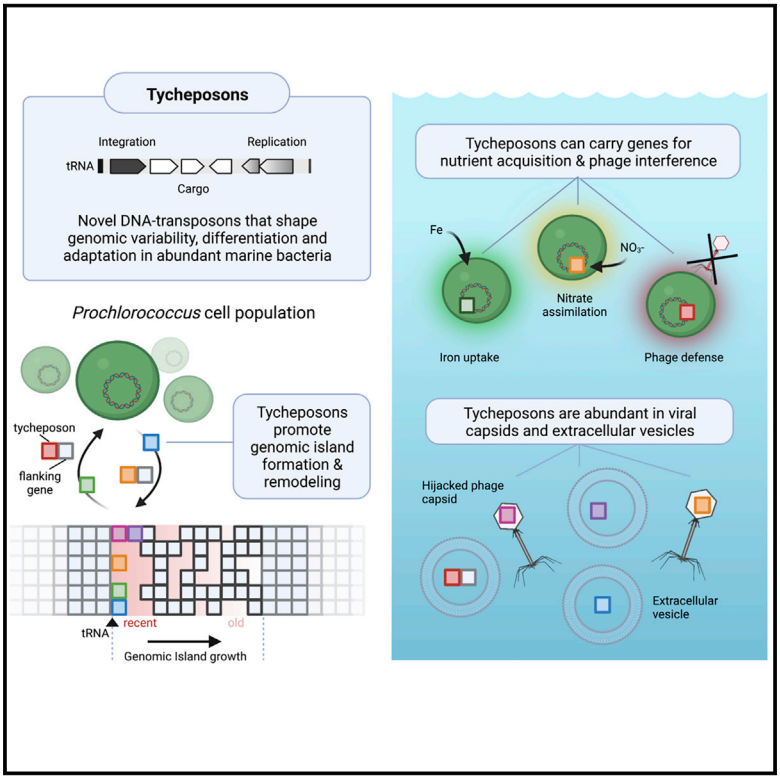
Cell

Novel integrative elements and genomic plasticity in ocean ecosystems

Graphical abstract



Highlights

- Tycheposons are novel DNA transposons promoting genomic adaptation in marine bacteria
- Tycheposons can be viral satellites or carry cargo such as nutrient-acquisition genes
- Tycheposons are abundant in viral capsids and extracellular vesicles in seawater
- Tycheposons accelerate genomic island formation and remodeling

Authors

Thomas Hackl, Raphaël Laurenceau, Markus J. Ankenbrand, ..., Edward F. DeLong, Steven J. Biller, Sallie W. Chisholm

Correspondence

t.hackl@rug.nl (T.H.), chisholm@mit.edu (S.W.C.)

In brief

Tycheposons, a group of mobile genetic elements, facilitate horizontal gene transfer in marine picocyanobacteria. Dispersed through viral capsids and extracellular vesicles, tycheposons promote genomic diversification and adaptation, accelerating microbial evolution across our planet's largest habitat.

Hackl et al., 2023, Cell *186*, 47–62
January 5, 2023
https://doi.org/10.1016/j.cell.2022.12.006

Cell

CellPress

Article

Novel integrative elements and genomic plasticity in ocean ecosystems

Thomas Hackl,[1,2,12,13,*] Raphaël Laurenceau,[1,12] Markus J. Ankenbrand,[1,3,12] Christina Bliem,[1] Zev Cariani,[1] Elaina Thomas,[1] Keven D. Dooley,[1] Aldo A. Arellano,[1] Shane L. Hogle,[1] Paul Berube,[1] Gabriel E. Leventhal,[1] Elaine Luo,[4] John M. Eppley,[4] Ahmed A. Zayed,[6,7] John Beaulaurier,[8] Ramunas Stepanauskas,[9] Matthew B. Sullivan,[5,6,7] Edward F. DeLong,[4] Steven J. Biller,[10] and Sallie W. Chisholm[1,11,*]

[1]Massachusetts Institute of Technology, Department of Civil and Environmental Engineering, Cambridge, MA 02139, USA
[2]Groningen Institute for Evolutionary Life Sciences, University of Groningen, 9700CC Groningen, the Netherlands
[3]University of Würzburg, Center for Computational and Theoretical Biology, 97070 Würzburg, Germany
[4]Daniel K. Inouye Center for Microbial Oceanography, Research and Education, University of Hawai'i Manoa, Honolulu, HI 96822, USA
[5]Department of Microbiology & Department of Civil, Environmental, and Geodetic Engineering, Ohio State University, Columbus, OH 43210, USA
[6]EMERGE Biology Integration Institute, Ohio State University, Columbus, OH 43210, USA
[7]Center of Microbiome Science, Ohio State University, Columbus, OH 43210, USA
[8]Oxford Nanopore Technologies Inc, San Francisco, CA 94501, USA
[9]Bigelow Laboratory for Ocean Sciences, East Boothbay, ME 04544, USA
[10]Wellesley College, Department of Biological Sciences, Wellesley, MA 02481, USA
[11]Massachusetts Institute of Technology, Department of Biology, Cambridge, MA 02139, USA
[12]These authors contributed equally
[13]Lead contact
*Correspondence: t.hackl@rug.nl (T.H.), chisholm@mit.edu (S.W.C.)
https://doi.org/10.1016/j.cell.2022.12.006

SUMMARY

Horizontal gene transfer accelerates microbial evolution. The marine picocyanobacterium *Prochlorococcus* exhibits high genomic plasticity, yet the underlying mechanisms are elusive. Here, we report a novel family of DNA transposons—"tycheposons"—some of which are viral satellites while others carry cargo, such as nutrient-acquisition genes, which shape the genetic variability in this globally abundant genus. Tycheposons share distinctive mobile-lifecycle-linked hallmark genes, including a deep-branching site-specific tyrosine recombinase. Their excision and integration at tRNA genes appear to drive the remodeling of genomic islands—key reservoirs for flexible genes in bacteria. In a selection experiment, tycheposons harboring a nitrate assimilation cassette were dynamically gained and lost, thereby promoting chromosomal rearrangements and host adaptation. Vesicles and phage particles harvested from seawater are enriched in tycheposons, providing a means for their dispersal in the wild. Similar elements are found in microbes co-occurring with *Prochlorococcus*, suggesting a common mechanism for microbial diversification in the vast oligotrophic oceans.

INTRODUCTION

Prochlorococcus is the smallest and numerically most abundant cyanobacterium in the oceans. It possesses a large pangenome and contains hypervariable genomic islands that have been linked to niche differentiation and phage defense.[1–4] Adaptations to light and temperature broadly define clades in the genus, which are subdivided into a mosaic of co-existing subpopulations in the wild.[5–7] This structure provides stability and resilience to the global population in the face of viral predation and changing environmental conditions.[8–10] While the importance of variable genomic islands in the ecology of *Prochlorococcus* is well known, how they are formed and acquire new genes remains an open question because most cells lack common means of horizontal gene transfer, such as conjugative systems[11] and genes for natural competence.[12] Island diversification via gene exchange with cyanophages has been observed for genes involved in photosynthesis, high-light adaptation, and other metabolic functions,[13–15] but evidence for prophage-mediated transduction is rare: while some cyanophages carry integrase genes and putative attachment sites linked to recombination hotspots in *Prochlorococcus* genomes,[16,17] only one partial prophage has been observed in hundreds of available genomes.[18] *Prochlorococcus* cells, moreover, appear devoid of any common mobile genetic elements (MGEs) including plasmids, transposons, insertion sequences (ISs), or integrative and conjugative

elements (ICEs),[19–21] with the exception of a few transposons and ISs previously identified in the most basal *Prochlorococcus* clade LLIV.[17] This overall limited ability to utilize canonical horizontal gene transfer mechanisms—seemingly inconsistent with a large and widely distributed pangenome[7,22–24]—motivated us to explore genomes of cultured and wild *Prochlorococcus* cells[21] for evidence of mechanisms promoting genomic island diversity in this group.

RESULTS

A hidden mobilome in *Prochlorococcus* genomic islands

While the term "genomic island" is sometimes used in reference to individual MGEs,[25] here we use it to identify large chromosomal regions with high interstrain variability and a relatively high density of flexible genes (i.e., those not shared by all genomes), which is consistent with early descriptions of genomic islands in *Prochlorococcus*.[1] In most cases, these regions do not represent mobile units but heterogeneous aggregations of horizontally transferred material.[26]

Searching for clues as to how island hotspots of variability arise and are maintained over evolutionary timescales in this genus, we annotated genomic islands in 623 *Prochlorococcus* genomes from cultured isolates and wild single cells[21] (Table S1). These genomes belong to ten distinct phylogenetic clades, groupings that serve as ecologically and evolutionarily relevant units of differentiation in this system.[5,19,27] We identified well-defined islands using a custom hidden-Markov-model (HMM)-based approach that distinguishes them based on their differential enrichment of flexible genes (Figure S1A). In all genomes, we found on average of 8–10 islands per clade, typically between 4 and 200 kbp in size, that comprise about one-quarter of all genes per genome. The islands harbor more than two-thirds of all flexible genes of the *Prochlorococcus* pangenome (Figure S1B; Table S2).

Consistent with more limited studies,[1,28] most *Prochlorococcus* islands we identified are directly adjacent to 7 tRNA genes ($proline_{tgg}$, $serine_{tga}$, $alanine_{ggc}$, $threonine_{ggt}$, $arginine_{tct}$, one of three $methionines_{cat}$, and the tmRNA gene—a bifunctional transfer-messenger RNA important for releasing stalled ribosomes) (Figure S2). Because tRNA genes are known integration hotspots for a variety of MGEs,[29] we reconsidered the traditional notion that *Prochlorococcus* genomes generally lack MGEs and carefully re-examined the available genomes for signatures of such entities.

First, we used straightforward approaches (see STAR Methods) to screen for the presence of well-known types of MGEs, including prophages, integrative plasmids, transposons, casposons, ISs, ICEs, and pathogenicity- and phage-inducible chromosomal islands (PICIs). These searches did not reveal any clearly recognizable MGEs other than some known ISs and transposons in the most basal *Prochlorococcus* clade (LLIV).[17] The key to unveiling *Prochlorococcus'* mobilome emerged from manually examining a particular, small genomic island (6.9 kbp) we had identified, which shared a notably high nucleotide-level identity (99%) in two genomes from different clades (HLI and HLII; Figure S1C), indicating a horizontal transfer event. This region coded for a large serine recombinase—an integrase-type protein common in MGEs—and eight additional hypothetical genes. Using advanced remote homology detection[30] we further examined the hypothetical genes and identified a putative transcriptional regulator (MerR-like), a putative major capsid protein (HK97-fold), a replicative helicase (DnaB-like), and three smaller genes coding for proteins with remote structural similarity to replication factors often found in phages and other MGEs. Together, these features were consistent with the presence of an MGE in this island, and we used its sequence to iteratively identify more related MGEs and ultimately curate a comprehensive set of hallmark proteins of *Prochlorococcus'* integrase-centered mobilome (Table S3). With these proteins, we identified 937 putative integrase-carrying MGEs across the set of 623 *Prochlorococcus* genomes, providing the basis for the following detailed study on these MGEs and their role in *Prochlorococcus* evolution and ecology.

Tycheposons: Novel DNA transposons with roles in phage interference and nutrient acquisition

As we describe in detail below, about half of the identified *Prochlorococcus* MGEs (501 of 937) constitute a cohesive new family of cargo-carrying DNA transposons (Figures 1 and 2). Some of these MGEs encode functions associated with conferring a fitness advantage under known stresses in the marine environment. We named these novel MGEs tycheposons, referring to the Greek deity Tyche, a guardian of fortune and prosperity and a daughter of Oceanus. As we argue with evidence below, tycheposons form an independent lineage of MGEs that (1) share a common set of hallmark genes facilitating mobility and replication, which are distinct from other MGEs (Figure 2); (2) contain a site-specific integrase—most often a site-specific tyrosine recombinase belonging to a lineage unique to tycheposons—which leaves integrated elements flanked by partial tRNA repeats (Figures 1 and 3); (3) have a conserved gene organization with an inward-facing integrase at one end and an optional, small replication module either next to the integrase or at the opposite end (Figure 1); and finally (4) carry cargo with clear implications in terms of adaptation to the local environment—e.g., facilitating nutrient acquisition and phage interference (Figure 1).

To arrive at this definition of tycheposons, we initially classified all *Prochlorococcus* MGEs in the context of known MGEs using a gene-sharing network approach.[39–41] We co-clustered all hallmark proteins of the *Prochlorococcus* mobilome (those with functions related to a mobile lifestyle, including recombination and DNA replication and packaging)—together with all proteins from NCBI viral RefSeq; all proteins from the mobileOG database (a comprehensive database based on 10 million hallmark proteins sequences of all major classes of MGEs)[42]; and an additional set of proteins representing PICIs which are MGEs with genomic architectures similar to some elements we identified, which is discussed in more detail below. We then computed a bipartite network linking MGEs by shared protein clusters to analyze how the different *Prochlorococcus* MGEs relate to each other as well as to known MGE types (Figure 2).

Based on this network, roughly half (501) of *Prochlorococcus* MGEs—the tycheposons—form a tightly connected cluster encompassing the vast majority of the MGE hallmark proteins providing the basis for defining these elements as a new and

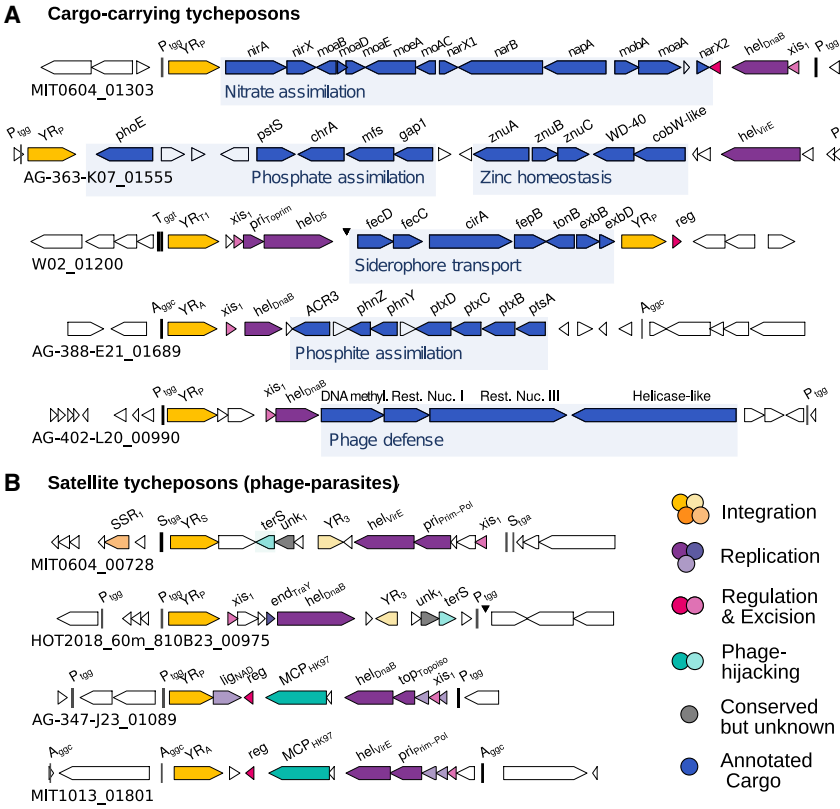


Figure 1. Structure and function of tycheposons in *Prochlorococcus*

(A and B) Examples of two types of tycheposons illustrating their modular structure: (A) cargo-carrying tycheposons selected here because of their clear ecological relevance in ocean ecosystems, and (B) satellite tycheposons, carrying either a *terS* or MCP viral-packaging gene likely used to hijack phage capsids for dispersal. Cargo modules were annotated with roles in nitrate assimilation,[31] siderophore transport,[18] phosphate assimilation,[32] zinc homeostasis,[33] and phosphite assimilation.[34]

Gene labels: full-length and partial tRNA genes are labeled with single-letter amino acid code and their anticodon, e.g., S_{tga}, tRNA-serine$_{TGA}$; YR, tyrosine recombinase; MCP, major capsid protein; terS, terminase small subunit; xis, excisionase; hel, helicase; top, topoisomerase; lig, ligase; reg, transcriptional regulator; pri, primase, primase/polymerase, or primase/helicase; end, endonuclease; SSR, small serine recombinase; unk, conserved unknown. Some annotations are further labeled with more specific subprofiles indicating specific families of helicases, for example (hel_{DnaB} or hel_{VirE}). More detailed information on the gene profiles is available in Table S3.

See also Tables S1 and S3.

independent lineage of MGEs. Only three of the tycheposon-associated protein clusters also contained sequences from well-known MGE classes: a cluster of major capsid proteins contained viral sequences; a truncated tyrosine recombinase found on plasmids and viruses; and a rare (6 proteins in tycheposons) but apparently promiscuous replicative helicase also found in plasmids, ICEs, and viruses. The other half of the MGEs in the network were cryptic elements, lacking identifiable hallmark genes other than integrases and excisionases, and we excluded them from analyses beyond those two genes unless otherwise stated.

In addition to their distinct gene pool, tycheposons differ from other known families of MGEs in a key feature of transposable elements classification: the enzyme that catalyzes their movement.[43] The most abundant prokaryotic transposable elements are IS-like transposons with their autonomous representatives—ISs and composite transposons—encoding transposases that typically lack site-selectivity and enable integration at different locations within the same genome. By contrast, *Prochlorococcus* tycheposons and cryptic elements encode site-specific integrases (910 out of 937 contain a phage-integrase-like tyrosine recombinase and the rest a large serine recombinase) that recognize short sequence motifs—so-called attachment sites.[44] Such integrases have been shown to catalyze the recombination between two attachment sites, facilitating either the integration of excised MGEs into the recipient genome (recombination between one attachment site on the MGE and one in the genome) or the excision of an already integrated MGE (recombination between the two attachment sites flanking the MGE after integration). The directionality of this reversible cut-and-paste transposition mechanism is often regulated by an excisionase, which, if expressed, inhibits reintegration.[44]

Among the different tyrosine recombinases of the *Prochlorococcus* mobilome, those encoded by tycheposons stand out further due to their integration sites and their phylogeny. For integration, they specifically target 40-nucleotide-long attachment sites in 7 tRNA genes that abut *Prochlorococcus*' major genomic islands (Table S4). With the attachment sites located at the start/end of the tRNA genes and the full-length genes reconstituted by the incoming attachment site, the targeted tRNA genes are not functionally disrupted. This integration process, which likely can also occur at partial tRNAs, leaves an integrated tycheposon flanked at least by partial tRNAs on both sides—a hallmark for discerning boundaries of complete tycheposons in genomic analyses (Figure 1).

With few exceptions, the phylogenies we reconstructed for tycheposon integrases and those of other MGEs, viruses, bacteria, and archaea[45] reveal that the former form a monophyletic clade, more closely related to each other than to integrases from any other system (Figure 3). This is unexpected because integrases often jump between viruses, MGEs, and hosts, resulting in more convoluted evolutionary histories.[46] Within the tycheposons, the integrases further cluster by the different tRNA genes they target. Each island has its specific, proximal tRNA and each tRNA its specific integrase; i.e., integrases are island specific. Moreover, based on our phylogenetic tree, tycheposon integrases account for more than 10% of the integrase diversity

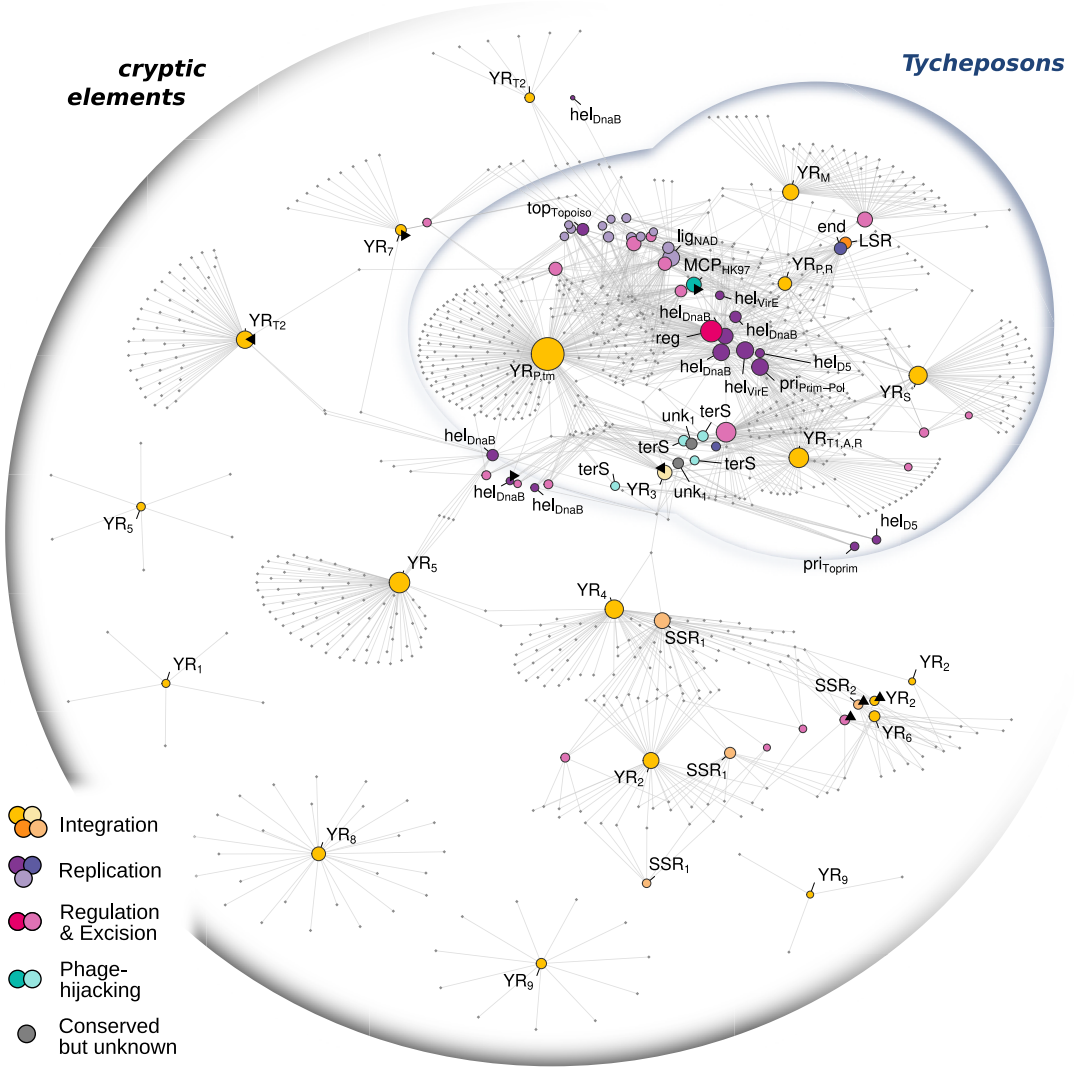


Figure 2. Bipartite gene-sharing network of the *Prochlorococcus* mobilome comprising tycheposons and cryptic MGEs

Visualization of part of a bipartite gene-sharing network based on protein clusters computed from *Prochlococococcus*' mobilome together with a comprehensive set of known MGEs and viral genomes (mobileOG-DB, PICIs, viral RefSeq). Two types of nodes are shown: putative MGEs (small gray dots) and protein clusters (colored by functional category, scaled by abundance in the dataset, labeled as in Figure 1). Connections are drawn between each MGE and the proteins they encode. Here, only *Prochlorococcus* MGEs are shown because (1) a complete visualization of all MGEs and viruses would be computationally infeasible and (2) only a handful of protein clusters of the *Prochlorococcus* mobilome shared a few connections with other MGEs (indicated by small black triangles); i.e., even in a more extensive network, *Prochlorococcus* MGEs form a distinct subgraph. Within this subgraph, tycheposons (blue outline) stand out as a tightly interlinked cluster rarely connected to other MGEs. Compared with known MGEs and viruses, only three clusters shared connections (one rare hel_{DnaB}-cluster and the YR_3 cluster contained proteins from both viruses and other MGEs, and the MCP cluster contained some proteins flagged as viral).
See also Tables S3 and S4.

across various domains of life (Figure 3) (see STAR Methods for tree-based diversity estimation), supporting their ancient origin, and suggesting that the tycheposons carrying them evolved independently over similar timescales.

In addition to an integrase module (~1 kbp) encoding the integrase and sometimes an excisionase, many tycheposons carry genes associated with replication (2–4 kb modules with polymerases, primases, and/or helicases) (Figure 1; Table S4); 44% of tycheposons with clear boundaries (i.e., flanked by attachment sites on both sides) carry diverse replicative helicases, with the most common types belonging to helicase superfamily 4 (RecA-fold, cd01125; rarely DnaB-like, cd00984) and superfamily 3 (VirE-like, PF05272). We also observed some members of helicase superfamily 6 (MCM-like, COG1241). Moreover, 15% of the tycheposons also carry a putative multifunctional primase-polymerase domain (PrimPol, cd04859), either on a gene adjacent to, or fused with, the helicase. These multifunctional primase-polymerase domains suggest that at least some tycheposons have the potential to self-synthesize. This would make them the second group of prokaryotic transposons with such a

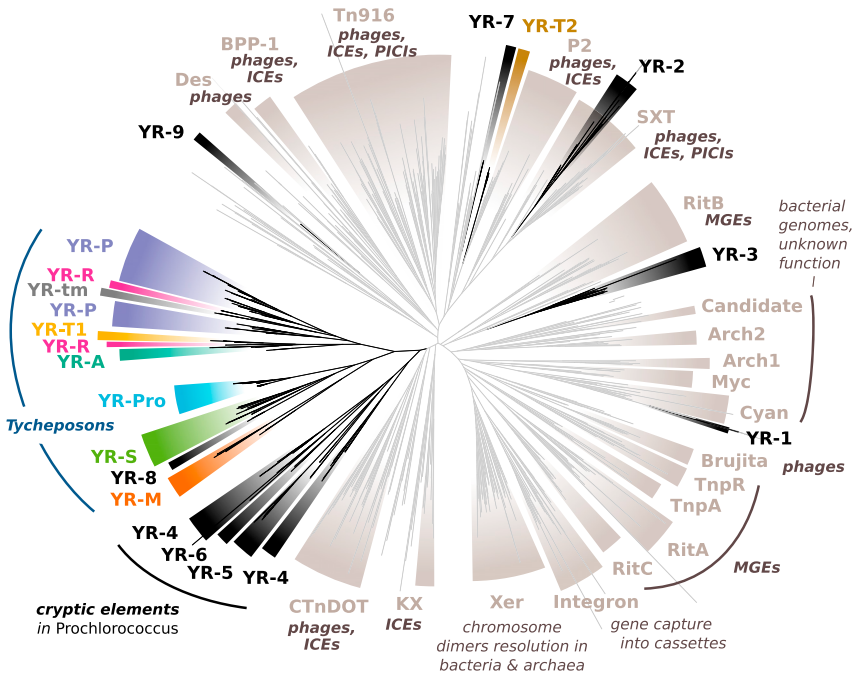


Figure 3. Phylogenetic placement of tycheposon tyrosine recombinases

Maximum likelihood phylogeny of a comprehensive, representative set of Xer-like tyrosine recombinases from bacterial, archaeal, and phage genomes as well as those identified in *Prochlorococcus* genomes in this study. Common types (gray) are labeled according to the classification introduced by Smyshlyaev et al.[35] Integrases found in phages, ICEs, or other MGEs and integrases associated with other functions are indicated by navy blue labels. Integrases from known PICIs (phage-inducible chromosomal islands[36–38]) match two known types: Tn916 and SXT. Most *Prochlorococcus* tRNA-associated (colored) and non-tRNA-associated (black) tyrosine recombinases form a monophyletic group, together with a *Prochlorococcus* core gene integrase (light blue) and are distinct from the previously described common types. 474 of the 501 tycheposons described here carry one of the tRNA-associated tyrosine recombinases.

See also Table S3.

capability after casposons, which encode a protein-primed type B DNA polymerase.[47,48]

Site-specific integrases and autonomous replication are not uncommon in MGEs, but they are typically associated with either eukaryotic elements[49] or large prokaryotic elements with distinctive lifestyles: lysogenic phages, which encode their own capsids; plasmids, which typically exist as extrachromosomal DNA molecules; and ICEs, which encode conjugation machinery that enable them to move between cells.[50,51] Tycheposons lack these complex functional traits and, based on their hallmark functions and overall genomic organization resemble viral satellites[52]—parasitic MGEs that depend on a helper phage for replication/mobilization, such as PICIs[36,53,54] (first described as SaPIs[55]) and PICI-like elements (PLEs).[56] Although phage satellites have recently been reported to be present in a considerable proportion of marine virus particles, the *Prochlorococcus* tycheposons reported here appear to be distinct from those that are found as concatemers in marine virus particles.[57]

The resemblance is particularly noticeable for a subset of "satellite tycheposons" (82 out of 501) which, despite not sharing genes with recognizable homology with PICIs, carry the same functional types of apparent phage-hijacking genes: either a predicted major capsid protein, a building block of a phage capsid, or a small terminase subunit, a part of the protein complex that pumps phage DNA into newly formed phage capsids (Figure 1B). Given this analogy in predicted structure and function (and supported by their prevalence in viral metagenomes—see below), we expect these tycheposons also to be viral satellites, exhibiting a lifestyle similar to that of PICIs—i.e., they reactivate from the host genome during phage infection, replicate, potentially interfere with the replication of the phage, and hijack newly produced phage capsids in order to spread to other cells, ultimately providing population-level resistance against phage predation. Interestingly, however, we never observed satellite tycheposons with additional cargo genes, while this appears to be common in PICIs.[36]

The majority of tycheposons in our dataset (419 out of 501) lack genes for viral packaging or interference and instead carry cargo of often up to 10 kbp (out of 213 tycheposons with clear boundaries: 80% carrying >2 kb of cargo, 42% carrying >10 kb of cargo) that appears to reflect abiotic selection pressures of the local environment (Figure 1A). In many cases, the cargo of these tycheposons is composed of genes of unknown function—a standing challenge for the annotation of wild microbes.[58] The genes we could annotate, however, revealed a broad spectrum of functions associated with ecologically relevant processes—in particular, nutrient acquisition. Remarkably, those functions include the three key nutrients that limit primary productivity in the global oceans: nitrogen, phosphorus, and iron[59] (Figure 1A). For example, we found complete multigene assimilation pathways for nitrate, phosphate, phosphite, and siderophore-mediated iron uptake—complex metabolic traits often encoded in the syntenic regions of the core genome of other bacteria and basal *Prochlorococcus* clades[31,60]—captured in tycheposons, turning them into mobile units with the potential to boost the fitness of streamlined *Prochlorococcus* genomes that had lost these traits in environments where a particular nutrient is a strong selective agent. These findings underpin the possible role of tycheposons in accelerating microbial adaptation and genome evolution, especially with respect to key metabolic traits, which are typically thought to be more rarely gained and lost.

To determine whether tycheposons were specific to *Prochlorococcus* or might be a feature of genomically streamlined ocean microbes, we examined over 2,000 of other marine bacterial genomes for similar elements. This search revealed related elements with genes falling within the tycheposon-specific

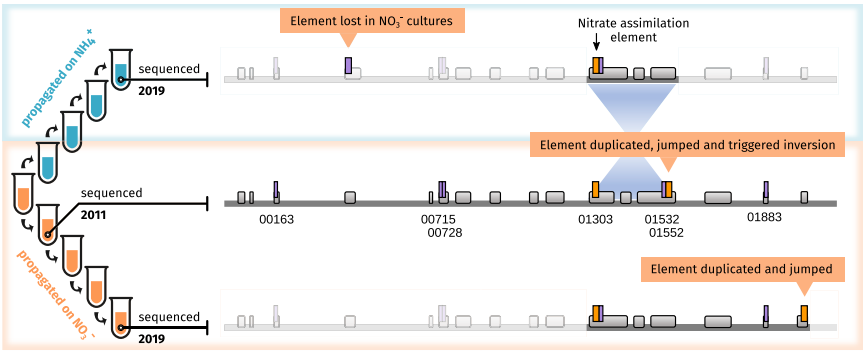


Figure 4. *Prochlorococcus* MIT0604 genome and genomic islands remodeling induced by its integrated tycheposons

Comparison of 3 *Prochlorococcus* MIT0604 genomes from two liquid cultures maintained through serial transfers for 10 years: MIT0604 (NO_3^-) was kept growing on nitrate as the sole nitrogen source while MIT0604 (NH_4^+) was transferred onto and maintained on ammonia as the sole nitrogen source. The sketch on the left gives a simplified representation of the sequence of propagation of the two lineages and when the genomes were sequenced. The reference MIT0604 genome sequenced in 2011 carries seven integrated tycheposons, including two identical copies of a tycheposon containing the nitrate-assimilation gene cluster (Figure 2) in two different genomic islands. The dark gray line represents the reference genome (1.78 Mbp) and regions that changed in the other lineages relative to the reference; faded regions remained unchanged. Gray boxes represent genomic islands, purple boxes represent different tycheposons and the orange boxes the tycheposons carrying the nitrate assimilation cluster (see details in Figure S4C). The genomes sequenced in 2019 furthermore showed that identical copies of a tycheposon in a single genome can trigger chromosomal inversions probably through homologous recombination[62] thereby revealing yet another mechanism by which these elements can promote genomic plasticity—in this case even at the chromosomal scale. Comparing the cultures maintained on different N sources, we observe that the duplication and movement of the nitrate-assimilation tycheposon happened in both lineages maintained on nitrate. Each jump occurred in two distinct genomic island locations, inserting next to a short segment of the same tRNA abutting the original copy, supporting our model of site-specific recombination.

gene-sharing network in a wide variety of bacterial groups, including Alpha-, Gamma- and Deltaproteobacteria (Figure 5C), where they potentially play similar roles in promoting genomic plasticity and adaptability.

Evidence of activity and mobility

Based on genomic data, tycheposons and cryptic elements also appeared highly transient; i.e., they were gained and lost—presumably through integration and excision—within time frames exceeding the resolution of our genome collection: 62% of MGEs were unique and therefore only present in a single genome. About 26% were present in two to five (and in rare cases up to 25) closely related genomes likely representing a single integration event in a common ancestor and subsequent vertical inheritance; 12% were present in at least two distantly related genomes, indicating independent transfer events. This high diversity and patchy distribution suggest a dynamic system with a large pool of elements that transiently visit the islands, thereby greatly contributing to intrapopulation heterogeneity (Figure S3A). (Elements were defined as the same if they shared at least 90% identity over 50% of the shorter element; we did not factor in location because independent integrations will occur at the same tRNA due to the site-specificity of the integrase.)

To get a more detailed picture of the activity of tycheposons, we turned to cultures of *Prochlorococcus* strain MIT0604, which is particularly interesting as it contains 7 MGEs (4 tycheposons and 3 cryptic elements). Two of these tycheposons share 100% sequence identity and encode the complete gene cluster for the assimilation of nitrate[31,60] (Figure S4C), a nutrient that often limits primary productivity in ocean ecosystems.[61] We exploited the fact that we had maintained this strain through serial transfer in two separate liquid cultures for 10 years: one propagated on media with NH_4^+ as the only N source, the other on the same media but with NO_3^- as the only N source. The latter culture was sequenced shortly after separation in 2011, and then both independent cultures were sequenced again in 2019. Comparison of the three genomes revealed significant genomic rearrangements, all of which centered around the tycheposons: we observed gain, loss, and duplication of tycheposons, as well as an associated chromosomal inversion between two identical copies of the nitrate-assimilation cluster they carry (Figure 4). Because the tycheposon attachment sites reside within genomic islands, the modifications were confined to islands, while the rest of the genome remained unaltered. Importantly, the rearrangements also point to the selective advantage provided by tycheposons carrying cargo with adaptive metabolic functions: the most conspicuous changes were linked to the tycheposon containing the nitrate-assimilation gene cluster when cultures were maintained with NO_3^- as the sole N source. That is, the cargo metabolic function—nitrate assimilation—has become an independent "plug-in" cassette that can be duplicated under selective pressure (NO_3^-, in this case), lost when useless (if NH_4^+ is the only N source), and in the wild, likely also more flexibly transferred between cells than would be a core genome trait.

We also looked for evidence of tycheposon and cryptic element mobility in four additional tycheposon-containing *Prochlorococcus* strains, based on PCR amplification of the circular and excised element intermediates. We found that all cultures were internally heterogeneous, with subpopulations of cells missing elements that excised or possessed tandem repeat junctions (Figures S4A and S4B) indicating that stochastic excision and integration events do occur.

To better understand what conditions would trigger tycheposon mobility, we measured the transcriptional response of an integrase associated with a cargo-carrying tycheposon in MIT0604 to a wide range of stressors (Figure S5A)—some of which a cell might experience in the environment and others artificial. The only significant relative increase in transcripts occurred in response to mitomycin C, a DNA-alkylating agent commonly used to trigger DNA damage, and well known to induce prophage excision. The mitomycin C treatment also induced other integrases in other strains (Figure S5B) and triggered detectable mobilization of some tycheposons and cryptic elements

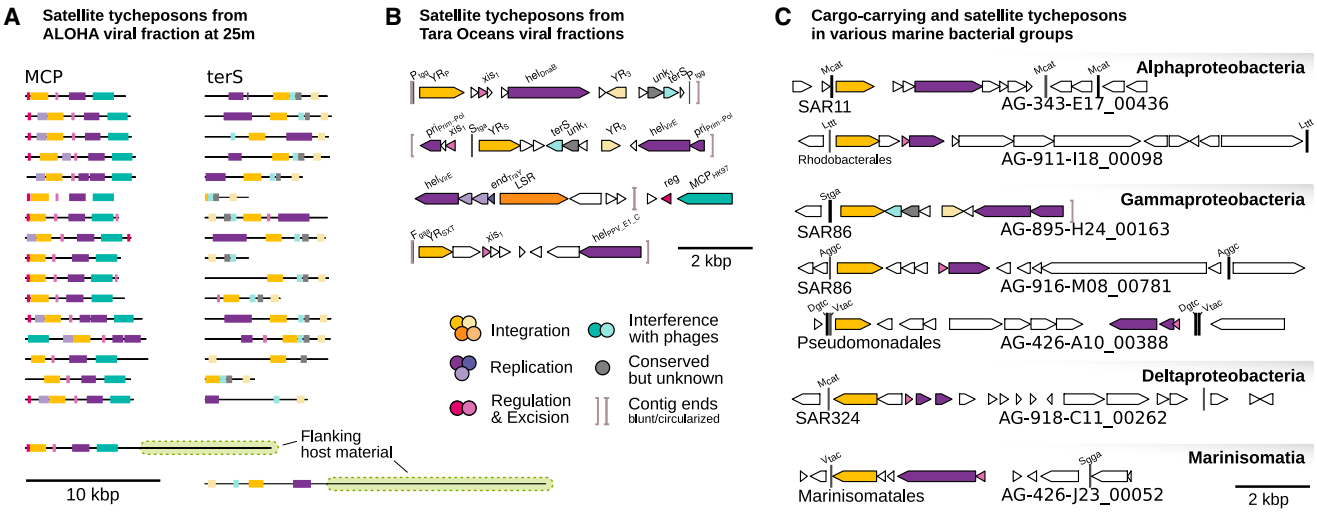


Figure 5. Tycheposons in viral particles from the environment and genomes of various marine bacteria

(A) Examples of nanopore reads from viral-fraction metagenomes from Station ALOHA appearing to be full-length tycheposons. Two reads also carry additional flanking material, likely representing imprecise excision events and hinting at the ability of the tycheposons to promote the transfer of adjacent host material.

(B) Satellite tycheposons in viral-fraction metagenomes from different Tara Ocean stations.

(C) Examples of satellite and cargo-carrying tycheposons in a variety of other marine bacterial groups.

Gene labels: tRNA genes and snippets are labeled with single-letter amino acid code and their anticodon, e.g., S_{tga}, tRNA-serin$_{TGA}$; YR, tyrosine recombinase; LSR, large serine recombinase; MCP, major capsid protein; terS, terminase small subunit; xis, excisionase; hel, helicase; top, topoisomerase; lig, ligase; reg, transcriptional regulator; pri, primase or primase/polymerase or primase/helicase; end, endonuclease; SSR, small serine recombinase; unk, conserved unknown. Some annotations are further labeled with their specific profile (e.g., hel$_{DnaB}$ or hel$_{VirE}$).

See also Table S3.

(Figure S4B), showing that the effect is not specific to the MIT0604 strain.

Further, a genome-wide transcriptome analysis of MIT0604 (Figure S5C) revealed that most integrases, putative excisionases, and replication genes all had elevated transcripts when subjected to mitomycin C, indicating a universal regulatory mechanism. We note that other DNA damaging treatments (Figure S5A), including UV shock, which is a known source of DNA damage in the surface ocean,[63–66] did not significantly increase relative transcript abundance of the integrase (Figure S5C). This suggests something distinctive about the mitomycin C inhibitory mechanism, which we suspect lies in its ability to cause lethal DNA crosslinks[67,68] that lead to replication fork arrests. Thus we postulate that this tight regulatory mechanism would ensure that mobility genes remain mostly silent, avoiding the toxicity generally linked to their activity[50,69] and are only induced in a fraction of the population,[70] even in times of stress such as nutrient starvation or high UV exposure.

Tycheposons in viral capsids and extracellular vesicles

Next, we asked how tycheposons might move from cell to cell in the dilute oceans. We first looked at viral particles as potential vectors by screening viral-fraction metagenomic libraries from ocean samples where *Prochlorococcus* is abundant. Previous work leveraging single-molecule nanopore sequencing provided evidence for PLEs packaged as concatemers in virus-like particles in the open ocean.[71] Further examination of both viral-fraction long nanopore reads[71] and short-read contigs[72,73] also revealed an abundance of satellite tycheposons, supporting our hypothesis that their phage-like packaging genes enable them to hijack phage capsids, promoting their dissemination and leading to the observed prevalence in marine viral metagenomes (Figures 5A and 5B). This postulate is further supported by most recent work showing that viral satellites including tycheposons are abundant within virus particles from marine plankton of the photic zone.[57]

We could not, however, identify cargo-carrying tycheposons, which lack phage-packaging genes, in the viral fraction samples, suggesting alternative transfer routes and motivating us to look elsewhere. We wondered if extracellular vesicles, known to contain DNA and to be released by *Prochlorococcus* and other marine microbes,[74] might serve as vectors. Comparing vesicle- and cellular-fraction metagenomic data from seawater, we observed a specific enrichment of almost all predicted tycheposon hallmark genes in vesicles compared with cells (Figure 6). Similarly, most segments of tRNAs acting as tycheposon attachment sites were also significantly enriched over other tRNA segments in vesicles. Unfortunately, we could not ascertain whether tycheposons in vesicles include cargo-carrying elements, because for the vesicle fraction we lack long-read data that can capture complete elements and our short-read data were too diverse and shallow to assemble into contigs. Nevertheless, the specific enrichments of both of these features strongly suggest that vesicles serve as a common means of dispersal for these MGEs, thereby casting new light on the importance of vesicles as vectors for horizontal gene transfer in microbial communities.

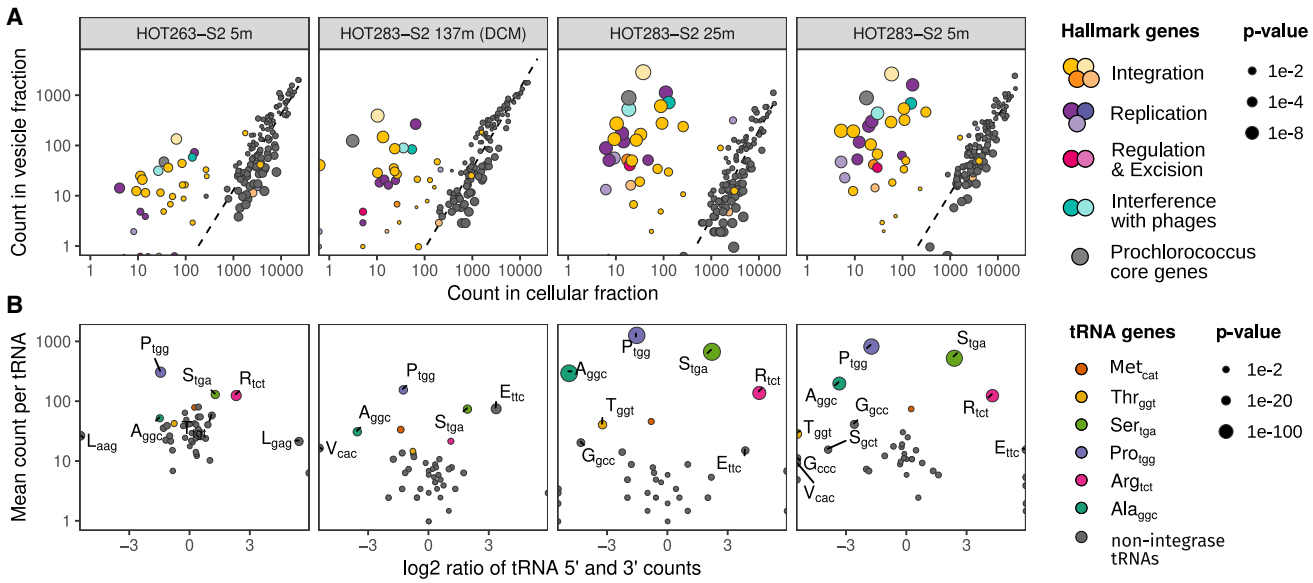


Figure 6. Differential abundance of tycheposon signatures in vesicle-fraction metagenomes from the ocean

(A) Read counts obtained for the same gene profiles from vesicle- and cellular-fraction metagenomes from the North Pacific Subtropical Gyre (dark gray, *Prochlorococcus* core gene profiles; colored, tycheposon hallmark genes; plot labels, Hawai'i Ocean Time-series cruise, station, and depth the samples were collected from, DCM, deep chlorophyll maximum). Larger points correspond to a higher degree of deviation (lower p value in edgeR differential abundance analysis) from the expected abundance ratio, approximately indicated by a linear fit to the core gene counts (dotted line).

(B) Abundances of tRNA 5′ and 3′ segments in vesicle-fraction metagenomes. Deviation from a 1:1 ratio on the x axis indicates an over- or underrepresentation of the respective ends of the tRNA sequence in vesicles, suggesting in the most extreme cases an excess of partial, integrase-targeted tRNAs of more than 10-fold (A_{gcc}, R_{tct}). Higher mean counts for both ends on the y axis indicate higher overall abundance. Larger points correspond to a higher degree of deviation scaled according to p values after Bonferroni correction.

See also Table S3.

Tycheposons as accelerators of genomic island formation

Finally, we investigated the broader impact of tycheposons on *Prochlorococcus*' genome organization and evolution. The clear co-localization of tycheposons and genomic islands—right next to the same seven tRNA genes—prompted us to further explore the underlying mechanisms of genomic island formation and possible connections to tycheposons integration (Figure 7).

One simple explanation would have been that islands in their entirety were made up of tycheposons (and possibly other MGEs). Such *additive islands* form through the accumulation of consecutively integrating MGEs at the same integration site.[75] Indeed, we found a few cases of multiple tycheposons located right next to each other, separated only by a partial copy of the island-tRNA (for example, MIT9202 and AG-355-A09 in Figure S6A). In addition, we also identified remnants of degenerated MGEs, such as fragmented integrases and more partial tRNAs scattered throughout the islands, suggestive of additional, older integration events.

However, with an average of 8–10 islands but only 1–2 MGEs per genome (tycheposons plus cryptic elements), most individual genomic islands did not contain any MGEs, and those that did were often much larger than the MGEs they contained (Figure S2). Overall, we found that putatively active MGEs (with clear boundaries and obvious degeneration) only made up 3.6% of the total island material. Partial and degenerated elements, which lack clear boundaries and are harder to delineate from surrounding material, appear to account for a similar fraction, based on similar numbers of hallmark genes present in both element and nonelement parts of islands. Thus, the major fraction of the islands appears not to be made up of tycheposons or cryptic elements but rather of additional transferred material.

We therefore wondered how this additional material was brought in and, in particular, why it appears to accumulate next to the tycheposon integration sites. Genomic plasticity, especially in the absence of MGEs, is typically associated with the exchange of material through homologous recombination. Homologous recombination can maintain presence/absence patterns in islands and can even play an important role in purging parasitic MGEs from genomes.[76] It can also lead to the emergence of variable genomic islands at specific locations.[77,78] In these cases, island-flanking core genes act as recombination anchors showing clearly elevated recombination rates. Looking for such patterns in *Prochlorococcus*, however, did not reveal elevated recombination rates in island-flanking regions (Figure S7). Thus, while homologous recombination likely plays a role in maintaining within-island variability, it does not seem to provide a good explanation as to why islands are forming specifically next to island tRNA genes.

Instead, we postulate that the additional island material is also acquired through site-specific recombination carried out by transiently visiting tycheposons. Specifically, we hypothesize that the material is brought in as flanking material temporarily captured onto a tycheposon by imprecise excision from a donor

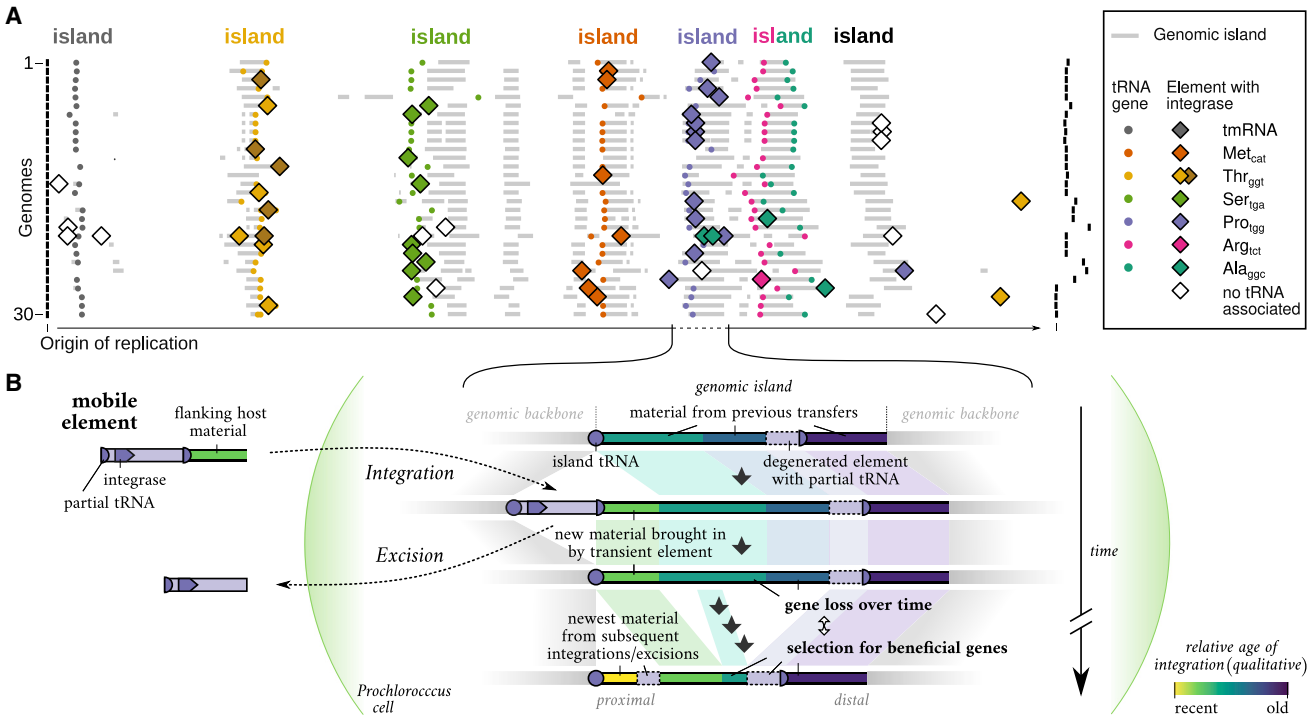


Figure 7. Chromosomal organization of genomic islands and associated mobile genetic elements in *Prochlorococcus*

(A) 30 selected circular *Prochlorococcus* genomes shown relative to their origin of replication (left- to rightmost vertical black bar). Vertical column-like features indicate the predicted genomic islands in conserved locations across the genomes (gray bars). Most islands are associated with one or two specific full-length tRNA genes (colored points, header color). These tRNAs are targeted by mobile genetic elements carrying integrases specific to the different islands (colored diamonds). The genomes shown here are a representative subset of a single *Prochlorococcus* clade (HLII) and are among the most complete genomes in the dataset. The genomes are ordered according to their phylogenetic relationships; i.e., the most closely related genomes are plotted next to each other. See also Figure S2.

(B) A model for genomic island formation promoted by mobile element activity. Mobile elements integrate and excise at the tRNA gene at the proximal end of the island. Genetic material brought in but not excised later, such as flanking DNA from other hosts and degenerated elements, accumulates next to the tRNA leading to gene gain and island growth. Gene gain is countered by gene loss and selection, preserving only beneficial acquisitions which may, in turn, become fixed in descending lineages. Due to the directionality of the process, the observed intrastrain heterogeneity is highest at the proximal end of the island right next to the tRNA and decreases toward the distal end of the island.

See also Figure S2 and Table S2.

genome, comparable with specialized transduction in lysogenic phages.[17,79] After integration into the recipient genome, the subsequent precise excision of the tycheposon would leave the transferred flanking material behind in its new host. Over time, such a process would lead to the accretion of material on one side of tycheposon-targeted tRNAs, and thus the formation of islands (Figure 7B).

Support for this model of island formation comes from three distinct observations: (1) we did find full-length tycheposons in viral metagenomes carrying sequences that appear to be adjacent host material (Figure 5A); (2) most islands exhibit a distinct polarity with content diversity being highest at the proximal end next to the tRNA where we expect new material to be added (Figure S6A); and (3) we observe a clear correlation between the size of islands (excluding MGEs) and the prevalence of tycheposons in these islands—consistent with the prediction that the more often an island is frequented by tycheposons, the more often flanking material can be mobilized via specialized transduction, leading to an increase in island size (Figure S6B).

As a consequence, *Prochlorococcus* tRNA-associated islands appear to comprise two gene pools with different transfer and turnover rates: element genes—which include hallmark genes for integration and replication as well as cargo genes—and island genes that are not part of a functional element. Element genes are highly mobile and can be gained and lost dynamically to match selective pressures quickly. Island genes apparently still have higher rates of transfer than core genes but appear to rely on more stochastic processes such as the co-transfer of element-flanking material or recombination with active elements. Moreover, a strong overlap in enriched functions between elements and adjacent island regions (Figure S3B) supports the idea that cargo carried on tycheposons can become part of the nonelement parts of islands, for example, due to degradation of the carrier element or captured onto elements via recombination processes.

Moreover, because the integrases determine which island a tycheposon will integrate into, the pool of tycheposons is also partitioned into subpopulations that affect different islands.

This mechanism has the potential to control which tycheposons and flanking genomic regions are more likely to recombine. It has previously been shown that different genes of similar ecological functions appear to occur in the same islands across different strains.[1,3] The island specificity of tycheposons could be one important factor for promoting this differentiation of islands with respect to broader ecological themes.

In summary, we conclude that the majority of material in *Prochlorococcus* islands is not derived from tycheposons directly but rather appears to have been brought in by them as flanking material, suggesting that this type of horizontal transfer is crucial to the island-formation process. While genomic data is not direct evidence for tycheposon-driven island formation via transduction, the observed island features—the sharp island delineation at tRNA genes, accretion of horizontally acquired material at the distal end of the islands (the end furthest from the island tRNA) and the increase of island size with increased MGE activity—provide strong support. In this model, the tRNAs function as landing sites and the transient MGEs as vectors, gradually contributing to the accumulation of transferred material and, thus, the formation of large, persistent islands.

DISCUSSION

Given that their streamlined genomes generally lack canonical modes of horizontal gene transfer, the mechanisms generating the diverse pangenomes of oligotrophic marine bacteria have been elusive. Using genomic, experimental, and field data, we present evidence that a unique set of MGEs—which we named tycheposons—is involved in creating this diversity. Tycheposons can carry functional modules that appear to be important for niche differentiation among subpopulations of cells, such as genes involved in the uptake of nitrogen, phosphorus, and iron—the three key nutrients limiting primary productivity in the global ocean[59] and important agents of natural selection over this vast ecosystem.[80]

A subset of the tycheposons appear to be viral satellites similar to PICIs[36,37] and are the first of their kind described in cyanobacteria. The different evolutionary histories of tycheposon and PICI genes and, in particular, their integrases suggest that their structural similarity could have arisen from convergent evolution. It is unclear whether satellite tycheposons were initially introduced in a *Prochlorococcus* ancestor and diverged into the larger family of cargo-carrying tycheposons, or if cargo-carrying tycheposons acquired phage-interference genes. Regardless, all tycheposons are interconnected by a shared set of integrases and replication genes. That the two broad functional categories of tycheposons involve both relief from growth limitation and mortality defense is elegant in its ecological simplicity.

The role of tycheposons appears to go beyond just the function of their cargo and places them at the center of a system for the formation and remodeling of *Prochlorococcus'* largest reservoir of variability—its genomic islands. Genes present in and near genomic islands suggest that the horizontally acquired regions that are not part of tycheposons were brought in with them. Moreover, tycheposons appear to facilitate rearrangements inside islands, promoting the genomic diversity that is characteristic of the population structure of *Prochlorococcus* cells in the wild, where hundreds to thousands of subpopulations varying by small gene cassettes co-exist.[5,6] Similar processes have been reported for *Prochlorococcus*' sister taxon, marine *Synechoccoccus*, in which tycheposons appear to be involved in the diversification of pigment types.[81] This potential to accelerate their hosts' differentiation and adaptation implicates tycheposons as important contributors to the ocean-wide stability of microbial populations such as the *Prochlorococcus* collective.[7]

Our metagenomic analysis of wild populations further shows that tycheposons are contained within both phage capsids and membrane vesicles, implicating these structures in the movement of tycheposons among cells. As particles that can diffuse through seawater, vesicles and phages are well suited to transport and deliver genetic information between microbial cells. Vesicle-mediated horizontal gene transfer—i.e., vesiduction[82]—occurs in diverse microbial systems[83–89] and may mediate exchanges between more distantly related taxa than do viruses,[90] which frequently exhibit narrow host ranges.[91] Given the abundance of vesicles in the oceans[74] and our evidence that they are enriched with tycheposons, we propose that vesicles are important vectors for the dispersal of these MGEs in marine ecosystems.

Tycheposons have left their signatures all over the genomes of marine microbes, implicating themselves as important agents of microbial diversification and adaptation. Though much remains to be learned about tycheposon mobilization, replication, and transfer, their potential for shaping the genetic structure of marine microbial populations opens an exciting new area of inquiry, shedding new light on the processes that govern evolution across the vast oligotrophic oceans and beyond.

Limitations of the study

Many of our findings are based on the analysis of big genomics data. These data provide an unprecedented window into the genomes of marine microbes and allow us to unveil large-scale patterns and correlations across a vast system. However, they do not provide insights into the underlying dynamics and mechanisms. While we have learned much about the activity and mobility of tycheposons from the stress- and lab-evolution experiment, other aspects remain elusive: we do not yet know how satellite tycheposons respond to infection with phages. Currently, we lack cultured phages able to infect *Prochlorococcus* cultures carrying satellite tycheposons. Similarly, while we do observe tycheposons in vesicles, we still lack knowledge about how they might be packaged into these particles and transferred into other cells. Finally, while the presence of key nutrient-uptake genes implies importance for adaptation and fitness, we have yet to confirm such effects through direct observations, for example, through chemostat experiments.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2022.12.006.

ACKNOWLEDGMENTS

This study was supported in part by the Simons Foundation (Life Sciences Project Award IDs 337262, 647135, 736564 to S.W.C., SCOPE award ID 329108 and 721246 to E.F.D. and S.W.C., SCOPE ALOHA award ID 721223 to E.F.D., Life Sciences Project Award IDs 827839 and 510023 to R.S.), Gordon and Betty Moore Foundation (award IDs 3777 to E.F.D. and 3790 to M.B.S.), the Department of Energy (248445 to M.B.S.), and the National Science Foundation (DBI 0424599, OCE-1153588, OCE-1356460 and IOS-1645061 to S.W.C., OCE-1829831 to M.B.S., and OIA-1826734 to R.S.). We thank Wei Ding for support with the panX pangenome software, Gera Smyshlyaev for access to the HMM used to classify known tyrosine-recombinases, Kathryn Kauffman for feedback on the analysis of viral satellites, and Kristina Haslinger, Daniel S. Fisher, and Jed Fuhrman for comments on the manuscript. We also thank Jamie Becker, Jessie Berta-Thompson, and Elena Kazamia for assistance with vesicle metagenome sampling.

AUTHOR CONTRIBUTIONS

T.H., R.L., and M.J.A. conceived the study. T.H. led the bioinformatics analyses. T.H., M.J.A., and E.T. carried out the computational analyses with contributions from S.L.H., P.B., and G.E.L. R.L. led the experimental work on tycheposon mobility and activity. R.L., C.B., and Z.C. carried out those experiments. S.J.B. led the vesicle-related experiments and carried them out together with K.D.D. and A.A.A. E.L., J.B., J.M.E., and E.F.D. developed a new method for single molecule nanopore sequencing of virus-like particles, and collected, sequenced, and assembled Station ALOHA time-series short-read picoplankton and virioplankton metagenomic contigs. A.A.Z. and M.B.S. generated and provided the viral-fraction metagenomic contigs from Tara Oceans. R.S. provided the Global Ocean Reference Genomes database. T.H., R.L., M.J.A., and S.W.C. wrote the manuscript with contributions from all co-authors. T.H. and S.W.C. supervised the project.

DECLARATION OF INTERESTS

J.B. is an employee of Oxford Nanopore Technologies and is a shareholder and/or share option holder.

Received: March 29, 2021
Revised: September 16, 2022
Accepted: December 5, 2022
Published: January 5, 2023

REFERENCES

1. Coleman, M.L., Sullivan, M.B., Martiny, A.C., Steglich, C., Barry, K., Delong, E.F., and Chisholm, S.W. (2006). Genomic islands and the ecology and evolution of Prochlorococcus. Science *311*, 1768–1770. https://doi.org/10.1126/science.1122050.
2. Rodriguez-Valera, F., Martin-Cuadrado, A.B., Rodriguez-Brito, B., Pašić, L., Thingstad, T.F., Rohwer, F., and Mira, A. (2009). Explaining microbial population genomics through phage predation. Nat. Rev. Microbiol. *7*, 828–836. https://doi.org/10.1038/nrmicro2235.
3. Avrani, S., Wurtzel, O., Sharon, I., Sorek, R., and Lindell, D. (2011). Genomic island variability facilitates Prochlorococcus-virus coexistence. Nature *474*, 604–608. https://doi.org/10.1038/nature10172.
4. Delmont, T.O., and Eren, A.M. (2018). Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. PeerJ *6*, e4320. https://doi.org/10.7717/peerj.4320.
5. Kashtan, N., Roggensack, S.E., Rodrigue, S., Thompson, J.W., Biller, S.J., Coe, A., Ding, H., Marttinen, P., Malmstrom, R.R., Stocker, R., et al. (2014). Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science *344*, 416–420. https://doi.org/10.1126/science.1248575.
6. Kashtan, N., Roggensack, S.E., Berta-Thompson, J.W., Grinberg, M., Stepanauskas, R., and Chisholm, S.W. (2017). Fundamental differences in diversity and genomic population structure between Atlantic and Pacific Prochlorococcus. ISME J. *11*, 1997–2011. https://doi.org/10.1038/ismej.2017.64.
7. Biller, S.J., Berube, P.M., Lindell, D., and Chisholm, S.W. (2015). Prochlorococcus: the structure and function of collective diversity. Nat. Rev. Microbiol. *13*, 13–27. https://doi.org/10.1038/nrmicro3378.
8. Ahlgren, N.A., Perelman, J.N., Yeh, Y.C., and Fuhrman, J.A. (2019). Multi-year dynamics of fine-scale marine cyanobacterial populations are more strongly explained by phage interactions than abiotic, bottom-up factors. Environ. Microbiol. *21*, 2948–2963. https://doi.org/10.1111/1462-2920.14687.
9. Ribalet, F., Swalwell, J., Clayton, S., Jiménez, V., Sudek, S., Lin, Y., Johnson, Z.I., Worden, A.Z., and Armbrust, E.V. (2015). Light-driven synchrony ofProchlorococcusgrowth and mortality in the subtropical Pacific gyre. Proc. Natl. Acad. Sci. USA *112*, 8008–8012. https://doi.org/10.1073/pnas.1424279112.
10. García-García, N., Tamames, J., Linz, A.M., Pedrós-Alió, C., and Puente-Sánchez, F. (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. ISME J. *13*, 1–15. https://doi.org/10.1038/s41396-019-0487-8.
11. Guglielmini, J., de la Cruz, F., and Rocha, E.P.C. (2013). Evolution of conjugation and type IV secretion systems. Mol. Biol. Evol. *30*, 315–331. https://doi.org/10.1093/molbev/mss221.
12. Johnston, C., Martin, B., Fichant, G., Polard, P., and Claverys, J.-P. (2014). Bacterial transformation: distribution, shared mechanisms and divergent control. Nat. Rev. Microbiol. *12*, 181–196. https://doi.org/10.1038/nrmicro3199.
13. Lindell, D., Sullivan, M.B., Johnson, Z.I., Tolonen, A.C., Rohwer, F., and Chisholm, S.W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proc. Natl. Acad. Sci. USA *101*, 11013–11018. https://doi.org/10.1073/pnas.0401526101.
14. Zeidner, G., Bielawski, J.P., Shmoish, M., Scanlan, D.J., Sabehi, G., and Béjà, O. (2005). Potential photosynthesis gene recombination between Prochlorococcus and Synechococcus via viral intermediates. Environ. Microbiol. *7*, 1505–1513. https://doi.org/10.1111/j.1462-2920.2005.00833.x.
15. Sullivan, M.B., Lindell, D., Lee, J.A., Thompson, L.R., Bielawski, J.P., and Chisholm, S.W. (2006). Prevalence and evolution of core photosystem II genes in marine cyanobacterial viruses and their hosts. PLoS Biol. *4*, e234. https://doi.org/10.1371/journal.pbio.0040234.

16. Sullivan, M.B., Coleman, M.L., Weigele, P., Rohwer, F., and Chisholm, S.W. (2005). Three Prochlorococcus Cyanophage genomes: signature features and ecological interpretations. PLoS Biol. *3*, e144. https://doi.org/10.1371/journal.pbio.0030144.

17. Sullivan, M.B., Krastins, B., Hughes, J.L., Kelly, L., Chase, M., Sarracino, D., and Chisholm, S.W. (2009). The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial "mobilome.". Environ. Microbiol. *11*, 2935–2951. https://doi.org/10.1111/j.1462-2920.2009.02081.x.

18. Malmstrom, R.R., Rodrigue, S., Huang, K.H., Kelly, L., Kern, S.E., Thompson, A., Roggensack, S., Berube, P.M., Henn, M.R., and Chisholm, S.W. (2013). Ecology of uncultured Prochlorococcus clades revealed through single-cell genomics and biogeographic analysis. ISME J. *7*, 184–198. https://doi.org/10.1038/ismej.2012.89.

19. Rocap, G., Larimer, F.W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N.A., Arellano, A., Coleman, M., Hauser, L., Hess, W.R., et al. (2003). Genome divergence in two Prochlorococcus ecotypes reflects oceanic niche differentiation. Nature *424*, 1042–1047. https://doi.org/10.1038/nature01947.

20. Biller, S.J., Berube, P.M., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., Awad, L., Roache-Johnson, K.H., Ding, H., Giovannoni, S.J., Rocap, G., et al. (2014). Genomes of diverse isolates of the marine cyanobacterium Prochlorococcus. Sci. Data *1*, 140034. https://doi.org/10.1038/sdata.2014.34.

21. Berube, P.M., Biller, S.J., Hackl, T., Hogle, S.L., Satinsky, B.M., Becker, J.W., Braakman, R., Collins, S.B., Kelly, L., Berta-Thompson, J., et al. (2018). Single cell genomes of Prochlorococcus, Synechococcus, and sympatric microbes from diverse marine environments. Sci. Data *5*, 180154. https://doi.org/10.1038/sdata.2018.154.

22. Bentkowski, P., Van Oosterhout, C., and Mock, T. (2015). A model of genome size evolution for prokaryotes in stable and fluctuating environments. Genome Biol. Evol. *7*, 2344–2351. https://doi.org/10.1093/gbe/evv148.

23. Karcagi, I., Draskovits, G., Umenhoffer, K., Fekete, G., Kovács, K., Méhi, O., Balikó, G., Szappanos, B., Györfy, Z., Fehér, T., et al. (2016). Indispensability of horizontally transferred genes and its impact on bacterial genome streamlining. Mol. Biol. Evol. *33*, 1257–1269. https://doi.org/10.1093/molbev/msw009.

24. Pachiadaki, M.G., Brown, J.M., Brown, J., Bezuidt, O., Berube, P.M., Biller, S.J., Poulton, N.J., Burkart, M.D., La Clair, J.J., Chisholm, S.W., et al. (2019). Charting the complexity of the marine microbiome through single-cell genomics. Cell *179*, 1623–1635.e11. https://doi.org/10.1016/j.cell.2019.11.017.

25. Boyd, E.F., Almagro-Moreno, S., and Parent, M.A. (2009). Genomic islands are dynamic, ancient integrative elements in bacterial evolution. Trends Microbiol. *17*, 47–53. https://doi.org/10.1016/j.tim.2008.11.003.

26. Kettler, G.C., Martiny, A.C., Huang, K., Zucker, J., Coleman, M.L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., et al. (2007). Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. PLoS Genet. *3*, e231. https://doi.org/10.1371/journal.pgen.0030231.

27. Malmstrom, R.R., Coe, A., Kettler, G.C., Martiny, A.C., Frias-Lopez, J., Zinser, E.R., and Chisholm, S.W. (2010). Temporal dynamics of Prochlorococcus ecotypes in the Atlantic and Pacific Oceans. ISME J. *4*, 1252–1264. https://doi.org/10.1038/ismej.2010.60.

28. Liu, H.-L., and Zhu, J. (2010). Analysis of the 3′ ends of tRNA as the cause of insertion sites of foreign DNA in Prochlorococcus. J. Zhejiang Univ. Sci. B *11*, 708–718. https://doi.org/10.1631/jzus.B0900417.

29. Williams, K.P. (2002). Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. Nucleic Acids Res. *30*, 866–875. https://doi.org/10.1093/nar/30.4.866.

30. Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., and Alva, V. (2018). A completely reimplemented MPI bioinformatics toolkit with a new HHpred server at its core. J. Mol. Biol. *430*, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007.

31. Berube, P.M., Biller, S.J., Kent, A.G., Berta-Thompson, J.W., Kelly, L., Roggensack, S.E., et al. (2014). Physiology and evolution of nitrogen acquisition in Prochlorococcus. ISME, 1195–1207. https://doi.org/10.1038/ismej.2014.211.

32. Martiny, A.C., Coleman, M.L., and Chisholm, S.W. (2006). Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. Proc. Natl. Acad. Sci. USA *103*, 12552–12557. https://doi.org/10.1073/pnas.0601301103.

33. Barnett, J.P., Millard, A., Ksibe, A.Z., Scanlan, D.J., Schmid, R., and Blindauer, C.A. (2012). Mining genomes of marine cyanobacteria for elements of zinc homeostasis. Front. Microbiol. *3*, 142. https://doi.org/10.3389/fmicb.2012.00142.

34. Martínez, A., Osburne, M.S., Sharma, A.K., DeLong, E.F., and Chisholm, S.W. (2012). Phosphite utilization by the marine picocyanobacterium Prochlorococcus MIT9301. Environ. Microbiol. *14*, 1363–1377. https://doi.org/10.1111/j.1462-2920.2011.02612.x.

35. Smyshlyaev, G., Barabas, O., and Bateman, A. (2021). Sequence analysis allows functional annotation of tyrosine recombinases in prokaryotic genomes. Mol. Syst. Biol. *17*, e9880. https://doi.org/10.1101/542381.

36. Fillol-Salom, A., Martínez-Rubio, R., Abdulrahman, R.F., Chen, J., Davies, R., and Penadés, J.R. (2018). Phage-inducible chromosomal islands are ubiquitous within the bacterial universe. ISME J. *12*, 2114–2128. https://doi.org/10.1038/s41396-018-0156-3.

37. Martínez-Rubio, R., Quiles-Puchalt, N., Martí, M., Humphrey, S., Ram, G., Smyth, D., Chen, J., Novick, R.P., and Penadés, J.R. (2017). Phage-inducible islands in the Gram-positive cocci. ISME J. *11*, 1029–1042. https://doi.org/10.1038/ismej.2016.163.

38. Hiramatsu, K., Ito, T., Tsubakishita, S., Sasaki, T., Takeuchi, F., Morimoto, Y., Katayama, Y., Matsuo, M., Kuwahara-Arai, K., Hishinuma, T., et al. (2013). Genomic basis for methicillin resistance in Staphylococcus aureus. Infect. Chemother. *45*, 117–136. https://doi.org/10.3947/ic.2013.45.2.117.

39. Iranzo, J., Krupovic, M., and Koonin, E.V. (2016). The double-stranded DNA virosphere as a modular hierarchical network of gene sharing. mBio *7*. e00978–e00916. https://doi.org/10.1128/mBio.00978-16.

40. Krupovic, M., Makarova, K.S., Wolf, Y.I., Medvedeva, S., Prangishvili, D., Forterre, P., and Koonin, E.V. (2019). Integrated mobile genetic elements in Thaumarchaeota. Environ. Microbiol. *21*, 2056–2078. https://doi.org/10.1111/1462-2920.14564.

41. Bellas, C.M., and Sommaruga, R. (2021). Polinton-like viruses are abundant in aquatic ecosystems. Microbiome *9*, 13. https://doi.org/10.1186/s40168-020-00956-0.

42. Brown, C.L., Mullet, J., Hindi, F., Stoll, J.E., Gupta, S., Choi, M., Keenum, I., Vikesland, P., Pruden, A., and Zhang, L. (2021). mobileOG-db: a manually curated database of protein families mediating the life cycle of bacterial mobile genetic elements. Preprint at bioRxiv. https://doi.org/10.1101/2021.08.27.457951.

43. Siguier, P., Gourbeyre, E., and Chandler, M. (2014). Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol. Rev. *38*, 865–891. https://doi.org/10.1111/1574-6976.12067.

44. Grindley, N.D.F., Whiteson, K.L., and Rice, P.A. (2006). Mechanisms of site-specific recombination. Annu. Rev. Biochem. *75*, 567–605. https://doi.org/10.1146/annurev.biochem.73.011303.073908.

45. Smyshlyaev, G., Bateman, A., and Barabas, O. (2021). Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. Mol. Syst. Biol. *17*, e9880. https://doi.org/10.15252/msb.20209880.

46. Rodriguez-Valera, F., Martin-Cuadrado, A.-B., and López-Pérez, M. (2016). Flexible genomic islands as drivers of genome evolution.

Curr. Opin. Microbiol. *31*, 154–160. https://doi.org/10.1016/j.mib.2016.03.014.

47. Krupovic, M., Makarova, K.S., Forterre, P., Prangishvili, D., and Koonin, E.V. (2014). Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol. *12*, 36. https://doi.org/10.1186/1741-7007-12-36.

48. Krupovic, M., Béguin, P., and Koonin, E.V. (2017). Casposons: mobile genetic elements that gave rise to the CRISPR-Cas adaptation machinery. Curr. Opin. Microbiol. *38*, 36–43. https://doi.org/10.1016/j.mib.2017.04.004.

49. Hackl, T., Duponchel, S., Barenhoff, K., and Weinmann, A. (2021). Endogenous virophages populate the genomes of a marine heterotrophic flagellate. Elife *10*, e72674.

50. Johnson, C.M., and Grossman, A.D. (2015). Integrative and conjugative elements (ICEs): what they do and how they work. Annu. Rev. Genet. *49*, 577–601. https://doi.org/10.1146/annurev-genet-112414-055018.

51. Sobecky, P.A., and Hazen, T.H. (2009). Horizontal gene transfer and mobile genetic elements in marine systems. Methods Mol. Biol. *532*, 435–453. https://doi.org/10.1007/978-1-60327-853-9_25.

52. Frígols, B., Quiles-Puchalt, N., Mir-Sanchis, I., Donderis, J., Elena, S.F., Buckling, A., Novick, R.P., Marina, A., and Penadés, J.R. (2015). Virus satellites drive viral evolution and ecology. PLoS Genet. *11*. e1005609. https://doi.org/10.1371/journal.pgen.1005609.

53. Novick, R.P., Christie, G.E., and Penadés, J.R. (2010). The phage-related chromosomal islands of Gram-positive bacteria. Nat. Rev. Microbiol. *8*, 541–551. https://doi.org/10.1038/nrmicro2393.

54. Penadés, J.R., and Christie, G.E. (2015). The phage-inducible chromosomal islands: A family of highly evolved molecular parasites. Annu. Rev. Virol. *2*, 181–201. https://doi.org/10.1146/annurev-virology-031413-085446.

55. Maiques, E., Ubeda, C., Tormo, M.A., Ferrer, M.D., Lasa, I., Novick, R.P., and Penadés, J.R. (2007). Role of staphylococcal phage and SaPI integrase in intra- and interspecies SaPI transfer. J. Bacteriol. *189*, 5608–5616. https://doi.org/10.1128/JB.00619-07.

56. O'Hara, B.J., Barth, Z.K., McKitterick, A.C., and Seed, K.D. (2017). A highly specific phage defense system is a conserved feature of the Vibrio cholerae mobilome. PLoS Genet. *13*. e1006838. https://doi.org/10.1371/journal.pgen.1006838.

57. Eppley, J.M., Biller, S.J., Luo, E., Burger, A., and DeLong, E.F. (. 2022;). Marine viral particles reveal an expansive repertoire of phage-parasitizing mobile elements. Proc. Natl. Acad. Sci. USA *119*. e22127221192022. https://doi.org/10.1073/pnas.2212722119.

58. Lobb, B., Tremblay, B.J.-M., Moreno-Hagelsieb, G., and Doxey, A.C. (2020). An assessment of genome annotation coverage across the bacterial tree of life. Microb. Genom. *6*. e000341. https://doi.org/10.1099/mgen.0.000341.

59. Moore, C.M., Mills, M.M., Arrigo, K.R., Berman-Frank, I., Bopp, L., Boyd, P.W., Galbraith, E.D., Geider, R.J., Guieu, C., Jaccard, S.L., et al. (2013). Processes and patterns of oceanic nutrient limitation. Nat. Geosci. *6*, 701–710. https://doi.org/10.1038/ngeo1765.

60. Berube, P.M., Rasmussen, A., Braakman, R., Stepanauskas, R., and Chisholm, S.W. (2019). Emergence of trait variability through the lens of nitrogen assimilation in Prochlorococcus. eLife *8*, e41043. https://doi.org/10.7554/eLife.41043.

61. Tyrrell, T. (1999). The relative influences of nitrogen and phosphorus on oceanic primary production. Nature *400*, 525–531. https://doi.org/10.1038/22941.

62. Achaz, G., Coissac, E., Netter, P., and Rocha, E.P.C. (2003). Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics *164*, 1279–1289. https://doi.org/10.1093/genetics/164.4.1279.

63. Llabrés, M., and Agustí, S. (2006). Picophytoplankton cell death induced by UV radiation: evidence for oceanic Atlantic communities. Limnol. Oceanogr. *51*, 21–29. https://doi.org/10.4319/lo.2006.51.1.0021.

64. Agustí, S., and Llabrés, M. (2007). Solar radiation-induced mortality of marine pico-phytoplankton in the oligotrophic ocean. Photochem. Photobiol. *83*, 793–801. https://doi.org/10.1111/j.1751-1097.2007.00144.x.

65. Kolowrat, C., Partensky, F., Mella-Flores, D., Le Corguillé, G., Boutte, C., Blot, N., Ratin, M., Ferréol, M., Lecomte, X., Gourvil, P., et al. (2010). Ultraviolet stress delays chromosome replication in light/dark synchronized cells of the marine cyanobacterium Prochlorococcus marinus PCC9511. BMC Microbiol. *10*, 204. https://doi.org/10.1186/1471-2180-10-204.

66. Mella-Flores, D., Six, C., Ratin, M., Partensky, F., Boutte, C., Le Corguillé, G., Marie, D., Blot, N., Gourvil, P., Kolowrat, C., et al. (2012). Prochlorococcus and Synechococcus have Evolved Different Adaptive Mechanisms to Cope with Light and UV Stress. Front. Microbiol. *3*, 285. https://doi.org/10.3389/fmicb.2012.00285.

67. Iyer, V.N., and Szybalski, W. (1964). Mitomycins and porfiromycin: chemical mechanism of activation and cross-linking of DNA. Science *145*, 55–58. https://doi.org/10.1126/science.145.3627.55.

68. Tomasz, M. (1995). Mitomycin C: small, fast and deadly (but very selective). Chem. Biol. *2*, 575–579. https://doi.org/10.1016/1074-5521(95)90120-5.

69. Turner, S.L., Bailey, M.J., Lilley, A.K., and Thomas, C.M. (2002). Ecological and molecular maintenance strategies of mobile genetic elements. FEMS Microbiol. Ecol. *42*, 177–185. https://doi.org/10.1111/j.1574-6941.2002.tb01007.x.

70. Minoia, M., Gaillard, M., Reinhard, F., Stojanov, M., Sentchilo, V., and van der Meer, J.R. (2008). Stochasticity and bistability in horizontal transfer control of a genomic island in Pseudomonas. Proc. Natl. Acad. Sci. USA *105*, 20792–20797. https://doi.org/10.1073/pnas.0806164106.

71. Beaulaurier, J., Luo, E., Eppley, J.M., Uyl, P.D., Dai, X., Burger, A., Turner, D.J., Pendelton, M., Juul, S., Harrington, E., et al. (2020). Assembly-free single-molecule sequencing recovers complete virus genomes from natural microbial communities. Genome Res. *30*, 437–446. https://doi.org/10.1101/gr.251686.119.

72. Gregory, A.C., Zayed, A.A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., et al. (2019). Marine DNA viral macro- and microdiversity from pole to pole. Cell *177*, 1109–1123.e14. https://doi.org/10.1016/j.cell.2019.03.040.

73. Luo, E., Eppley, J.M., Romano, A.E., Mende, D.R., and DeLong, E.F. (2020). Double-stranded DNA virioplankton dynamics and reproductive strategies in the oligotrophic open ocean water column. ISME J. *14*, 1304–1315. https://doi.org/10.1038/s41396-020-0604-8.

74. Biller, S.J., Schubotz, F., Roggensack, S.E., Thompson, A.W., Summons, R.E., and Chisholm, S.W. (2014). Bacterial vesicles in marine ecosystems. Science *343*, 183–186. https://doi.org/10.1126/science.1243457.

75. López-Pérez, M., Gonzaga, A., and Rodriguez-Valera, F. (2013). Genomic diversity of "deep ecotype" Alteromonas macleodii isolates: evidence for pan-Mediterranean clonal frames. Genome Biol. Evol. *5*, 1220–1232. https://doi.org/10.1093/gbe/evt089.

76. Croucher, N.J., Mostowy, R., Wymant, C., Turner, P., Bentley, S.D., and Fraser, C. (2016). Horizontal DNA transfer mechanisms of bacteria as weapons of intragenomic conflict. PLOS Biol. *14*. e1002394. https://doi.org/10.1371/journal.pbio.1002394.

77. López-Pérez, M., Martin-Cuadrado, A.-B., and Rodriguez-Valera, F. (2014). Homologous recombination is involved in the diversity of replacement flexible genomic islands in aquatic prokaryotes. Front. Genet. *5*, 147. https://doi.org/10.3389/fgene.2014.00147.

78. Oliveira, P.H., Touchon, M., Cury, J., and Rocha, E.P.C. (2017). The chromosomal organization of horizontal gene transfer in bacteria. Nat. Commun. *8*, 841. https://doi.org/10.1038/s41467-017-00808-w.

79. Sato, K., and Campbell, A. (1970). Specialized transduction of galactose by lambda phage from a deletion lysogen. Virology *41*, 474–487. https://doi.org/10.1016/0042-6822(70)90169-8.

80. Coleman, M.L., and Chisholm, S.W. (2010). Ecosystem-specific selection pressures revealed through comparative population genomics. Proc. Natl. Acad. Sci. USA *107*, 18634–18639. https://doi.org/10.1073/pnas.1009480107.

81. Grébert, T., Garczarek, L., Daubin, V., Humily, F., Marie, D., Ratin, M., Devailly, A., Farrant, G.K., Mary, I., Mella-Flores, D., et al. (2022). Diversity and evolution of pigment types in marine Synechococcus cyanobacteria. Genome Biol. Evol. *14*, evac035. https://doi.org/10.1093/gbe/evac035.

82. Soler, N., and Forterre, P. (2020). Vesiduction: the fourth way of HGT. Environ. Microbiol. *22*, 2457–2460. https://doi.org/10.1111/1462-2920.15056.

83. Dorward, D.W., Garon, C.F., and Judd, R.C. (1989). Export and intercellular transfer of DNA via membrane blebs of Neisseria gonorrhoeae. J. Bacteriol. *171*, 2499–2505. https://doi.org/10.1128/jb.171.5.2499-2505.1989.

84. Klieve, A.V., Yokoyama, M.T., Forster, R.J., Ouwerkerk, D., Bain, P.A., and Mawhinney, E.L. (2005). Naturally occurring DNA transfer system associated with membrane vesicles in cellulolytic Ruminococcus spp. of ruminal origin. Appl. Environ. Microbiol. *71*, 4248–4253. https://doi.org/10.1128/AEM.71.8.4248-4253.2005.

85. Gaudin, M., Krupovic, M., Marguet, E., Gauliard, E., Cvirkaite-Krupovic, V., Le Cam, E., Oberto, J., and Forterre, P. (2014). Extracellular membrane vesicles harbouring viral genomes. Environ. Microbiol. *16*, 1167–1175. https://doi.org/10.1111/1462-2920.12235.

86. Renelli, M., Matias, V., Lo, R.Y., and Beveridge, T.J. (2004). DNA-containing membrane vesicles of Pseudomonas aeruginosa PAO1 and their genetic transformation potential. Microbiology (Reading) *150*, 2161–2169. https://doi.org/10.1099/mic.0.26841-0.

87. Kolling, G.L., Simon, L., and Matthews, K.R. (2000). Vesicle-mediated transfer of virulence genes from *Escherichia coli* O157: H7 to other enteric bacteria. Appl. Environ. Microbiol. *66*, 4414–4420.

88. Erdmann, S., Tschitschko, B., Zhong, L., Raftery, M.J., and Cavicchioli, R. (2017). A plasmid from an Antarctic haloarchaeon uses specialized membrane vesicles to disseminate and infect plasmid-free cells. Nat. Microbiol. *2*, 1446–1455. https://doi.org/10.1038/s41564-017-0009-2.

89. Tran, F., and Boedicker, J.Q. (2017). Genetic cargo and bacterial species set the rate of vesicle-mediated horizontal gene transfer. Sci. Rep. *7*, 8813. https://doi.org/10.1038/s41598-017-07447-7.

90. Nazarian, P., Tran, F., and Boedicker, J.Q. (2018). Modeling multispecies gene flow dynamics reveals the unique roles of different horizontal gene transfer mechanisms. Front. Microbiol. *9*, 2978. https://doi.org/10.3389/fmicb.2018.02978.

91. Kauffman, K.M., Hussain, F.A., Yang, J., Arevalo, P., Brown, J.M., Chang, W.K., VanInsberghe, D., Elsherbini, J., Sharma, R.S., Cutler, M.B., et al. (2018). A major lineage of non-tailed dsDNA viruses as unrecognized killers of marine bacteria. Nature *554*, 118–122. https://doi.org/10.1038/nature25474.

92. Thompson, A.W., Huang, K., Saito, M.A., and Chisholm, S.W. (2011). Transcriptome response of high- and low-light-adapted Prochlorococcus strains to changing iron availability. ISME J. *5*, 1580–1594. https://doi.org/10.1038/ismej.2011.49.

93. Shimada, A., Kanai, S., and Maruyama, T. (1995). Partial sequence of ribulose-1,5-bisphosphate carboxylase/oxygenase and the phylogeny of Prochloron and Prochlorococcus (Prochlorales). J. Mol. Evol. *40*, 671–677. https://doi.org/10.1007/BF00160516.

94. Penno, S., Campbell, L., and Hess, W.R. (2000). Presence of phycoerythrin in two strains of Prochlorococcus (cyanobacteria) isolated from the subtropical North Pacific Ocean. J. Phycol. *36*, 723–729. https://doi.org/10.1046/j.1529-8817.2000.99203.x.

95. Cubillos-Ruiz, A., Berta-Thompson, J.W., Becker, J.W., van der Donk, W.A., and Chisholm, S.W. (2017). Evolutionary radiation of lanthipeptides in marine cyanobacteria. Proc. Natl. Acad. Sci. USA *114*, E5424–E5433.

96. Biller, S.J., Coe, A., Martin-Cuadrado, A.-B., and Chisholm, S.W. (2015). Draft genome sequence of Alteromonas macleodii Strain MIT1002, isolated from an enrichment culture of the marine cyanobacterium Prochlorococcus. Genome Announc. *3*. e00967–e00915. https://doi.org/10.1128/genomeA.00967-15.

97. Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., and Tyson, G.W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. *25*, 1043–1055. https://doi.org/10.1101/gr.186072.114.

98. Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. Bioinformatics *30*, 2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

99. Ding, W., Baumdicker, F., and Neher, R.A. (2018). panX: pan-genome analysis and exploration. Nucleic Acids Res. *46*, e5. https://doi.org/10.1093/nar/gkx977.

100. Eddy, S.R. (2011). Accelerated profile HMM Searches. PLoS Comput. Biol. *7*. e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

101. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One *11*. e0163962. https://doi.org/10.1371/journal.pone.0163962.

102. Nakamura, T., Yamada, K.D., Tomii, K., and Katoh, K. (2018). Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics *34*, 2490–2492. https://doi.org/10.1093/bioinformatics/bty121.

103. Capella-Gutiérrez, S., Silla-Martínez, J.M., and Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics *25*, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348.

104. Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One *5*, e9490. https://doi.org/10.1371/journal.pone.0009490.

105. Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things): phytools: R package. Methods Ecol. Evol. *3*, 217–223. https://doi.org/10.1111/j.2041-210X.2011.00169.x.

106. Yu, G., Lam, T.T.-Y., Zhu, H., and Guan, Y. (2018). Two methods for mapping and visualizing associated data on phylogeny using ggtree. Mol. Biol. Evol. *35*, 3041–3043. https://doi.org/10.1093/molbev/msy194.

107. Gao, F., and Zhang, C.T. (2008). Ori-Finder: A web-based system for finding oriC s in unannotated bacterial genomes. BMC Bioinformatics *9*, 79. https://doi.org/10.1186/1471-2105-9-79.

108. Tang, H., Zhang, X., Miao, C., Zhang, J., Ming, R., Schnable, J.C., Schnable, P.S., Lyons, E., and Lu, J. (2015). ALLMAPS: robust scaffold ordering based on multiple maps. Genome Biol. *16*, 3. https://doi.org/10.1186/s13059-014-0573-1.

109. R Core Team (2013). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

110. Himmelmann L. HMM: HMM-hidden markov models. R package version 1.0. https://cran.r-project.org/package=HMM. 2016.

111. Steinegger, M., and Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat. Biotechnol. *35*, 1026–1028. https://doi.org/10.1038/nbt.3988.

112. Guo, J., Bolduc, B., Zayed, A.A., Varsani, A., Dominguez-Huerta, G., Delmont, T.O., Pratama, A.A., Gazitúa, M.C., Vik, D., Sullivan, M.B., et al. (2021). VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. Microbiome *9*, 37. https://doi.org/10.1186/s40168-020-00990-y.

113. Bertelli, C., Laird, M.R., Williams, K.P., Simon Fraser University Research Computing Group, Lau, B.Y., Hoad, G., Winsor, G.L., and Brinkman, F.S.L. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Res. *45*, W30–W35. https://doi.org/10.1093/nar/gkx343.

114. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., and Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics *20*, 473. https://doi.org/10.1186/s12859-019-3019-7.

115. Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. Bioinformatics *30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531.

116. Rambaut, A. (2012). FigTree v1. 4. http://tree.bio.ed.ac.uk/software/figtree/.

117. Hyatt, D., Chen, G.-L., Locascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

118. Laslett, D., and Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res. *32*, 11–16. https://doi.org/10.1093/nar/gkh152.

119. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

120. Wickham, H. (2011). ggplot2. WIREs Comp. Stat. *3*, 180–185. https://doi.org/10.1002/wics.147.

121. Pedersen, T.L. (2019). Ggraph: an implementation of grammar of graphics for graphs and networks. https://ggraph.data-imaginist.com/.

122. Letunic, I., and Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. *47*, W256–W259. https://doi.org/10.1093/nar/gkz239.

123. Ruan, J., and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. Nat. Methods *17*, 155–158. https://doi.org/10.1038/s41592-019-0669-3.

124. Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., et al. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics *28*, 1647–1649. https://doi.org/10.1093/bioinformatics/bts199.

125. Darling, A.E., Mau, B., and Perna, N.T. (2010). Progressivemauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One *5*, e11147. https://doi.org/10.1371/journal.pone.0011147.

126. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

127. Nattestad, M., Chin, C.-S., and Schatz, M.C. (2016). Ribbon: visualizing complex genome alignments and structural variation. Preprint at bioRxiv. https://doi.org/10.1101/082123.

128. Chin, C.-S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods *10*, 563–569. https://doi.org/10.1038/nmeth.2474.

129. Chen, I.-M.A., Markowitz, V.M., Chu, K., Palaniappan, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Andersen, E., Huntemann, M., et al. (2017). IMG/M: integrated genome and metagenome comparative data analysis system. Nucleic Acids Res. *45*, D507–D516. https://doi.org/10.1093/nar/gkw929.

130. Markowitz, V.M., Chen, I.-M.A., Palaniappan, K., Chu, K., Szeto, E., Pillay, M., Ratner, A., Huang, J., Woyke, T., Huntemann, M., et al. (2014). IMG of the integrated microbial genomes comparative analysis system. Nucl. Acids Res. *42*, D560–D567. https://doi.org/10.1093/nar/gkt963.

131. Buchfink, B., Xie, C., and Huson, D.H. (2015). Fast and sensitive protein alignment using DIAMOND. Nat. Methods *12*, 59–60. https://doi.org/10.1038/nmeth.3176.

132. Woodcroft, B.J., Boyd, J.A., and Tyson, G.W. (2016). OrfM: a fast open reading frame predictor for metagenomic data. Bioinformatics *32*, 2702–2703. https://doi.org/10.1093/bioinformatics/btw241.

133. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

134. Bushnell, B.J.R. (2014). BBMap short read aligner, and other bioinformatic tools. http://sourceforge.net/projects/bbmap/.

135. Lin, M., and Kussell, E. (2019). Inferring bacterial recombination rates from large-scale sequencing datasets. Nat. Methods *16*, 199–204. https://doi.org/10.1038/s41592-018-0293-7.

136. Alexa A., Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. R Package Version 2. https://bioconductor.org/packages/release/bioc/html/topGO.html. 2010.

137. Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., and Zhang, B. (2019). WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Res. *47*, W199–W205. https://doi.org/10.1093/nar/gkz401.

138. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

139. Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

140. Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550. https://doi.org/10.1186/s13059-014-0550-8.

141. Moore, L.R., Coe, A., Zinser, E.R., Saito, M.A., Sullivan, M.B., Lindell, D., Frois-Moniz, K., Waterbury, J., and Chisholm, S.W. (2007). Culturing the marine cyanobacterium Prochlorococcus. Limnol. Oceanogr. Methods *5*, 353–362. https://doi.org/10.4319/lom.2007.5.353.

142. Biller, S.J., Berube, P.M., Dooley, K., Williams, M., Satinsky, B.M., Hackl, T., Hogle, S.L., Coe, A., Bergauer, K., Bouman, H.A., et al. (2018). Marine microbial metagenomes sampled across space and time. Sci. Data *5*, 180176. https://doi.org/10.1038/sdata.2018.176.

143. Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., and White, O. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. *29*, 41–43. https://doi.org/10.1093/nar/29.1.41.

144. Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., and Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat. Biotechnol. *36*, 996–1004. https://doi.org/10.1038/nbt.4229.

145. Gao, F., Luo, H., and Zhang, C.T. (2013). DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. Nucleic Acids Res. *41*, D90–D93. https://doi.org/10.1093/nar/gks990.

146. Dufresne, A., Ostrowski, M., Scanlan, D.J., Garczarek, L., Mazard, S., Palenik, B.P., Paulsen, I.T., de Marsac, N.T., Wincker, P., Dossat, C., et al. (2008). Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. *9*, R90. https://doi.org/10.1186/gb-2008-9-5-r90.

147. El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., et al. (2019). The Pfam protein families database in 2019. Nucleic Acids Res. *47*, D427–D432. https://doi.org/10.1093/nar/gky995.

148. Roux, S., Enault, F., Hurwitz, B.L., and Sullivan, M.B. (2015). VirSorter: mining viral signal from microbial genomic data. PeerJ *3*, e985. https://doi.org/10.7717/peerj.985.

149. Quick, J. (2018). Ultra-long read sequencing protocol for RAD004 v3 (protocols.io.mrxc57n) https://doi.org/10.17504/protocols.io.mrxc57n.

150. Wilson, K. (2001). Preparation of Genomic DNA from Bacteria. Current Protocols in Molecular Biology *Chapter 2*. Unit 2.4.. https://doi.org/10.1002/0471142727.mb0204s56

151. Pfaffl, M.W. (2001). A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res. *29*, e45. https://doi.org/10.1093/nar/29.9.e45.

152. Laurenceau, R., Bliem, C., Osburne, M.S., Becker, J.W., Biller, S.J., Cubillos-Ruiz, A., and Chisholm, S.W. (2019). Toward a genetic system in the marine cyanobacterium Prochlorococcus https://doi.org/10.1101/820027.

153. Osburne, M.S., Holmbeck, B.M., Frias-Lopez, J., Steen, R., Huang, K., Kelly, L., Coe, A., Waraska, K., Gagne, A., and Chisholm, S.W. (2010). UV hyper-resistance in Prochlorococcus MED4 results from a single base pair deletion just upstream of an operon encoding nudix hydrolase and photolyase. Environ. Microbiol. *12*, 1978–1988. https://doi.org/10.1111/j.1462-2920.2010.02203.x.

154. Bushnell, B. (2014). BBTools Software Package. http://sourceforge.net/projects/bbmap.

155. Khil, P.P., and Camerini-Otero, R.D. (2002). Over 1000 genes are involved in the DNA damage response of Escherichia coli. Mol. Microbiol. *44*, 89–105.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Prochlorococcus strain MIT0604	Biller et al.[74]	N/A
Prochlorococcus strain MIT9202	Thompson et al.[92]	N/A
Prochlorococcus strain MIT9312	Coleman et al.[1]	N/A
Prochlorococcus strain MIT9215	Kettler et al.[26]	N/A
Prochlorococcus strain SB	Shimada et al.[93]	N/A
Prochlorococcus strain PAC1	Penno et al.[94]	N/A
Prochlorococcus strain MIT1013	This study	N/A
Prochlorococcus strain MIT1306	Cubillos-Ruiz et al.[95]	N/A
Alteromonas macleodii strain MIT1002	Biller et al.[96]	N/A
Deposited data		
Global Ocean Reference Genomes	EBI/NCBI	https://www.ebi.ac.uk/ena/data/view/PRJEB33281
Station ALOHA viral-fraction nanopore reads	NCBI Sequence Read Archive	SRX7079550
Vesicle-fraction metagenomes	NCBI Sequence Read Archive	SRP272691
RNA-sequencing reads	NCBI GEO	PRJNA719560
Oligonucleotides		
See Table S5 for RTqPCR reaction primers and End-point PCR primers		
Software and algorithms		
checkm v1.0.7	Parks et al.[97]	N/A
Guppy v3.0.4	Oxford Nanopore Technologies, Ltd.	N/A
PROKKA v1.12-beta	Seemann[98]	N/A
panX sha-0a4dfce	Ding et al.[99]	N/A
HMMER3 v3.2.1	Eddy[100]	N/A
seqkit v10.2	Shen et al.[101]	N/A
mafft v7.310	Nakamura et al.[102]	N/A
trimAl v1.4	Capella-Gutiérrez et al.[103]	N/A
msa-concatenate/msa-trim/msa-codon sha-2334c67	https://github.com/thackl/phylo-scripts/	N/A
FastTree v2.1.10	Price et al.[104]	N/A
Phytools	Revell[105]	N/A
ggtree v2.5.0	Yu et al.[106]	N/A
Ori-Finder v1.0	Gao and Zhang[107]	N/A
ALLMAPS v0.7.7	Tang et al.[108]	N/A
R v3.5.1	R Core Team[109]	N/A
HMM v1.0	Himmelmann[110]	N/A
MMseq2 sha-45111b	Steinegger and Söding[111]	N/A
VirSorter2	Guo et al.[112]	N/A
IslandViewer4	Bertelli et al.[113]	N/A
HHPred/HHSearch v3.1.0	Steinegger et al.[30] and Zimmermann et al.[114]	N/A
AliView	Larsson[115]	N/A
FigTree	Rambaut[116]	N/A
prodigal v2.6.3	Hyatt et al.[117]	N/A
ARAGORN v1.2.38	Laslett and Canback[118]	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
BLAST+ v2.8.1	Altschul et al.[119]	N/A
ggplot2	Wickham[120]	N/A
https://github.com/thackl/pro-tycheposons	This study (https://doi.org/10.5281/zenodo.4642441)	N/A
ggraph v2.0.2	Pedersen[121]	N/A
iTOL	Letunic and Bork[122]	N/A
wtdbg2 sha-8926622	Ruan and Li[123]	N/A
Geneious	Kearse et al.[124]	N/A
Mauve	Darling et al.[125]	N/A
minimap2	Li[126]	N/A
Ribbon	Nattestad et al.[127]	N/A
SMRT Analysis 2.3.0	Chin et al.[128]	N/A
IMG Annotation Pipeline version 4	Chen et al.[129] and Markowitz et al.[130]	N/A
ProPortal CyCOGs 6.0	Berube et al.[21]	N/A
DIAMOND v0.9.4	Buchfink et al.[131]	N/A
orfm v0.7.1	Woodcroft et al.[132]	N/A
edgeR v3.20	Robinson et al.[133]	N/A
bbduk v38.16	Bushnell[134]	N/A
mcorr v20180102	Lin and Kussell[135]	N/A
topGO v2.34.0	Alexa and Rahnenfuhrer[136]	N/A
GOView	Liao et al.[137]	N/A
Burrows-Wheeler Aligner v0.7.16a-r1181	Li and Durbin[138]	N/A
HTSeq package v0.11.2	Anders et al.[139]	N/A
DESeq2 v1.24.0	Love et al.[140]	N/A

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Thomas Hackl (t.hackl@rug.nl).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table. Raw *Prochlorococcus* isolate and single-cell assemblies are available through public databases such as NCBI Genbank, the reference-scaffolded assemblies with our lifted annotations are available from GitHub (https://github.com/thackl/pro-tycheposons) and Zenodo (https://doi.org/10.5281/zenodo.4642441). Global Ocean Reference Genomes can be accessed at https://osf.io/pcwj9 and as EBI/NCBI bioproject: PRJEB33281. RNA-Seq data of *Prochlorococcus* strain MIT0604 are available from NCBI GEO: PRJNA719560. The Tara Oceans viral-fraction short-read metagenome contigs can be accessed through the Data Commons portal of iVirus under GOV2.0 (filename: Tara_assemblies.tar.gz). Station ALOHA viral-fraction nanopore reads are available from the NCBI Sequence Read Archive: SRX7079550. Vesicle-fraction metagenomes are available from the NCBI Sequence Read Archive: SRP272691. All original code has been deposited at GitHub (https://github.com/thackl/pro-tycheposons) and Zenodo (https://doi.org/10.5281/zenodo.4642441). Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Strain isolation

Prochlorococcus strain MIT1013 was isolated from seawater obtained on the BiG-RAPA (Biogeochemical Gradients – Role in Arranging Planktonic Assemblages) expedition aboard the R/V Melville (MV1015) during the late austral spring of 2010 (18 November

2010 – 14 December 2010). Seawater was collected at Station 7 (Latitude: -26.25; Longitude -104) using a Niskin bottle rosette (Cast 68, 10 December 2010, 15:03 GMT) from a depth of 150m, corresponding to the subsurface chlorophyll maximum. Seventeen mL of seawater was aliquoted into an acid-washed 28 mL screw cap polycarbonate tube and amended with 20 μM ammonium chloride, 1 μM sodium phosphate, 1 mM sodium bicarbonate, 0.117 μM ethylenediaminetetraacetic acid, 0.117 μM iron (III) chloride, 0.009 μM manganese (II) chloride, 0.0008 μM zinc (II) sulfate, 0.0005 μM cobalt (II) chloride, 0.0003 μM sodium molybdate, 0.001 μM sodium selenite, and 0.001 μM nickel (II) chloride. The MIT1013 strain has been deemed unialgal based on observations of a single *Prochlorococcus* population using flow cytometry and by the presence of a single 16S–23S rRNA internal transcribed spacer (ITS) sequence as determined by direct sequencing of its ITS PCR amplicon. The complete genome sequence of MIT1013 was additionally determined. Cells were grown to mid-exponential phase and pelleted by centrifugation.

Culture conditions

Prochlorococcus cells (axenic cultures, except for MIT1013 and PAC1, and MIT0604 when specified) were grown under constant light flux (30–40 μmol photons m^{-2} s^{-1}) at 24°C in natural seawater-based Pro99 medium containing 0.2-μm-filtered Sargasso Sea water, amended with Pro99 nutrients (N, P, and trace metals).[141] Growth was monitored using bulk culture fluorescence measured with a 10AU fluorometer (Turner Designs).

Two independently growing cultures of MIT0604: the 'nitrate-culture' corresponds to regular Pro99 media culture described above, while the 'ammonia-culture' was grown in Pro99 media with the 800 μM ammonium chloride (NH4Cl) omitted and replaced by 800 μM sodium nitrate (NaNO3), leaving ammonium as the only nitrogen source.

Prochlorococcus isolate strains used in this study. HOTS: Hawaii Ocean Time Series station, Pacific oligotrophic gyre.

Strain	Clade	Location of isolation, depth	Reference
MIT0604	HLII	HOTS, 175m	Biller et al.[20]
MIT9202	HLII	Tropical Pacific, 79m	Thompson et al.[92]
MIT9312	HLII	Gulf Stream, 135m	Coleman et al.[1]
MIT9215	HLII	Equatorial Pacific, surface	Kettler et al.[26]
SB	HLII	Western Pacific, 40m	Shimada et al.[93]
PAC1	LLI	HOTS, 100m	Penno et al.[94]
MIT1013	LLI	Eastern South Pacific Subtropical Gyre, 150m	This study
MIT1306	LLIV	HOTS, 150m	Cubillos-Ruiz et al.[95]

METHOD DETAILS

Genomic datasets

Prochlorococcus genomes

For our analyses, we selected a set of 623 publicly available *Prochlorococcus* genome assemblies, 73 obtained from cultured isolates, 540 generated through single-cell sequencing and 10 extracted from metagenome assemblies (Table S1). We quality-screened the assemblies using checkm v1.0.7.[97] The selected assemblies have a minimum completeness of 25% and median completeness of 74%; isolates and single-cell genomes have less than 4% contamination, metagenome-assembled genomes less than 8%. Raw assemblies are available through public databases such as NCBI Genbank, the reference-scaffolded assemblies with our lifted annotations are available from GitHub (https://github.com/thackl/pro-tycheposons) and Zenodo (https://doi.org/10.5281/zenodo.4642441).

Global Ocean Reference Genomes (GORG) Tropics

This dataset consists of 12,715 single amplified genomes (SAGs) of Bacteria and Archaea, which were obtained through a randomized cell selection from 28 globally distributed samples of tropical and subtropical, epipelagic ocean water.[24] GORG-Tropics SAGs represent all major lineages of surface ocean prokaryoplankton. Used as a reference database, GORG-Tropics recruits an average of 40% reads from tropical and subtropical epipelagic metagenomes with >95% nucleotide identity. The dataset can be accessed at https://osf.io/pcwj9 and https://www.ebi.ac.uk/ena/data/view/PRJEB33281.

Vesicle-fraction metagenomes

We generated paired metagenomes from the cellular and vesicle/small particle fractions of four oligotrophic water samples. All samples were collected at Station ALOHA (22.75 °N, 158 °W) in the North Pacific Subtropical Gyre on cruises HOT263 (June 2014; 5m depth) and HOT283 (April 2016; 5m, 25m, and 137m depth). Each set of samples was derived from a total of ~100-200 L of water retrieved from Niskin bottles. For cellular metagenomes, 3-6 L of water was filtered onto a 0.2 μm Sterivex filter (Millipore) and preserved. As previously described in Biller et al.,[142] DNA was later extracted using a phenol/chloroform-based extraction. Lysing Matrix

E beads (MP Biomedicals), 400 ul Phenol:Chloroform:IAA (25:24:1) and 400 ul 2x TENS buffer (100 mM Tris-HCL pH 8.0, 40 mM EDTA, 200 mM NaCl, 2% SDS for 2x buffer) were added to a microcentrifuge tube containing the filter and then vigorously agitated using a beadbeater for 40 seconds. After spinning at 19,000 xg for 5 minutes, the aqueous phase was transferred into a Phase Lock Gel tube (5 Prime), mixed with an equal volume of chloroform, and then spun at ~27,000 x g for 5 minutes. The supernatant was removed and mixed with an equal volume of AMPure XP beads (Beckman Coulter), and incubated at room temperature for 10 minutes. Beads were washed twice with 75% ethanol, dried, and resuspended in 20 uL ultrapure glass distilled water (Teknova). Total DNA yield was quantified using the PicoGreen assay (ThermoFisher) with yields ranging from ~10-2600 ng total DNA.

The <0.2 μm fraction of the remaining water was concentrated on a 100 kDa tangential flow filter and further fractionated using an Optiprep gradient; vesicle-enriched fractions were identified, DNA outside the vesicles removed using TURBO DNAse, and the remaining DNA (presumably within vesicles) was extracted as previously described[74]: Vesicles were lysed in GES buffer (50 mM guanidinium thiocyanate, 1 mM EDTA, and 0.005% (w/v) sarkosyl; final concentration) at 37 °C for 30 minutes. DNA was purified using DNA Clean & Concentrator-5 columns (Zymo Research) per the manufacturer's instructions, using a 5:1 ratio of DNA binding buffer, and eluted in ultrapure water. Sequencing libraries for the purified cellular and vesicle/particle-associated DNA were prepared from ~1 ng of DNA using the NextEra XT kit (Illumina) and 150+150nt paired-end sequences generated by an Illumina NextSeq 500 at the MIT BioMicro Center. All data are available from the NCBI Sequence Read Archive (SRP272691).

Tara Oceans viral-fraction metagenome contigs

The short-read viral data were obtained from 131 samples collected during the Tara Oceans and Tara Oceans Polar Circle expeditions and span water layers from the surface (5m deep) to the mesopelagic ocean (up to 1000m deep). The full description of the collection locations and depths as well as the viral enrichment, DNA sequencing, and contig assembly protocols for these samples can be found in the methods section of the Global Ocean Viromes 2.0 (GOV2.0) dataset.[72] GOV2.0's deep ocean samples from the Malaspina expedition (n=14) were not included in this study. All the assembled contigs from the 131 Tara samples were screened for tycheposons prior to any bioinformatic viral selection carried out in Gregory et al.[72] to establish the GOV2.0 dataset. These assembled contigs can be accessed through the Data Commons portal of iVirus under GOV2.0 (filename: Tara_assemblies.tar.gz)

Station ALOHA viral-fraction nanopore reads

The 25 m deep sample was collected on the HOT-314 cruise on August 5, 2019 at Station ALOHA (22°45' N, 158° W; http://hahana.soest.hawaii.edu/hot/). Sample collection, filtration, viral concentration, extraction, and sequencing have all been previously described.[71] Briefly, the seawater was pre-filtered by peristaltic pumping through a 0.22 um filter (Sterivex GV) then concentrated by tangential flow filtration (TFF) over a 30 kDa filter (Biomax 30 kDa) membrane, catalog #: P3B030D01, Millipore). Following multiple rounds of centrifugal concentration, lysis and DNA purification were performed in a single tube using the Qiagen Genomic-tip 20/G protocol following manufacturer's recommendations. Virus-enriched samples from 110 L of 0.22um pre-filtered seawater yielded a total of 3.2 μg of purified, high molecular weight DNA. Sequencing was conducted on a GridION X5 with FLO-MIN106 (R 9.4.1) flowcells (Oxford Nanopore Technologies, Ltd.). The resulting 701,515 reads were basecalled using Guppy v3.0.4, generating 10.38Gb of sequencing data with a read N50 length of 29.70 Kb. All data are available from the NCBI Sequence Read Archive (SRX7079550).

Computational analyses

Gene prediction and ortholog detection

To ensure consistent gene annotations, all 623 *Prochlorococcus* genomes (see section Genomic datasets for details) were reannotated using PROKKA v1.12-beta.[98] The annotated genes were clusters into groups of orthologs using panX sha-0a4dfce[99] with customized settings to account for the incompleteness of the single-cell genomes in the collection: core genes are defined by being present in at least 70% of all genomes and in every genome of >98% completeness (-cg 0.70 -csf strains_complete_98plus.txt).

Phylogenetic tree reconstruction

The phylogenetic reference tree for all 623 *Prochlorococcus* genomes analyzed was constructed using the maximum likelihood method from 109 concatenated single-copy core proteins. The markers were identified using HMMER3 v3.2.1[100] based on the 120 TIGRFAM[143] profiles previously described as ubiquitous bacterial single-copy core genes ("bac120").[144] 11 proteins found to be not single-copy in this specific data set were excluded. Sequence files were manipulated with seqkit v10.2[101] and individual protein alignments were generated with mafft v7.310,[102] trimmed with trimAl v1.4[103] (-gappyout), and concatenated with msa-concatenate (https://github.com/thackl/phylo-scripts/, sha-2334c67). The maximum likelihood phylogeny was inferred with FastTree v2.1.10.[104] The tree was rooted at the LLIV clade and visualized using the R packages phytools[105] and ggtree v2.5.0.[106]

Reference-based genome-scaffolding

All 34 complete genomes comprising only a single contig were considered as references. For better comparability, all reference genomes were oriented and rotated – they are circular chromosomes – to start with the origin of replication near the dnaN gene on the plus strand. This is consistent with the convention used for most published *Prochlorococcus* genomes. OriC prediction was performed with the command line version of Ori-Finder v1.0[107] with *dnaA* box sequence set to "TTTTCCACA" as suggested for cyanobacteria.[145] The start positions of 11 genomes were adjusted, and 3 of them were also reverse complemented. For all other genomes, the closest reference genome was determined by the smallest cophenetic distance in the reference phylogenetic tree described above. For each pair of draft genome and closest reference, 1x1 anchor maps were created by matching genes belonging to the same cluster of orthologous genes (see above). These anchor maps were used to orient and order contigs in a way that maximizes collinearity using ALLMAPS v0.7.7.[108] If necessary the contig spanning the beginning and end of the reference sequence was split

into two parts at the position of the *dnaN* gene. Gaps between contigs were estimated by minimizing the sum of absolute distances between corresponding genes from the 1x1 map. Contigs were weighted by the log of their length and minimum gap size was set to 100bp. The optimization was implemented using the "L-BFGS-B" method in R v3.5.1.[109] See section *Prochlorococcus* genomes for access to the data.

Genomic island predictions

To automate the annotation of genomic islands across the entire dataset, we devised an HMM-based approach that uses frequencies of orthologous genes as input. We define the frequency of a gene and its respective orthogroup, as the number of strains the gene is present in. Duplications within the same genome are ignored. We are working off the observation that islands are enriched in non-core genes, and hence their composition in terms of gene frequencies is different from non-island regions.[1] We used previously described genomic islands from 2 *Prochlorococcus* HLI strains (MED4, MIT9515), 2 *Prochlorococcus* HLII strains (MIT9312, MIT9215), 1 *Prochlorococcus* LLII strain (SS120) and 1 *Prochlorococcus* LLIV strain (MIT9313)[1,3,146] to generate four profiles of the island and non-island gene frequencies. Based on comparisons of these profiles we then defined different states for the HMM: core, flex, and inconclusive depending on their prevalence in or outside of islands. Using the R-package HMM v1.0[110] we built four HMMs with two hidden states (island, non-island). Start, transition, and emission probabilities were estimated from the literature annotations. For all genomes, each gene was assigned the category (core, flex, and inconclusive) depending on the gene frequency as described above. Using these labels in the order they appear on the scaffolded sequence (see above) as observations the Viterbi algorithm was applied to predict the hidden state for each gene (island, non-island). This gene-level resolution of genomic islands was used in further analyses.

We first reference-scaffolded and reorganized all assemblies to obtain single-chromosome scaffolds with consistent start and orientation. Using coordinates of known genomic islands[1,3,146] and abundances of orthologous gene clusters[99] we trained a Hidden-Markov-Model to predict islands across all genomes (Figure S1A).

Searches for known mobile genetic elements

Initially, we screened for the presence of common MGEs using similarity searches (HMMER3,[100] MMseq2[111]) against a comprehensive collection of MGE protein and profile databases (NCBI viral RefSeq, NCBI Plasmid RefSeq, ACLAME, ICEBerg, COMPASS, immedb, and ISfinder, pVOG[42]). Moreover, we ran automated annotation tools (VirSorter2,[112] IslandViewer4[113]). However, none of these searches returned clear hits except to some already known transposons and insertion sequences in LLIV *Prochlorococcus*. After identifying putative horizontally transferred sequences based on genomic island annotations, we used remote homology detection with HHPred/HHSearch v3.1.0[30,114] to assign functional predictions to otherwise unannotated genes.

Identification of in mobilome hallmark genes

To enable a comprehensive *Prochlorococcus*-focused search for integrative MGEs we compiled and curated protein HMM-profiles of genes typically found in candidate MGEs using an iterative, explorative approach: Starting with a handful of manually annotated high-confidence MGE candidates we identified relevant orthogroups (see section Gene prediction and ortholog detection) based on one of the two following criteria:

a) orthogroups with some annotations associated with the excision-replication-packaging life-cycle typical for MGEs (integrases, primases, helicases, capsid genes, terminases), and
b) orthogroups appearing in multiple candidates with syntenically conserved patterns.

From these orthogroups we created and curated, redundancy-reduced alignments and HMM-profiles using combinations of the following tools: mafft v7.310,[102] msa-trim (https://github.com/thackl/phylo-scripts/, sha-2334c67), AliView,[115] FastTree v2.1.10,[104] FigTree,[116] and HMMER3 v3.2.1.[100] Orthogroups with good reciprocal hits in all-versus-all comparisons and overall consistent multiple sequence alignment, when aligned all together, were merged. We then scanned all annotated proteins of our collection of 623 *Prochlorococcus* genomes using these profiles and rudimentary versions of the R-scripts described in more detail in the next section. We identified, ranked and visualized genomic regions with multiple hits to different hallmark genes within close proximity to each other (multiple hits within a 10-20kb window). From the thus obtained highest ranking clusters we selected new high-confidence candidates and used those together with previously selected candidates to repeat the identification of hallmark genes, expand the overall size of the gene set and refine the existing profiles.

In addition to curating profiles for proteins often found in putative MGEs in *Prochlorococcus*, we also added generic Pfam profiles[147] with a strong overlap to some of the other hallmark profiles and gathered proteins from four sets of published PICIs.[36–38,56] We grouped these genes into clusters based on the annotations provided in the respective publications, manually curated the alignments, and merged clusters into single profiles if they had good reciprocal hits in all-vs-all comparisons and consistent multiple sequence alignment when aligned together.

Automated mobilome detection and annotation

Ultimately we devised a small pipeline to automate the detection of MGEs in genomic data sets. The scripts are available from GitHub (https://github.com/thackl/pro-tycheposons) and Zendodo (https://doi.org/10.5281/zenodo.4642441), and perform the following steps:

- Gene prediction with prodigal v2.6.3[117] if no annotations are provided
- Full-length tRNA annotation with ARAGORN v1.2.38[118] if no external tRNA database is provided

- Detection of MGE hallmark genes and viral hallmark genes (VirSorter profiles[148]) using HMMER3
- Partial tRNA annotations using full-length tRNAs and BLAST+ v2.8.1[119]
- Attachment site detection in integrase-flanking regions with BLAST+
- Scoring of candidate MGEs based on the presence of different hallmark genes, attachments sites and tRNAs
- Visualization of MGEs using ggplot2[120] and https://github.com/thackl/gggenomes

For this study, we analyzed three datasets with this pipeline:

1) 623 *Prochlorococcus* genomes
2) 2344 single-cell assemblies of the Global Ocean Reference Genomes (GORG)
3) up to 1000 randomly selected 5-20 kbp long contigs from 262 different viral-fraction samples from Tara Oceans (128,566 contigs in total)

For more information on the datasets, see section Genomic datasets. The results from the analyses are available at GitHub (https://github.com/thackl/pro-tycheposons) and Zendodo (https://doi.org/10.5281/zenodo.4642441).

Co-clustering of mobilome hallmark genes

To analyze all *Prochlorococcus* MGEs in the context of known MGEs we used a gene-sharing network approach.[39–41] We sensitively co-clustered all hallmark proteins of the *Prochlorococcus* mobilome (those with functions related to a mobile life-style: recombination, DNA-replication and packaging) together with

- all proteins from NCBI viral RefSeq (https://ftp.ncbi.nlm.nih.gov/refseq/release/viral/, accessed 2021-02-05)
- all proteins from the mobileOG database – a comprehensive database based on 10 million hallmark proteins sequences of all major classes of MGEs (includes ICEBerg, ACLAME, NCBI Plasmid RefSeq, COMPASS, immedb, and ISfinder and pVOG)[42]
- and a comprehensive set of proteins representing phage-inducible chromosomal islands.[36–38,56]

The clustering was performed with MMseq2 sha-45111b ('easy-cluster –evalue 1e-4 –coverage .7').[111]

We further assigned functions to those clusters by matching them to our manually generated profiles. Overall, the automated clustering process showed good congruence with our manually curated protein clusters, although, we note that a with the simplistic approach of a universally applied cutoff, we cannot capture the different characteristics of the diverse proteins set optimally, as indicated by both collapsed clusters of larger conserved proteins (such as integrases) as well as split clusters in particular for likely fast-evolving short co-factors (such as excisionases). The sequences and clustering results are available at GitHub (https://github.com/thackl/pro-tycheposons) and Zenodo (https://doi.org/10.5281/zenodo.4642441).

Construction of gene sharing networks

To assess the structure within the mobile gene pool and the relatedness of the MGEs carrying those genes, we constructed two gene sharing networks. Firstly, we generated a simple gene-sharing network based on the manually curated hallmark protein profiles of the *Prochlorococcus* mobilome. In that network, protein profiles are nodes, and we connect every pair of profiles that co-occur on the same MGE, weighted by how often we observe that co-occurrence. Rare connections (count <3) are ignored. This network provides information about the internal structure of the mobilome, i.e. how many different types (unconnected subgraphs) of MGEs are present in the data.

Secondly, we generated a bipartite network based on automatically generated protein clusters computed from *Prochlorococcus'* mobilome plus a comprehensive set of viral and the MGE protein sets. In this network, we have two types of nodes: protein clusters and MGEs/viral genomes. Connections are drawn between MGEs/viral genomes and the proteins they contain. This second network explicitly highlights the delineation of the tycheposons compared to the realm of known MGEs. The networks were generated in R with ggraph v2.0.2.[121]

Phylogeny of tyrosine recombinases

To put the tyrosine recombinases identified on most of the new tycheposons and cryptic elements into context with previously described integrases, we first compiled a representative set of known recombinase proteins. We used a HMM database of recently described Xer-like tyrosine-recombinases,[35] which was kindly provided to us by the authors, to collect protein sequences from UniRef50 (https://www.uniprot.org/help/uniref). We then combined these sequences with a non-redundant subset of the integrase sequences found in our elements (max pairwise identity of 40%). We aligned the sequences with mafft v7.310[102] (–genafpair), computed a phylogenetic tree with FastTree v2.1.10[104] and visualized it with iTOL[122] followed by manual curation. For the comparison of the phylogenetic diversity contributed by the new integrases, we divided the sum of branch length of the new integrase clade by the total sum of branch lengths in the full tree. Before classifying new putative integrases as such, we also checked each subtype alignment for the presence of the characteristic residues at the catalytic sites to gain more confidence in the prediction of their functionality.

Oxford Nanopore genome assembly

The assemblies of two independently growing cultures of MIT0604 were generated using wtdbg2 sha-8926622[123] from nanopore reads longer than 50,000 bp for the nitrate-culture and 30,000 bp for the ammonia-culture, respectively. In both cases, a single contig matching the complete chromosome of MIT0604 was obtained and extracted for further analysis from the assembly. The contigs

were each reorganised to match the start and orientation of the reference Illumina assembly generated in 2011. The assemblies and read data set were analysed for rearrangements using Geneious,[124] Mauve,[125] minimap2[126] and Ribbon.[127] The processed assemblies are available for download at https://github.com/https://github.com/thackl/pro-tycheposons.

Experimental procedures

Oxford Nanopore sequencing

Two independent cultures of *Prochlorococcus* MIT0604 were sequenced using Oxford Nanopore technologies. Samples were prepared following an ultra-long read sequencing protocol,[149] and whole genomes were sequenced using the Rapid sequencing kit (SQK-RAD004) on a MinION sequencer according to the manufacturer's instructions (Oxford Nanopore).

Prochlorococcus MIT1013 genome sequencing

DNA was isolated by phenol/chloroform extraction.[150] PacBio library preparation and sequencing was carried out by the MIT BioMicro Center and the UMass Worcester Medical School's Deep Sequencing Core Facility. Assembly of PacBio reads was performed using the hierarchical genome assembly process (Protocol = RS_HGAP_Assembly.2) as implemented in SMRT Analysis 2.3.0[128] with the following parameters adjusted: Minimum Polymerase Read Quality = 0.85 and Genome Size = 2000000 bp (default settings were used for all other parameters). A single *Prochlorococcus* contig was identified as well as a 6513 bp contig most closely related to Marinobacter sp., a common heterotrophic contaminant of xenic cultures of *Prochlorococcus*. Overlapping ends of the *Prochlorococcus* contig were identified using BLAST, and the assembled contig was manually circularized. The circular assembly was corrected using the RS_Resequencing.1 protocol in SMRT Analysis 2.3.0 with the following parameters: Minimum Polymerase Read Quality = 0.85 and Consensus Algorithm = Quiver. This genome was deposited with IMG (accession number 2681812904), annotated using IMG Annotation Pipeline version 4,[129,130] and included in ProPortal CyCOGs 6.0.[21]

RTqPCR of integrase genes

Reverse-Transcription qPCR analysis of integrase genes was performed on biological triplicates exposed to the treatment, compared to untreated controls. For RNA preparation, cells were collected by centrifugation (12,000g for 12 min, 20°C) and immediately resuspended in 500 μL of TRI reagent (Zymo Research). RNA was isolated using the Direct-zol RNA MicroPrep kit (Zymo Research) according to the manufacturer's instructions. A DNA removal step was added after elution using the TURBO DNA-free Kit (ThermoFisher). RNA samples were subsequently concentrated using RNA the Clean & Concentrator kit (Zymo Research), followed by reverse transcription using the SuperScript™ III First-Strand Synthesis System (ThermoFisher). Finally, triplicate qPCRs were performed using the QuantiTect Probe PCR Kit (Qiagen) – using the primer sets detailed in Table S5 – and differential gene expression was calculated following the comparative CT (2-DDCT) method,[151] with the expression of gene rnpB as the endogenous reference.

Description of the treatments shown in Figure S5A: 'Alteromonas' = addition of the helper strain *Alteromonas* MIT1002[96] at $5x10^6$ cells mL^{-1} for 1h; 'Pyruvate' = addition of 5 mM pyruvate for 2h; 'Glucose' = addition of 5 mM glucose for 2h; 'Nitrate' = culture grown in Pro99 media with nitrate substituted to ammonium as the nitrogen source; 'N starvation' = cells are washed twice and resuspended in Pro99 media devoid of any nitrogen source, and incubated for 48h, while control cultures are washed but resuspended in replete Pro99; 'P starvation' = same process, using Pro99 media devoid of phosphate; 'Stationary phase' = cultures are left until they reach the stationary phase, while control cultures are harvested in exponential phase; 'Metal toxicity' = trace metals present in Pro99 are added at 5 times their concentration (toxic level) for 1h; 'Copper' = addition of $CuCl_2$ at 100 pM in the culture for 1h; 'Arsenate' = addition of Na_3AsO_4 at 10 μM for 16h; 'H_2O_2' = Addition of hydrogen peroxide at 0.5 μM for 1h; 'DCMU' = addition of DCMU (Diuron herbicide) at 4 μM for 1h; 'Chloramphenicol' = addition of chloramphenicol at 1 μM (lethal dose) for 1h; 'Ciprofloxacin' = addition of ciprofloxacin at 1 μM (lethal dose) for 1h[152]; 'Mitomycin C' = addition of mitomycin C at 20 μM (lethal dose) for 2h; 'Cold shock' = cultures were cooled to 14°C for 1h; 'pH 9.3' = addition of a 0.1 M NaOH solution at 1.4 mM to reach pH 9.3, for 1h; 'pH 6.5' = addition of a 0.1M HCl solution to 2.1 mM to reach pH 6.5, for 1h; 'Dark' = culture tubes are placed in dark for 1h; 'High Light' = culture tubes are placed at a light intensity of 150 μE for 1h; 'UV shock' = cells were irradiated at for 30 sec at 254 nm (UV 100 μW cm^{-2}) in an uncovered sterile glass petri dish (150 mm diameter) containing 25 mL of culture. Cells were placed back in 25 mL culture tubes inside the incubator for 30 min, while control cultures were subjected to the same conditions, omitting the UV irradiation[153]; 'UV acclimation' = Cultures were acclimated to a light level of 50 μE in diel regime, receiving an extra 2h daily UV at midday from a Rayminder 15W UV lamp shining 302-316 nm UVB (16 in distance from lamp). Control cultures were grown in diel regime at 25 μE light intensity and without UV. RNA was harvested at midday in exponentially growing cultures[65]; 'Exogenous DNA' = 1 μg of pUC19 plasmid DNA was electroporated in concentrated cell samples,[152] which were then left to recover for 24h before RNA extraction. The control cultures were also electroporated in the absence of plasmid DNA.

RNA-sequencing experiments

Nine 30 mL cultures of *Prochlorococcus* MIT0604 strain were grown to exponential phase. 3 cultures were treated with mitomycin C (Sigma-Aldrich) at a final concentration of 15 μg mL^{-1} for 2 hours; 3 cultures were applied a UV shock by irradiating cells at room temperature for 30 s at 254 nm (UV 100 μW cm^{-2}) in an uncovered sterile glass Petri dish (150 mm diameter) – then placed back into their initial culture condition for 1h[153]; the 3 remaining cultures were kept as control with no treatment. Cells were harvested by centrifugation at 12,000 g for 12 min, 20°C and RNA was extracted using the mirVana microRNA (miRNA) extraction kit (Ambion, Carlsbad, CA, USA). All strand-specific transcriptome sequencing (RNA-seq) libraries were constructed using the KAPA RNA HyperPrep kit (Illumina) and used the RiboZero kit (Illumina) for ribosomal RNA depletion. Sequencing was carried out on an Illumina

NextSeq 500 instrument at the BPF Next-Gen Sequencing Core Facility at Harvard Medical School, with a High-Output 75-cycle kit to obtain Single-Read 75bp reads.

Tycheposon mobility in lab isolates

Excision of elements in lab isolates was probed using end-point PCR specific to the chromosomal vacated site ('excised'), or the circular/tandem repeats state of the elements (Figure S4). Primers are listed in Table S5. 25 mL duplicate cultures were grown to exponential phase and sampled at t_0 (before the addition of mitomycin C) and t_{2h} (2h post-addition of mitomycin C at 10 $\mu g\ mL^{-1}$ final concentration). For sampling, 10 mL of cells were harvested by centrifugation (7,000 g for 30 min, 20°c), and DNA was extracted using the DNeasy Blood & Tissue kit (Qiagen). PCR reactions were performed using the Quick Load Taq 2X master mix (New England Biolabs) with 2 min elongation time and 52°C annealing temperature. PCR products were directly loaded on 1% agarose gels for visualization. Each band displayed on the figure was checked by Sanger sequencing and corresponded to the expected amplicon.

QUANTIFICATION AND STATISTICAL ANALYSIS

Tycheposon-containing reads in viral-fraction

To annotate tycheposon and cryptic element hallmark genes in raw nanopore reads, which have error-rates too high to predict open reading frames, we used a different strategy: We converted the alignments used to generate the HMM-profiles for hallmark genes into a protein reference database and used DIAMOND v0.9.4[131] in long-read mode to align nanopore reads to this database. We then identified and visualized reads with multiple hits to different hallmark genes.

Tycheposon signatures in vesicle- and cellular-fraction

To assess the relative abundance of tycheposon and cryptic element hallmark genes in cellular- and vesicle-fraction metagenomes, we recruited reads translated into amino-acid space using orfm v0.7.1[132] with HMMER3 (–evalue 1e-20) to all element hallmark gene profiles and 109 *Prochlorococcus* core gene profiles. See section Phylogenetic tree reconstruction for *Prochlorococcus* for details on the generation of the single-copy core marker gene set. We then analyzed the obtained counts for differential abundance with edgeR v3.20.[133] We only considered profiles with a minimum count of 5 in at least two cellular- and two viral-fraction samples. In the absence of replicates, we estimated the dispersion from all core genes across all samples, assuming that this would provide us with a reasonable yet rather conservative estimate. We normalized for library size (method="TMM") and tested for differences between the two fractions using the exact test and the dispersion estimated from the core genes.

Differential abundance of attachment sites

To assess the abundance of tRNA sequences that might serve as attachment site in the vesicle- and cellular-fraction samples we applied a two-step process: First, we screened for reads with at least one almost exact 39bp match using bbduk v38.16[134] (k=39 edist=2) to a non-redundant reference database of marine tRNA genes we compiled using the GORG single-cell genomes (see section Global Ocean Reference Genomes). We then blasted those reads with settings optimized for short almost exact matches (-task blastn -reward 1 -penalty -4 -gapopen 5 -gapextend 2 -perc_identity 94 -evalue 10e-5) against both 5' and 3' halves of all reference tRNAs. We then further analyzed the resulting counts in R: we filtered for a minimum alignment length of 38bp and tested for tRNAs with differentially abundant 5' and 3' regions with Fisher's exact test and the Bonferroni correction to adjust p-values for multiple testing. The resulting count distributions were visualized with ggplot2.[120]

Island size and frequency of MGE integration

To test if island size correlates with the frequency of tycheposon integrations, we computed the median sizes of the islands adjacent to the 7 tycheposon-targeted tRNA genes excluding integrated MGEs for each *Prochlorococcus* clade. We then compared these island sizes to the number of observed integrated tycheposons per clade and determined P- and R^2-values through a linear fit.

Estimation of homologous recombination rates

Genomic island formation is often driven by homologous recombination between flanking core genes.[46,78] To test if this is also the case for the islands in *Prochlorococcus* we estimated recombination rates for genomic regions relative to their distance to genomic islands. First, we identified backbone core genes (orthogroups not found in islands) individually for the 5 large monophyletic clades HLI, HLII/VI, LLI, LLII/III and LLIV. We then estimated a proxy for the average distance of each cluster to the closest island across the entire clade. For that, we took the 25% quantile of all gene-to-closest-island distances we obtained from each individual genome. Next, we generated protein alignments for all of those clade-wise clusters with mafft v7.310,[102] mapped back the nucleotide codons onto the amino-acid alignments (msa-codon https://github.com/thackl/phylo-scripts/) and trimmed positions with more than 70% gaps with trimAl v1.4[103] (-gt .3). We then concatenated all alignments ordered by their estimated distance to the closest island, with island-flanking genes at the beginning and genes furthest from islands at the end. Finally, we partitioned those clade-wise concatenated alignments into 99999 nucleotide long blocks. For each of those blocks, we estimated recombination rates and related variables using mcorr v20180102,[135] and compared those block-wise results with respect to the proximity of contained genes to genomic islands.

Functional enrichment analysis

For the functional enrichment analysis, all genes were split into three sets: backbone, island, elements. Detected hallmark genes and those smaller than 201bp were excluded. The sets are mutually exclusive so genes within elements within islands are only assigned the element category. Pairwise functional enrichment analyses were performed between backbone-element and between backbone-island using the R-package topGO v2.34.0.[136] Significantly enriched GO terms (p<1e-10, only GO terms with at least 100 occurrences are considered) in the two analyses were compared using GOView[137] for each of the GO categories: molecular function, cellular component and biological process.

RNA-sequencing analyses

Adapters were trimmed from the raw Illumina data with bbduk v38.16,[154] with settings ktrim=r, k=23, mink=11, hdist=1. Low-quality regions were removed from the adapter-trimmed sequences using bbduk v38.16,[154] with parameters qtrim=rl, trimq=6. The trimmed RNA-seq reads were aligned to a reference file containing the MIT0604 genome (available from https://github.com/https://github.com/thackl/pro-tycheposons/) with the Burrows-Wheeler Aligner v0.7.16a-r1181,[138] using the BWA-backtrack algorithm. To determine the number of reads that aligned to each annotated ORF in the "sense" and "antisense" orientations, we parsed the mappings using the HTSeq package v0.11.2[139] with default parameters and the "nonunique all" option. We compiled the counts of reads that aligned to each ORF (excluding rRNAs and tRNAs because library preparation included ribosomal depletion) across replicates. We identified differentially expressed genes using the DESeq2 R package v1.24.0.[140] Using the standard DESeq2 functions and workflow, we normalized samples by library sequencing depth and estimated the dispersion of each gene. Differential expression tests were performed on mitomycin C vs. control and UV shock vs. control comparisons with the Wald test, using a negative binomial generalized linear model. P-values were corrected for multiple testing with the Benjamini–Hochberg procedure. As was suggested by the DESeq2 authors,[140] genes with an adjusted p-value of <0.1 were considered to have significantly different expression between a given pair of treatments. We visualized the differential expression results with ggplot2.[120]

Supplemental figures

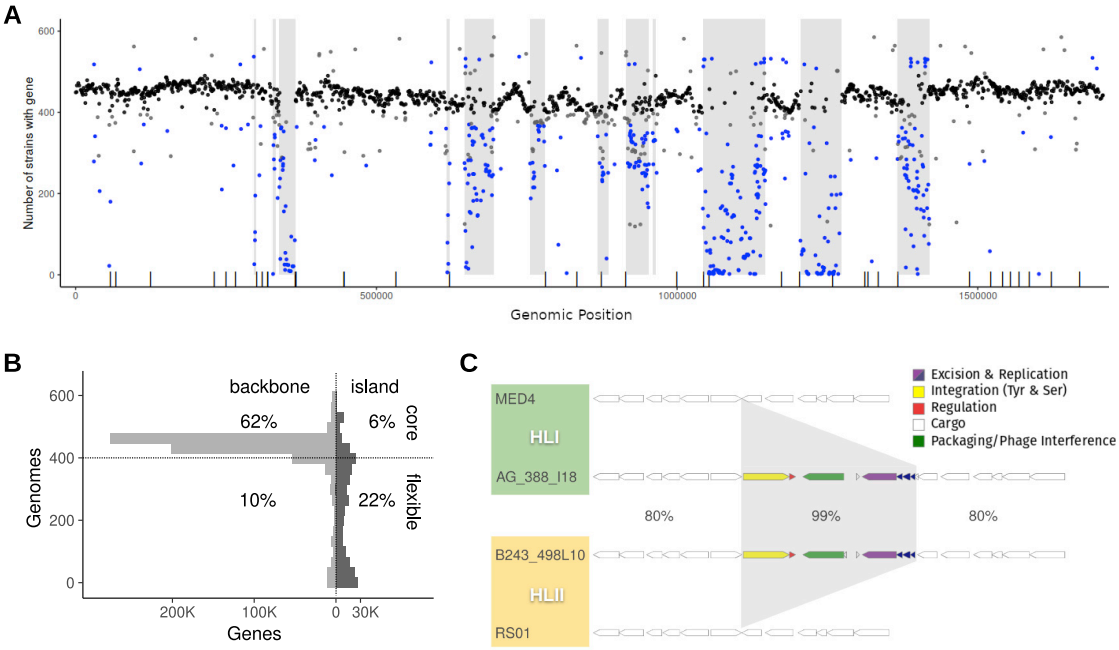


Figure S1. HMM-based genomic island prediction using abundances of orthologous genes, related to STAR Methods

(A) Result of the prediction on strain *Prochlorococcus* MIT9312 (HLII). Each gene is depicted as a dot at its genomic position with the number of strains that possess this gene on the y axis. Genes are colored by their class (core, black; flex, blue; unsure, gray). The sequence of genes and their class is fed to the clade-specific HMM as observations and hidden states are predicted through the Viterbi algorithm. The resulting island regions are depicted as vertical gray bars. Additionally, the location of tRNA genes is shown as black ticks on the x axis.

(B) A histogram showing the distribution of genes contained in the genomic backbone and genomic islands with respect to the number of genomes the respective genes are observed on the y axis. Overall, we analyzed 623 genomes, with median completeness of 74%. For the purpose of this summary and to account for the incompleteness of the genomes and statistical variance, we define core genes as genes present in at least 65% of the genomes (dotted line). Based on this definition, more than $^{2}/_{3}$ of *Prochlorococcus* flexible genes are contained in genomic islands we annotated.

(C) A putative horizontally transferred MGE present in two different *Prochlorococcus* genomes from two different clades (HLI:AG-388-I18 and HLII:B243-498L10) but missing from closely related isolates of the same respective clades. Predicted gene functions are colored-coded: yellow, large serine recombinase; red, transcriptional regulator; green, major capsid protein; purple, helicase; blue, putative replication factors.

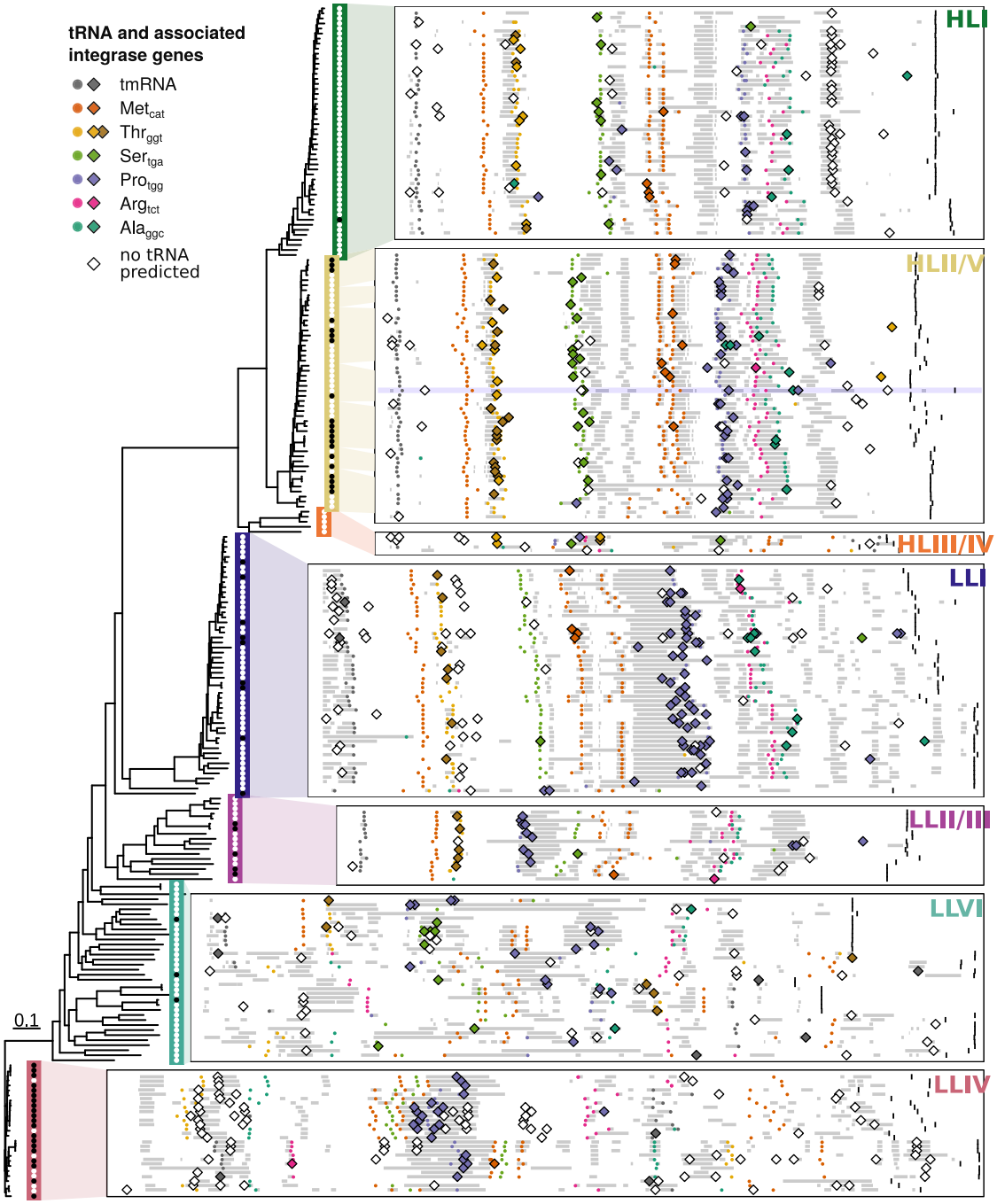


Figure S2. Chromosomal organization of genomic islands and mobile genetic elements in *Prochlorococcus*, related to Figure 7 and STAR Methods

284 finished or reference-scaffolded circular *Prochlorococcus* shown relative to their origin of replication (left- to rightmost black mark). Vertical column-like features indicate the predicted genomic islands in conserved locations across the genomes (gray bars). Most islands are associated with one or two specific full-length tRNA genes (colored points). These tRNAs are targeted by mobile genetic elements carrying integrases specific to the different islands (colored diamonds). Only the 50 most complete genomes of each clade (in the case of HLII/VI, 57) of all the 623 genomes used in this study are shown. The genomes are ordered according to their phylogenetic relationships (black dots, cultured isolates; white dots, single-cell amplified genomes), i.e., the most closely related genomes are plotted next to each other and are grouped in different panels corresponding to known *Prochlorococcus* clades and grades of the low-light- (LL) and high-light- (HL) adapted ecotypes. The genome of strain MIT0604 used for experimental analyses is highlighted in blue.

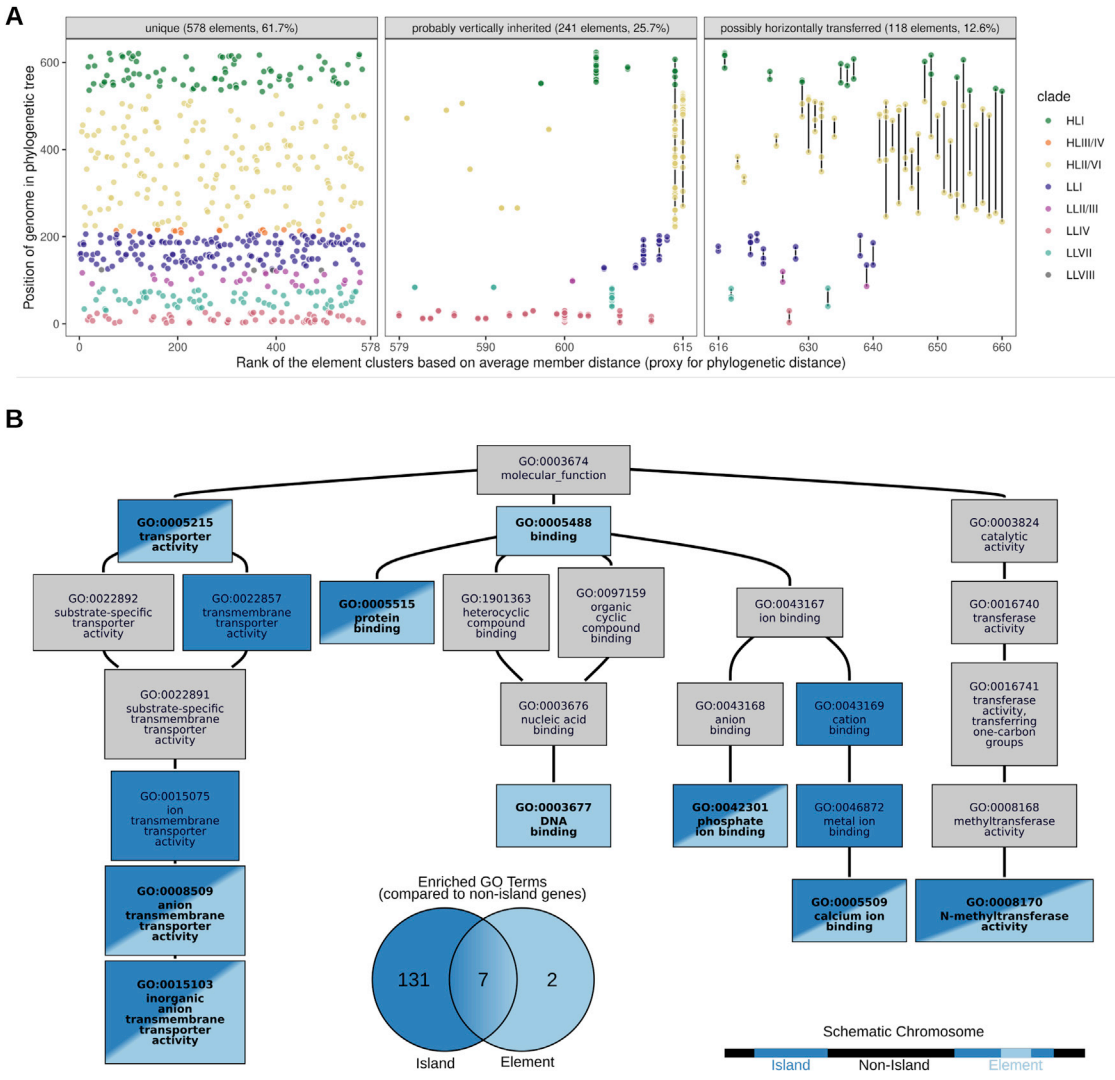


Figure S3. Distribution of mobile elements across 623 *Prochlorococcus* genomes and functional overlap between elements and islands, related to STAR Methods

(A) Elements are clustered (colored points linked by vertical black lines) based on significant pairwise similarity on the nucleotide level (at least 90% identity over 50% of the shorter element). The panels further group the mobile elements into three categories: elements found in only a single genome; elements found in some closely related genomes, likely representing vertical transmission; and elements found in multiple genomes with a noticeable difference in the phylogenetic position of their hosts, indicating possible horizontal transfer events.

(B) A subset of gene-ontology (GO) terms in the molecular function subontology. Enriched terms (compared to nonisland genes) are colored in shades of blue (see "functional enrichment analysis" in STAR Methods for details). Light blue indicates enrichment of nonhallmark genes (genes not related to recombination and DNA replication) on elements, while dark blue indicates that these terms are also enriched on genomic islands (excluding elements). There are another 127 terms in this subontology that are enriched on genomic islands but not on elements (not shown).

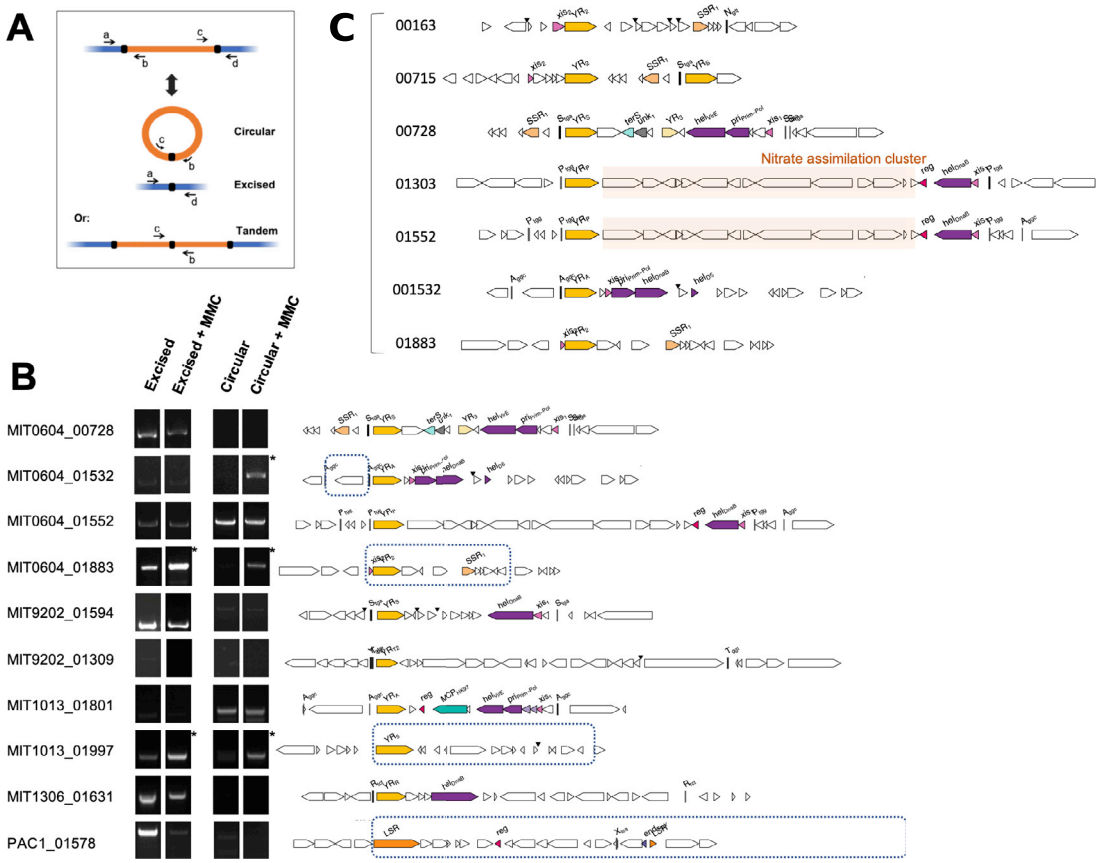


Figure S4. Detection of integration and excision of elements in cultures, related to STAR Methods

(A) Cartoon of the PCR strategy to detect element excision, showing a, b, c, d primer design listed in Table S5.

(B) PCR products run on an agarose gel. The gel exposure varies from one PCR amplicon to another but is identical for the same amplicon ± mitomycin C. When the direct tRNA repeat borders of the element are not apparent, a blue dotted box indicates the predicted borders. Most elements show that the starting population is heterogeneous (excision or tandem repeats of elements amplify from a subpopulation of cells). Only a few elements show mobilization by mitomycin C (indicated by a star), producing a likely circular intermediate. Interestingly, primers for the element MIT0604_01532 amplify a segment of DNA that does not contain the integrase, demonstrating the ability of elements to move segments of DNA in *trans*, as long as attachment sites are present.

(C) Overview of the 7 different elements found in the *Prochlorococcus* MIT0604 reference strains. Elements 01303 and 01552 are the two identical copies of the tycheposon carrying the complete nitrate assimilation cluster[31] (orange box). The gene colors and labels are identical to Figure 1.

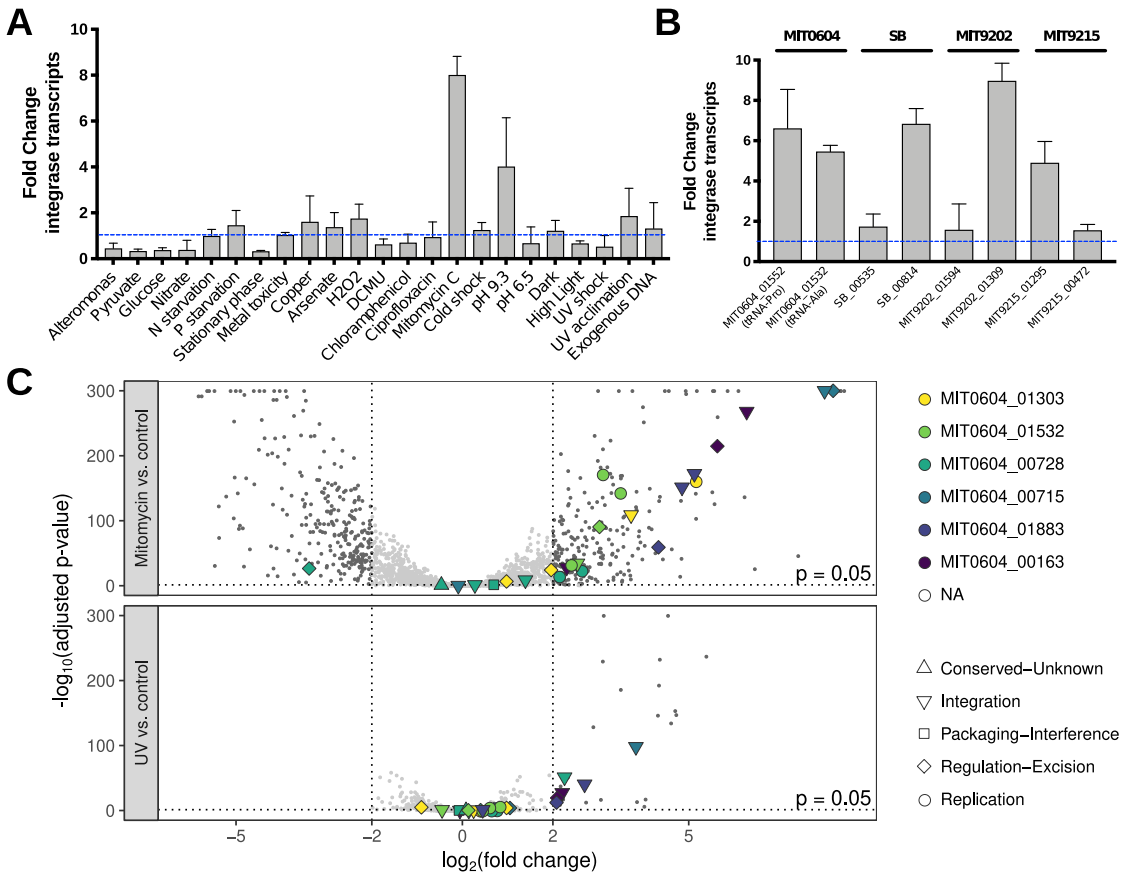


Figure S5. Transcriptional response of tycheposon hallmark genes under DNA damage stress and shock treatments, related to STAR Methods

(A) Reverse-transcription qPCR analysis of MIT0604_01303 integrase was performed on biological triplicates exposed to the treatment, compared to untreated controls. The dotted blue line indicates no change from treatment to control. See STAR Methods for details about treatments.

(B) Reverse-transcription qPCR analysis of 8 integrase genes from 4 different *Prochlorococcus* strains treated with mitomycin C. The dotted blue line indicates no change from treatment to control. Five integrases, distantly related to each other, were upregulated in response to mitomycin C, suggesting that this type of DNA damage is a general induction cue for tycheposon elements.

(C) Volcano plot showing the $\log_2$-fold change of gene expression of cultures treated with either mitomycin C (top) or a UV shock (bottom) relative to the control treatment. The y axis shows the false-discovery-rate-corrected p values as estimated by DESeq2 (Wald test using a negative binomial generalized linear model with Benjamini-Hochberg correction). Colored points indicate hallmark genes for each strain, with the shape indicating their function. Overall, the mitomycin C treatment produced a much stronger signal than UV (a phenomenon observed in other bacteria[155]) indicating that the cells are better equipped to handle UV-induced damage. The majority of tycheposon hallmark genes are strongly upregulated by mitomycin C but not UV shock. In the mitomycin C treatment, several tycheposons show co-upregulation of their hallmark genes (integrase, excisionase, and primase/helicase).

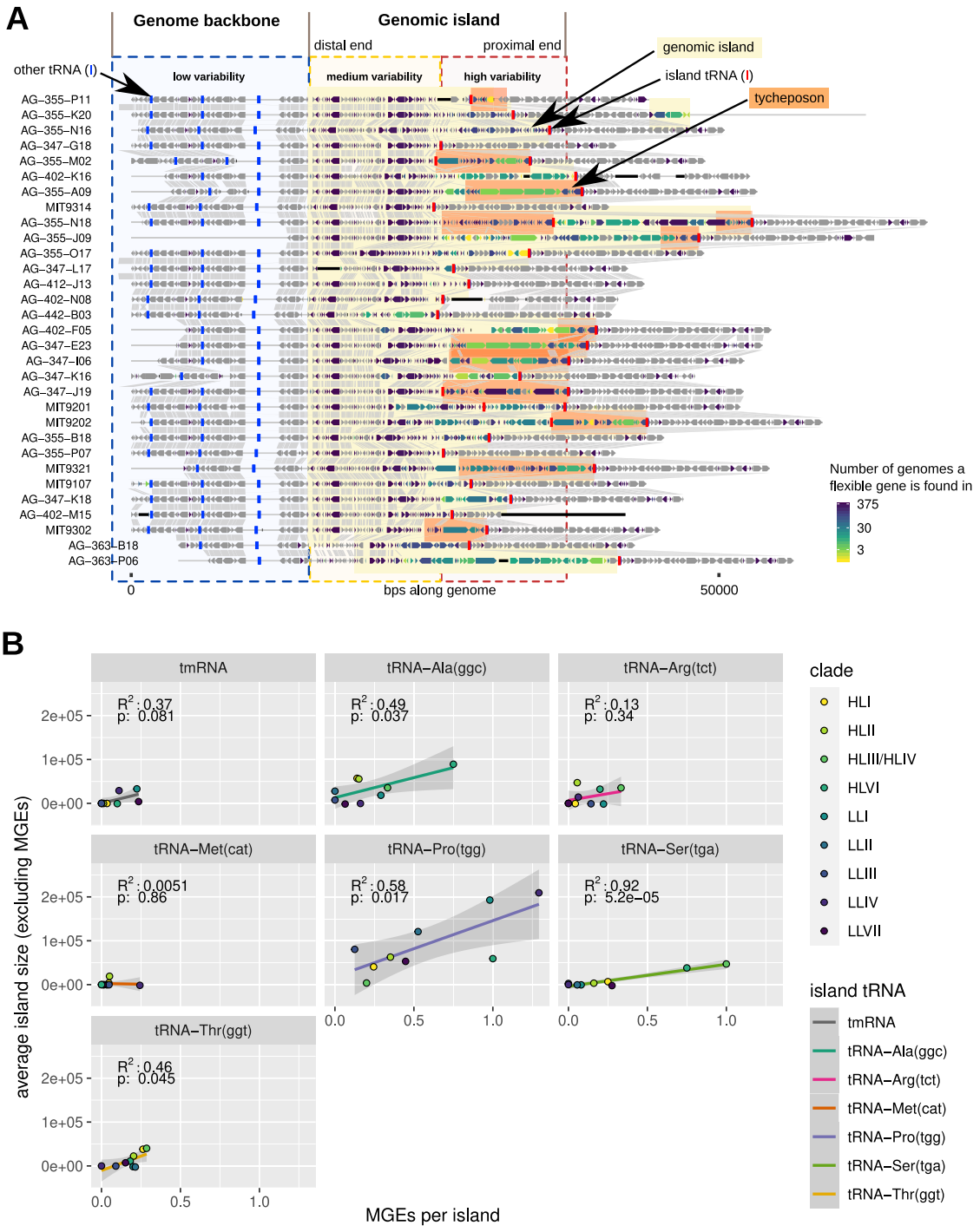


Figure S6. Increasing genomic-island content variability and genomic-island sizes linked to tycheposon activity, related to STAR Methods

(A) A gene synteny map showing the genomic regions around the threonine$_{ggt}$-island in a selection of HLII-*Prochlorococcus* strains. Each row is a genome. Filled arrows indicate protein-coding genes, colors their abundance in the set of 623 genomes (gray, core genome). Orthologous genes are linked by vertical gray bars. Red and blue vertical bars indicate tRNA genes, with the red tRNA being the integration site for mobile elements (dark yellow boxes) "visiting" the island. The three large blocks high-light regions of increasing genomic variability, from the most conserved backbone genome (blue), through the distal end of the island with medium variability (yellow, furthest from island tRNA), to the proximal end with high variability and in some cases integrated mobile elements (red, abutting Thr$_{ggt}$-tRNA gene). (B) Correlation between island sizes excluding integrated MGEs and the number of integrated MGEs per tRNA-associated island and clade. The observed positive correlation between inferred tycheposon activity (taking tycheposons per island as a proxy) and the amount of non-MGE material in islands supports the hypothesis that non-MGE island material is brought in as flanking material by tycheposons.

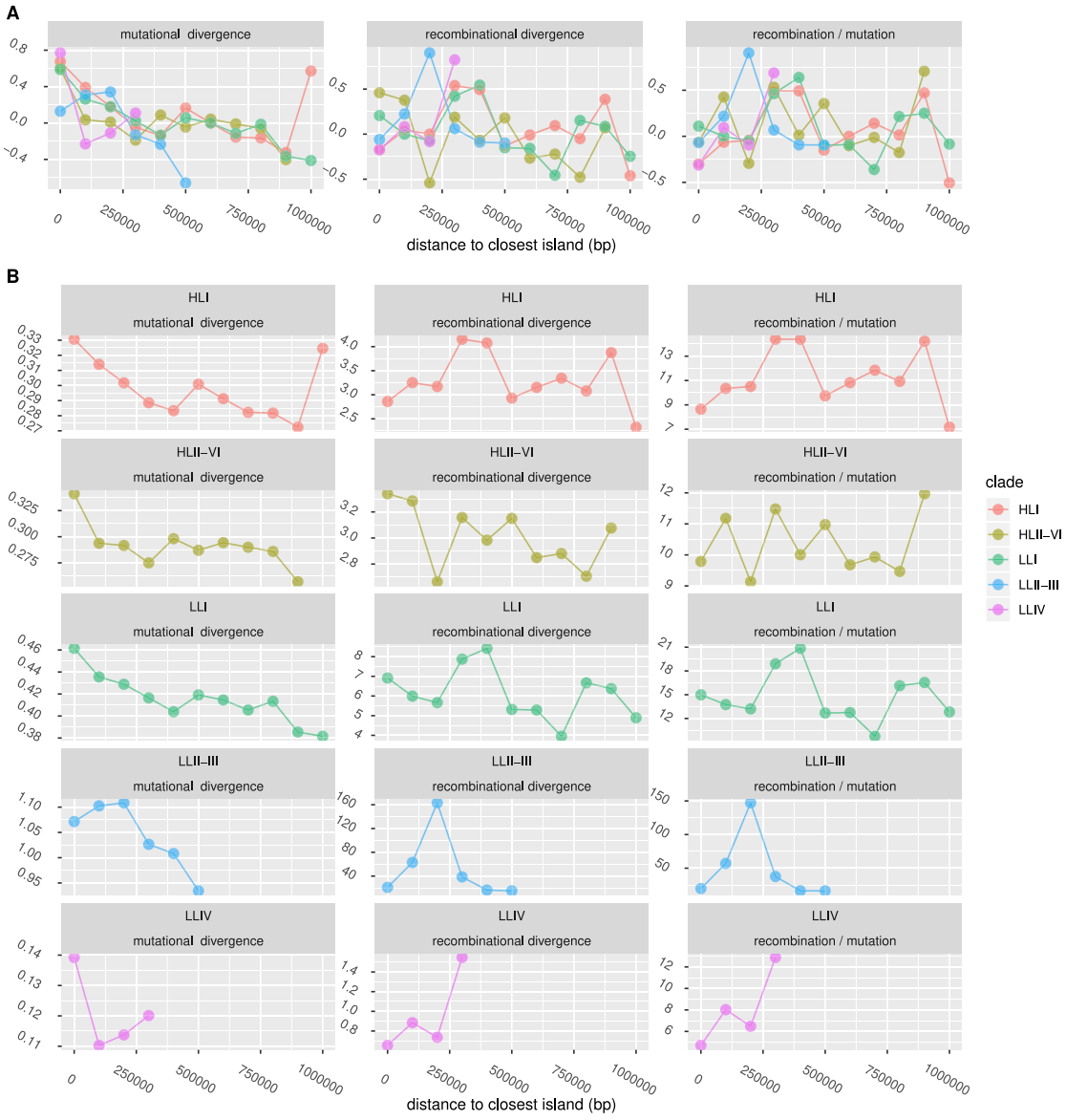


Figure S7. Homologous recombination in island-flanking and non-island-flanking genomic regions, related to STAR Methods

(A) Across-clade comparison of three recombination-related parameters—mutational divergence, recombinational divergence, and the ratio of the two—estimated using mcorr.[135] Values are derived from clade-specific estimates (see panel B) and were mean centered and rescaled for better visual comparison of trends across clades. Parameters were estimated on 99,999-bp-long partitions from a concatenated alignment of backbone genes, ordered by their proximity to genomic islands, from island-flanking genes (first partition) to genes furthest from islands (last partition).

(B) Absolute values of the same three recombination-related parameters for each individual clade. We note that in regions closest to genomic islands (first partition), the impact of recombination on divergence relative to mutation is among the lowest overall observed values, indicating that these locations are not hotspots of homologous recombination.