.1098/rsbl.2013.0838

27. Hartl DL, Lozovskaya ER, Lawrence JG. Nonautonomous transposable elements in prokaryotes and eukaryotes. Genetica. 1992;86: 47–53.

28. Consuegra J, Gaffé J, Lenski RE, Hindré T, Barrick JE, Tenaillon O, et al. Insertion-sequence-mediated mutations both promote and constrain evolvability during a long-term experiment with bacteria. Nat Commun. 2021;12: 980. doi:10.1038/s41467-021-21210-7

29. Wilke CO. Quasispecies theory in the context of population genetics. BMC Evol Biol. 2005;5: 44–8. doi:10.1186/1471-2148-5-44

30. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol Syst Biol. 2006;2: 2006.0008. doi:10.1038/msb4100050

31. Burke GR, Normark BB, Favret C, Moran NA. Evolution and Diversity of Facultative Symbionts from the Aphid Subfamily Lachninae. Appl Environ Microbiol. 2009;75: 5328–5335. doi:10.1128/AEM.00717-09

32. Degnan PH, Moran NA. Evolutionary genetics of a defensive facultative symbiont of insects: exchange of toxin-encoding bacteriophage. Mol Ecol. 2008;17: 916–929. doi:10.1111/j.1365-294X.2007.03616.x

33. McFall-Ngai MJ. The Importance of Microbes in Animal Development: Lessons from the Squid-Vibrio Symbiosis. Annu Rev Microbiol. 2014;68: 177–194. doi:10.1146/annurev-micro-091313-103654

34. Rainey PB, De Monte S. Resolving Conflicts During the Evolutionary Transition to Multicellular Life. Annu Rev Ecol Evol Syst. 2014;45: 599–620. doi:10.1146/annurev-ecolsys-120213-091740

35. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4: 406–425.

36. Haubold B, Klötzl F, Pfaffelhuber P. andi: fast and accurate estimation of evolutionary distances between closely related genomes. Bioinformatics. 2015;31: 1169–1175. doi:10.1093/bioinformatics/btu815

37. Ochman H, Wilson AC. Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. J Mol Evol. 1987;26: 74–86.

38. Moxon ER, Rainey PB, Nowak MA, Lenski RE. Adaptive evolution of highly mutable loci in pathogenic bacteria. Curr Biol. 1994;4: 24–33. doi:10.1016/s0960-9822(00)00005-1

39. Bertels F, Fortmann-Grote C, Rainey P. REPIN population analysis in 4 *Dokdonia* genomes. 1st ed. Zenodo; 2020. doi:10.5281/zenodo.4117577

40. Shyamala V, Schneider E, Ames GF. Tandem chromosomal duplications: role of REP sequences in the recombination event at the join-point. EMBO J. 1990;9: 939–946.

41. Yang Y, Ames GF. DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. Proc Natl Acad Sci U S A. 1988;85: 8850–8854.

42. Qian Z, Macvanin M, Dimitriadis EK, He X, Zhurkin V, Adhya S. A New Noncoding RNA Arranges Bacterial Chromosome Organization. mBio. 2015;6: 635. doi:10.1128/mBio.00998-15

43. Qimron U, Marintcheva B, Tabor S, Richardson CC. Genomewide screens for *Escherichia coli* genes affecting growth of T7 bacteriophage. Proc Natl Acad Sci U S A. 2006;103: 19039–19044. doi:10.1073/pnas.0609428103

44. Silby M, Cerdeno-Tarraga A, Vernikos G, Giddens S, Jackson R, Preston G, et al. Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. Genome Biol. 2009;10: R51. doi:10.1186/gb-2009-10-5-r51

45. Tobes R, Pareja E. Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. BMC Genom. 2006;7: 62.

46. Espéli O, Moulin L, Boccard F. Transcription attenuation associated with bacterial repetitive extragenic BIME elements. J Mol Biol. 2001;314: 375–386. doi:10.1006/jmbi.2001.5150

47. Newbury SF, Smith NH, Robinson EC, Hiles ID, Higgins CF. Stabilization of translationally active mRNA by prokaryotic REP sequences. Cell. 1987;48: 297–310.

48. Higgins CF, McLaren RS, Newbury SF. Repetitive extragenic palindromic sequences, mRNA stability and gene expression: evolution by gene conversion? A review. Gene. 1988;72: 3–14.

49. Liang W, Rudd KE, Deutscher MP. A Role for REP Sequences in Regulating Translation. Mol Cell. 2015;58: 431–439. doi:10.1016/j.molcel.2015.03.019

50. Wickham H. ggplot2. Springer; 2016.

51. Wilke CO. Streamlined Plot Theme and Plot Annotations for “ggplot2” [R package cowplot version 0.8.0]. 0 ed. Comprehensive R Archive Network (CRAN); 2016. Available: https://CRAN.R-project.org/package=cowplot

52. Yu G, Lam TT-Y, Zhu H, Guan Y. Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. Mol Biol Evol. 2018;35: 3041–3043. doi:10.1093/molbev/msy194

53. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35: 526–528. doi:10.1093/bioinformatics/bty633

54. Bertels F, Fortmann-Grote C, Rainey P. REPIN population analysis in 130 *Neisseria meningitidis* and *N. gonorrhoeae* genomes. 1st ed. Zenodo; 2020. doi:10.5281/zenodo.4049438

55. Correia FF, Inouye S, Inouye M. A 26-base-pair repetitive sequence specific for Neisseria gonorrhoeae and Neisseria meningitidis genomic DNA. J Bacteriol. 1986;167: 1009–1015. doi:10.1128/jb.167.3.1009-1015.1986

56. Roberts SB, Spencer-Smith R, Shah M, Nebel J-C, Cook RT, Snyder LAS. Correia Repeat Enclosed Elements and Non-Coding RNAs in the *Neisseria* Species. Microorganisms. 2016;4: 31. doi:10.3390/microorganisms4030031

57. Bertels F, Fortmann-Grote C, Rainey PB. REPIN population analysis in 42 *Pseudomonas chlororaphis* genomes. 1st ed. Zenodo; 2020. doi:10.5281/zenodo.4043615

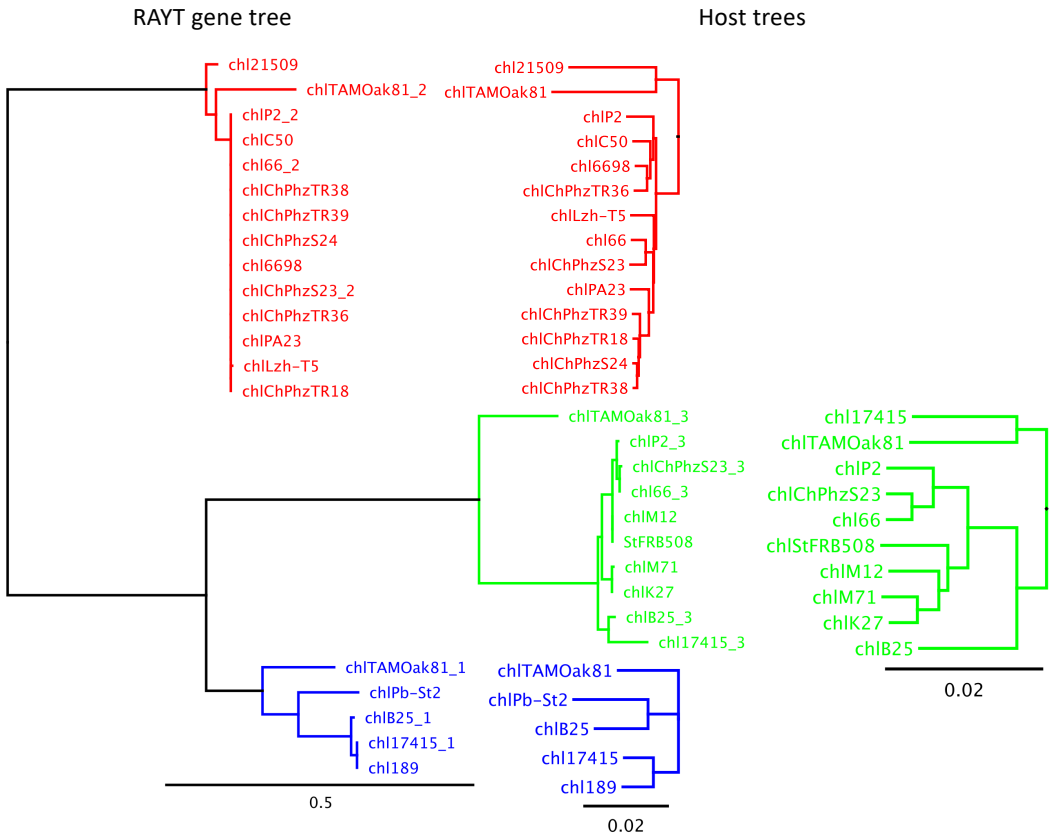
58. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5: 113. doi:10.1186/1471-2105-5-113

59. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 2003;52: 696–704.

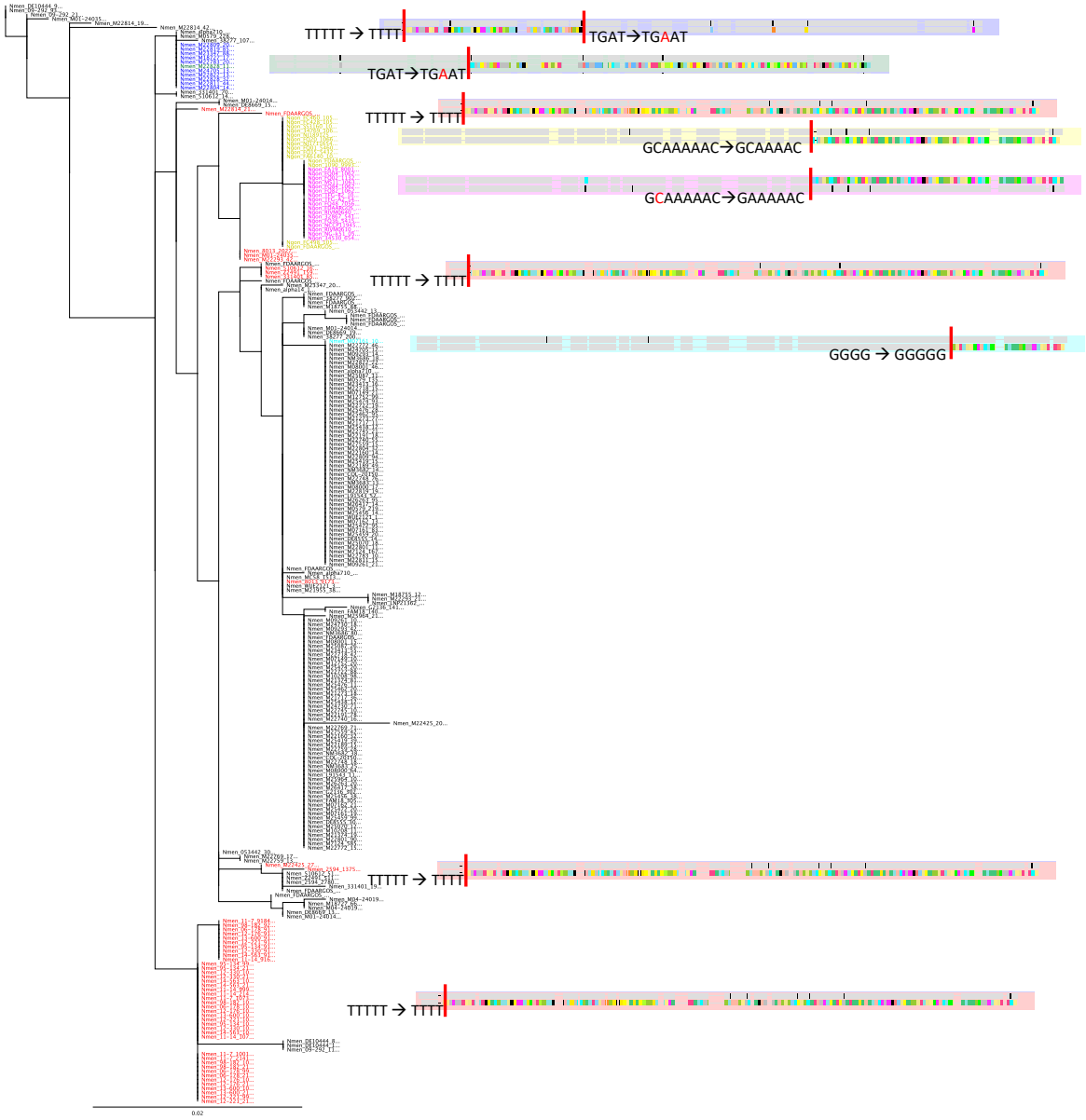
60. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1006/jmbi.1990.9999

61. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biol. 2010;59: 307–321. doi:10.1093/sysbio/syq010

Supplementary Figures



Supplementary Figure 1. The figure shows the RAYT phylogeny on the left and the corresponding host phylogeny for each of the RAYTs on the right side. The host phylogenies were determined from complete genome sequences using andi [36] and PHYML [59]. The RAYT phylogeny was built from alignments generated with MUSCLE [58] and the phylogeny was inferred with PHYML [59].



Supplementary Figure 2. Phylogeny of all *Neisseria* RAYTs. RAYT genes are named by the *Neisseria* strain in which they are contained. All RAYTs from *N. gonorrhoeae* contain a frameshift mutation (yellow and purple). Different colours indicate different frameshift mutations. Note that the red frameshift mutation occurs multiple times independently across the phylogeny and that occasionally frameshift mutations revert. Red bars indicate the position of the frameshift in the RAYT gene. The phylogeny was built from a MUSCLE [58] alignment using PHYML [61].