Title

Trichodesmium genome maintains abundant, widespread noncoding DNA in situ, despite oligotrophic lifestyle

Journal

Proceedings of the National Academy of Sciences of the United States of America, 112(14)

ISSN

0027-8424

Authors

Walworth, Nathan
Pfreundt, Ulrike
Nelson, William C
et al.

Publication Date

2015-04-07

DOI

10.1073/pnas.1422332112

Peer reviewed

Trichodesmium genome maintains abundant, widespread noncoding DNA in situ, despite oligotrophic lifestyle

Nathan Walworth[a], Ulrike Pfreundt[b], William C. Nelson[c], Tracy Mincer[d], John F. Heidelberg[a], Feixue Fu[a], John B. Waterbury[e], Tijana Glavina del Rio[f], Lynne Goodwin[f], Nikos C. Kyrpides[f], Miriam L. Land[g], Tanja Woyke[f], David A. Hutchins[a], Wolfgang R. Hess[b], and Eric A. Webb[a,1]

[a]Department of Biological Sciences, University of Southern California, Los Angeles, CA 90089; [b]Genetics and Experiment Bioinformatics, University of Freiburg, 79098 Freiburg, Germany; [c]Fundamental and Computational Sciences, Pacific Northwest National Laboratory, Richland, WA 99352; [d]Marine Chemistry and Geochemistry Department and [e]Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543; [f]Joint Genome Institute, Walnut Creek, CA 94598; and [g]Oak Ridge National Laboratory, Oak Ridge, TN 37831

Edited by Edward F. DeLong, University of Hawaii, Manoa, Honolulu, HI, and approved February 10, 2015 (received for review November 26, 2014)

Understanding the evolution of the free-living, cyanobacterial, diazotroph *Trichodesmium* is of great importance because of its critical role in oceanic biogeochemistry and primary production. Unlike the other >150 available genomes of free-living cyanobacteria, only 63.8% of the *Trichodesmium erythraeum* (strain IMS101) genome is predicted to encode protein, which is 20–25% less than the average for other cyanobacteria and nonpathogenic, free-living bacteria. We use distinctive isolates and metagenomic data to show that low coding density observed in IMS101 is a common feature of the *Trichodesmium* genus, both in culture and in situ. Transcriptome analysis indicates that 86% of the noncoding space is expressed, although the function of these transcripts is unclear. The density of noncoding, possible regulatory elements predicted in *Trichodesmium*, when normalized per intergenic kilobase, was comparable and twofold higher than that found in the gene-dense genomes of the sympatric cyanobacterial genera *Synechococcus* and *Prochlorococcus*, respectively. Conserved *Trichodesmium* noncoding RNA secondary structures were predicted between most culture and metagenomic sequences, lending support to the structural conservation. Conservation of these intergenic regions in spatiotemporally separated *Trichodesmium* populations suggests possible genus-wide selection for their maintenance. These large intergenic spacers may have developed during intervals of strong genetic drift caused by periodic blooms of a subset of genotypes, which may have reduced effective population size. Our data suggest that transposition of selfish DNA, low effective population size, and high-fidelity replication allowed the unusual "inflation" of noncoding sequence observed in *Trichodesmium* despite its oligotrophic lifestyle.

marine microbiology | oligotrophic | evolution genomics | nitrogen fixation

The low availability of N (and fixed carbon) in the midlatitude upper oceans provides an important niche for autotrophic organisms that can fix atmospheric nitrogen, which can exert control over global primary production (1–3). Nitrogen fixation is a prokaryotic process with a high-energy demand, and oceanic cyanobacteria are known to be significant sources of this "new" nitrogen (nitrogen that is fixed from the atmosphere or NO_3 advected from depth) (4, 5). Molecular field data have shown that a handful of cyanobacterial diazotrophs responsible for oligotrophic nitrogen fixation can reach relatively high cell numbers (6–12) and be of significant biogeochemical importance (13–16). These include photosynthetic free-living forms, such as the filamentous *Trichodesmium* and the unicellular *Crocosphaera* and *Cyanothece*, and photosynthetic and nonphotosynthetic symbiotic forms, such as heterocystous *Richelia* and *Candidatus* Atelocyanobacterium thalassa (3, 5, 17).

Trichodesmium cells can grow either as trichomes (i.e., filaments) or aggregates and form three types of classically described colonies, including radial puffs, vertically aligned fusiform tufts, and bowties (e.g., refs. 18–20). *Trichodesmium* spp. form blooms throughout the nitrogen-limited Atlantic and Pacific Oceans (21), as well as the Arabian and Red Seas (19, 22), and in the North Pacific Subtropical Gyre, they dominate a recurrent annual phytoplankton bloom (23). The up-to-multimillimeter-sized *Trichodesmium* colony environment can be an oasis of fixed N and C in the oligotrophic oceans (24, 25) and has been observed to contain a varied assemblage of organisms, ranging from prokaryotes and unicellular eukaryotes to juvenile copepods and decapods (26–31) with metabolisms including heterotrophs and anoxygenic and oxygenic phototrophs, as well as mixotrophic eukaryotes. Thus, in contrast to most other unicellular oligotrophs, *Trichodesmium* can either live in a colonial habitat dominated by extensive physical interactions with both sister cells and other taxa and/or as free trichomes that can constitute a significant fraction of the *Trichodesmium* water column biomass (32). These varying ecological lifestyles can partition *Trichodesmium* into different subpopulations with dynamic states including multiple morphologies, genotypes, and varied physical interactions. Furthermore, a specific population may interchange between trichome- and colony-dominated biomass.

Significance

The free-living cyanobacterium *Trichodesmium* is a major source of new nitrogen and fixed carbon to the tropical and subtropical oceans, but despite its importance, we know little about the molecular mechanisms it uses to succeed in its oligotrophic habitat. Here we show that its gene-sparse genome is littered with large, conserved, expressed intergenic spaces, which is atypical for most known free-living prokaryotes. Paradoxically, although its genome is enriched in predicted transposases and repeat sequences, it exhibits conserved intragenus synteny and similar intergenic architecture relative to its sympatric, gene-dense relatives *Prochlorococcus* and *Synechococcus*. This observation demonstrates a successful alternative to the genomic streamlining strategy observed in other free-living oligotrophs such as *Prochlorococcus* or *Pelagibacter*.

Author contributions: N.W., T.M., D.A.H., W.R.H., and E.A.W. designed research; N.W., U.P., W.C.N., T.M., J.F.H., F.F., J.B.W., T.G.d.R., L.G., N.C.K., M.L.L., T.W., and E.A.W. performed research; N.W., U.P., W.C.N., T.M., J.F.H., W.R.H., and E.A.W. analyzed data; and N.W., U.P., W.C.N., T.M., J.F.H., F.F., J.B.W., T.G.d.R., L.G., N.C.K., M.L.L., T.W., D.A.H., W.R.H., and E.A.W. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The Fastq files have been deposited into the NCBI Sequence Read Archive (SRA), www.ncbi.nlm.nih.gov/sra (accession nos. SAMN02199363 and SAMN02199364). The genomes of IMS101, 2175, and H94 have been deposited into the NCBI WGS database, www.ncbi.nlm.nih.gov (accession nos. SAMN02598485, SAMN03421191, and SAMN03421272.

[1]To whom correspondence should be addressed. Email: eawebb@usc.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1422332112/-/DCSupplemental.

Trichodesmium erythraeum IMS 101 (hereafter IMS101) was isolated in 1991 by Prufert-Bebout et al. from the Gulf Stream off the coast of North Carolina (33), and since then a handful of other strains have been isolated by Waterbury and others (20, 34, 35); however, virtually nothing is known about the similarity of its unusual genome architecture to the genome structure of natural populations. Additionally, efforts to cryopreserve *Trichodesmium* have been unsuccessful, which presents a challenge due to the requirement of constant culturing and maintenance.

In recent years, numerous studies on unicellular, oligotrophic microbes have observed relatively high coding percentages and genomic streamlining in sympatric, nondiazotrophic cyanobacteria (e.g., refs. 36–41) as well as other dominant, marine heterotrophic bacteria (e.g., refs. 42–44). In these aforementioned studies, streamlining has been used to describe both gene loss and general genome reduction primarily as a result of selection for greater metabolic efficiency in nutrient-poor regimes rather than mainly deriving from population-level processes. Accordingly, population-level processes suggest that the reduced impact of genetic drift as a result of elevated effective population sizes exhibited by free-living microbial taxa is sufficient to allow for selection against accumulation of excess DNA within a population (45, 46). This latter population-level stance argues that microbial genome reduction has been primarily a result of weakened nonadaptive forces rather than any one strong selective force increasing metabolic efficiency.

Hence, fitting with a small predicted effective population size as well as the absence of nitrogen limitation, many cyanobacterial diazotrophs (e.g., *Crocosphaera, Cyanothece*, *Cylindrospermopsis,* etc.) do not have the conventionally defined streamlined genomes seen in other marine taxa (40, 42, 45–47). Similar to IMS101, these genomes are enriched in predicted insertion sequences, repeats, and regulatory proteins (41, 48, 49), yet despite this fact still retain much larger coding percentages (≥80%) than *Trichodesmium*. Until now, the low gene density and large intergenic spacers have only been observed in the genome of IMS101 that has been maintained in culture for approximately two decades. Here, we explore the degree of genome architecture conservation between spatiotemporally segregated *Trichodesmium* isolates as well as natural populations.

A defining feature of *Trichodesmium* ecology involves populations with dynamic, multicellular morphotypes (e.g., single trichomes or different colony types), where one of these forms may dominate the population (including blooms) at different times (e.g., refs. 19, 31, and 32). Therefore, we hypothesize that high transposon load coupled to periodic nonadaptive, bloom-driven reductions in effective population size from a subset of morphological genotypes harboring different epibiotic consortia may have substantially contributed to production and maintenance of large intergenic regions, proliferation of repetitive DNA, and subsequent selection for noncoding regulatory regions. This trajectory may have ultimately allowed *Trichodesmium* to evolve one of most intergenic-rich genomes of any free-living prokaryote.

Results and Discussion

IMS101 Genome Does Not Exhibit Streamlining. Genomic streamlining has been associated with vitamin and amino acid auxotrophies, simplified carbon and nitrogen metabolism, a low abundance of pseudogenes, limited motility, and little to no selfish DNA (e.g., insertion sequences, transposons, etc.) (36–38, 42, 50–52). Compared with four sympatric reference marine picoplankton genomes (*Synechococcus* WH8102 and CC9311 and *Prochlorococcus* SS120 and MED4), the IMS101 genome displays characteristics that are inconsistent with streamlining. IMS101 has a large suite of regulatory proteins (COG K, transcriptional regulators subgroup; 141 vs. 55 average of the four genomes), motility-related proteins (COG N; 56 vs. 6.5 average), increased number of transport-related proteins (COG P; 125 vs. 75.25 average), signal transduction proteins (COG T; 146 vs. 41.75), a large number of transposase sequences (165 vs. zero), and numerous pseudogenes (625 vs. 3). Normalizing the number of these genes in IMS101 either by genome size (IMS101 is 3.7 times greater than the picoplankton average) or gene count (IMS101 is 2.2 times greater than the picoplankton average) shows that regulators and transporters roughly scaled with the increases in size, while transposases, motility genes, signal transducers, and pseudogenes (Figs. S1 and S2) were enriched in IMS101. These data suggest that unlike the recently described photofermentative, cyanobacterial symbiont, *Candidatus* Atelocyanobacterium thalassa (17, 38), and marine *Synechococcus* and *Prochlorococcus,* the metabolism predicted for *Trichodesmium* is not minimized.

Based on Integrated Microbial Genomes (IMG) COG analysis, *Crocosphaera* (strains WH8501, WH0003, and WH0401, average values reported below) and *Trichodesmium* retain similar numbers of signal transduction proteins (131 vs. 146) and transporter-related proteins (125 vs. 110), whereas *Crocosphaera* possesses slightly more transcriptional regulator proteins (29 vs. 13) and motility proteins (33 vs. 29) (Dataset S1). In terms of gene content, *Crocosphaera* does not necessarily exhibit a streamlined profile but its coding percentage is still that of an average free-living prokaryote (~75–80%), whereas *Trichodesmium* also retains similar protein content to *Crocosphaera* but has a substantially reduced coding percentage.

Accordingly, the IMS101 genome encodes 5,076 proteins (per the IMG annotation https://img.jgi.doe.gov/), yielding a coding percentage of ~60%, whereas its sympatric, picoplanktonic cousins *Prochlorococcus*, marine *Synechococcus*, *Crocosphaera watsonii*, and *Cyanothece* (41, 48, 53, 54), and all of the 45 currently sequenced members of the Oscillatoriales (the cyanobacterial order with which IMS101 is phylogenetically placed) have coding percentages >75%. The Oscillatoriales demonstrates variation both in gene count and genome size (average gene number = 5,663.6 ± 1,150.1; average genome size = 6,346,139 ± 1,207,411), and although the number of genes in IMS101 is within this range, the coding percentage and the genome size are at opposite ends of the spectrum, respectively (Dataset S1). Furthermore, principal component analysis (PCA) of cyanobacterial genome features segregates *Trichodesmium* from the rest of the taxa based on its substantially lower coding percentage relative to other characteristics (Fig. S1 and Dataset S1). Similarly, a PCA including only cyanobacteria that possess annotated transposases also shows *Trichodesmium* segregating away from other genomes opposite the axis of coding percentage, whereas *C. watsonii* WH 8501 segregates away from others based on the 1,000+ annotated transposases in its genome unlike other sequenced *Crocosphaera* genomes (41) (Fig. S1*B*). These data imply that, although IMS101 has a "normal" number of genes for a filamentous, diazotrophic cyanobacterium, its noncoding space is unique.

Long Intergenic Regions Are Conserved in Trichodesmium. IMS101 was in culture for >10 y (33) before genome sequencing was initiated in 2003. To determine whether the unusual genomic characteristics observed in IMS101 were in common in the *T. erythraeum* species, we generated draft genome sequence from a more recently isolated strain of *T. erythraeum* [strain 21–75 (2175) isolated from the Tropical Atlantic in 6/2006] that was only in culture for ~1 y before sequencing (Datasets S2 and S3). Both *T. erythraeum* strains have relatively large genomes (7–7.78 Mbps), low GC content (~33–34%), and a reduced protein-coding percentage (~61%). We also obtained a partial genome sequence from *T. thiebautii* H94, a Hawaiian isolate from 2004 that is representative of the other major *Trichodesmium* clade currently in culture (20), and although this 2009 sequencing run returned low coverage of the H94 genome, the contigs that assembled showed a similarly low coding percentage of ~61% (Dataset S4). Thus, from this limited genomic analysis of three cultured isolates, it appears that the low coding percentage and large genome observed in IMS101 are commonplace in the genus.

High levels of synteny were also observed between the two *T. erythraeum* isolates, even through the long, noncoding intergenic regions. According to a MAUVE alignment, the IMS101 and 2175 genomes contain 28 colinear blocks ranging in size from 1,700 bp up to ~2.5 Mbp (Fig. 1*A*).

Additionally, although 2175 is a draft genome, we were able to recover almost all (98%) IMS101 protein-encoding and intergenic regions (96%) using BLASTn (Dataset S5). Homologous intergenic regions between IMS101 and 2175 averaged 510 bp in length with a median of 380 bp. In contrast, the intergenic regions unique to each strain were considerably shorter, with an average length of 53 bp and a median of 21 bp. These results suggest that the unusually long intergenic regions may confer some sort of selectable advantage that ensures their conservation or, conversely, a robust maintenance mechanism that does not act upon the shorter regions.

To place these noncoding characteristics in the context of other bacteria across a broad range of representative phylogenetic origins (55), we performed PCA on multiple genomic features (Dataset S1), including the relative size and distribution of intergenic regions. Fig. 1*B* shows both *Trichodesmium* genomes segregating along the axis associated with the median intergenic spacer length, suggesting that maintenance or accumulation of noncoding DNA is distributed throughout the genome rather than confined to a few intergenic regions. This evidence contrasts with previous observations of predominant regulatory protein accumulation in free-living bacteria rather than noncoding DNA (47, 55, 56), and it further emphasizes the unique size and distribution of *Trichodesmium* intergenic regions relative to other genome features in bacteria.

Environmental Populations of Trichodesmium Have Low Gene Density. To extend our analysis of *Trichodesmium* genome structure to environmental populations, a colony-enriched metagenome was generated from *Trichodesmium* colonies in the North Atlantic Subtropical Gyre. Fragment recruitment of metagenomic reads against IMS101 shows nearly complete coverage in mixed natural *Trichodesmium* populations (Fig. 1*C*, *Upper*). Although many of the assembled metagenomic contigs (n = 460,494) were relatively small (N50 = 1,217), a considerable amount of larger contigs was generated (n = 1,032; N50 = 4,335; max length = 12,688) as well. When these larger contigs are mapped to the IMS101 scaffold using nucmer, the subsequent alignment plot strongly suggests genome synteny between IMS101 and natural populations in situ (Fig. 1*C*, *Lower*). Furthermore, the metagenomic dataset contained ~94% of the IMS101 and 2175 intergenic sequences. Undetected intergenic sequences for both IMS101 and 2175 again had small average intergenic lengths of 85 and 57 bp and medians of 27.5 and 21 bp, respectively. The near-complete in situ detection of each genome's longer intergenic sequences in contrast to the short averages and medians of the undetected intergenic sequences further suggests that in natural populations, longer intergenic regions may be selectively maintained, lending evidence to their potential physiological importance to the in situ ecology of *Trichodesmium* spp.

Intergenic Regions and Repetitive Elements. The intergenic regions of the IMS101 genome contain numerous DNA repeats, ranging from very small noncoding elements [e.g., highly interspersed palindromic sequences (35, 57) or other repeating sequences (58, 59)] to larger, gene-encoding insertion sequences (e.g., ref. 60). Because many of the repeating elements can overlap and/or be nested inside of each other, we assessed the contribution of intergenic repeats to the total intergenic space of 2,801,094 bp (*SI Materials and Methods*). We counted an intergenic sequence as a repeat if it occurred two or more times in the genome. Hence, when comparing IMS101 intergenic regions against the IMS101 scaffold by using BLASTn and summing the length of each repeat sequence hit, it yielded ~4.1 Mb of nested overlapping repeat sequence. However, when these nested repeats were consolidated into discrete nonoverlapping repeat sequences, only one-third of IMS101 noncoding DNA (804,807 bp) consisted of repetitive elements, or ~10% of the total genome (*SI Materials and Methods*). Similarly, approximately one-third of 2175 noncoding DNA (799,980 of 2,708,763 bp of total intergenic space) consisted of nonoverlapping repetitive elements

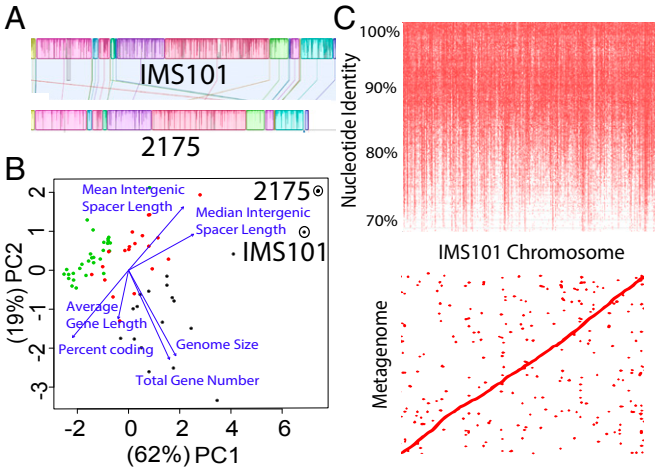

Fig. 1. Differences between *Trichodesmium* genome content and other bacteria and conservation of *Trichodesmium* genome both in culture and in situ. (*A*) MAUVE alignment showing high synteny between the finished IMS101 genome and the 2175 draft genome. (*B*) PCA on multiple genomic features segregates IMS101 and 2175 along the median intergenic spacer length axis relative to small (<2 Mb, green), medium (2–4 Mb, red), and large (>4 Mb, black) bacterial genomes (55), indicating that noncoding sequence accumulation is distributed throughout the genome. (*C*) *Upper* is a fragment recruitment map of enriched *Trichodesmium* metagenomic reads from the North Atlantic Subtropical Gyre mapped to the IMS101 genome. *Lower* is a mummer plot generated from a nucmer alignment between the IMS101 chromosome and the large metagenomic contigs (N50 = 4,335; max = 12,688), suggesting genome synteny of the IMS101 genome in natural populations (red diagonal) along with repeats (dots) in different positions around the genome.

(11.5% of the 2175 genome). Some of these repetitive elements are likely to be mobile, because 58% of the 625 IMS101 pseudogenes are interrupted by a repeat sequence. Regardless of the function of these repeats, their conservation across time and space from bottlenecked isolates to natural communities strongly suggests that selection for maintenance of these elements exists within this genus.

The distribution of repeats around a genome can give insights into the mechanism by which they propagate [i.e., DNA replication slippage for tandem and transposition/recombination for distributed repeats, respectively (58, 61, 62)]. Here, we compared the distribution of putative transposase genes and insertion sequences in the IMS101 genome to the top 10 intergenic regions containing the most abundant repeats identified in our pipeline (Table S1). Although the predicted insertion sequences are distributed around the genome, areas of increased density were apparent at ~3, 8, 9, and 12 o'clock on the genome (Fig. 2). Focusing on the ~7.49- to 7.54-Mbp regional cluster, we used BLASTn to identify locations containing numerous overlapping sequence elements, including insertion sequences, predicted transposase-related genes, and repetitive elements (Fig. 2, *Inset*), suggesting that this region may be a recombination "hot spot," with both DNA polymerase slippage and transposition causing genetic elements to be stacked on top of each other. Further, we identified sequences from this region in a publicly available metatranscriptome containing *Trichodesmium* colonies that is geospatially distinct from our metagenome samples (South Pacific vs. North Atlantic, respectively) (27). These results either suggest that this genomic region is generally conserved and active in the genus, or at least that the single copy elements comprising the region in IMS101 are active and conserved at high identity in other members of the genus, even if the arrangement observed in region 7.49–7.54 is not.

In a gene-centric study comparing *Crocosphaera* genomes (41), it was observed that most strains did not contain highly repetitive ORFs, with the exception of *Crocosphaera watsonii* WH 8501,

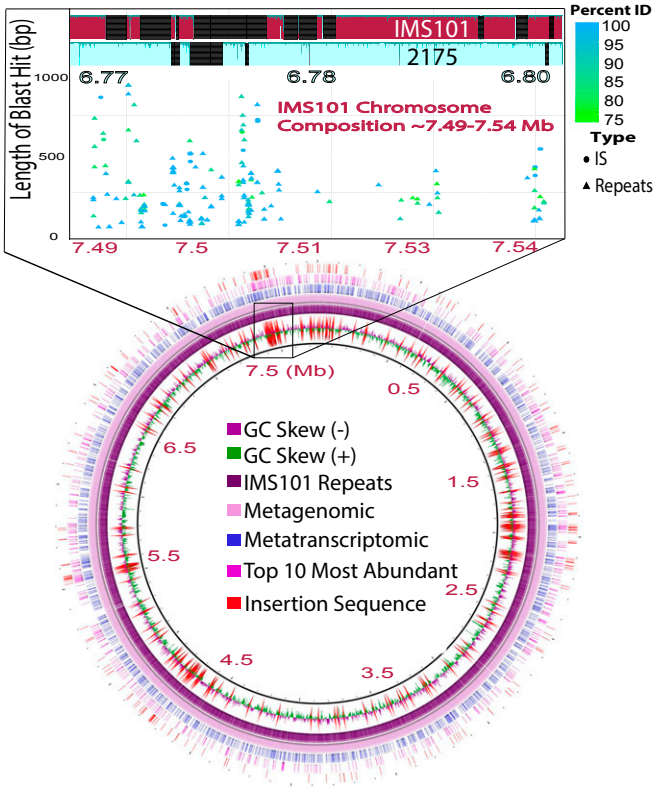

Fig. 2. Genomic map of the 10 intergenic regions possessing the most repeated sequences in the IMS101 genome. These IMS101 intergenic regions were mapped to the *Trichodesmium* metagenome and publicly available metatranscriptomes. For the circular map, going from out to in are insertion sequences (IS) (labeled red), the top 10 most abundant intergenic repeats, metatranscriptomic reads, metagenomic reads, IMS101 repeats, and GC skew with insertion sequence locations overlaid onto it shown in red. The scatterplot shows the IS and repeat composition of a repeat hot spot between ~7.49 and 7.54 Mb. *Inset* is the IMS101 genome alignment of this hotspot with 2175 showing nonhomologous (shaded boxes) regions between the two chromosomes aligning over the segment containing numerous repeats and IS sites.

and only WH 8501 had substantially more annotated transposases than the rest of the *Crocosphaera* genomes. This evidence is also seen in the PCA analysis including transposase and paralog numbers per genome (Fig. S1*B*), in which no *Crocosphaera* genomes segregate along the "Paralogs" axis. *C. watsonii* WH 8501 segregates away from other genomes along the "Transposase" axis, whereas the other *Crocosphaera* genomes remain more tightly clustered within the plot. Hence, although WH 8501 possesses a greatly enhanced transposon load relative to *Trichodesmium*, it did not develop the large noncoding regions observed in *Trichodesmium*.

Intergenic Regions and ncRNA Elements. In the absence of coding genes, it is possible that noncoding structural RNAs, regulatory RNAs, or ribozymes within the intergenic regions are the selectable traits driving conservation of these regions, which have been recognized as important components in cyanobacterial expression networks (63). Six known structural RNA elements were identified (using RFAM) within IMS101 intergenic regions (Table S2). We also used a pipeline that has previously identified cyanobacterial noncoding RNAs (ncRNAs) (63, 64) to look for *Trichodesmium* structural conservation between intergenic regions among IMS101 and 2175 and their in situ counterparts within metagenomic reads (*SI Materials and Methods*). For an in situ sympatric comparison, the same procedure was done with either *Prochlorococcus* ($n = 5$) or *Synechococcus* ($n = 6$) genomes, along with all of the assembled Sargasso Sea sequences from the Global Ocean Sampling dataset (65). Although it would have been informative to run this pipeline with *Crocosphaera* genomes, we feel our results would not have been comparable for several reasons (*SI Text*). However, of the 4,033 predicted ncRNAs in *Trichodesmium* (see below), only 0.6% were detected in a *Crocosphaera* metatranscriptome (66), as well as *Crocosphaera* and *Cyanothece* genomes using BLASTn (Dataset S6). This comparison suggests that much of the development and expansion of noncoding sequences in *Trichodesmium* may be uniquely specific to the evolution of the genus rather than shared with sympatric diazotrophs that also do not display streamlining.

The pipeline predicted 365 putative noncoding regulatory elements in *Synechococcus* environmental sequences and 190 in *Prochlorococcus* (Fig. 3*A*). This trend is consistent with previous studies showing widespread *Prochlorococcus* streamlining relative to *Synechococcus* evidenced in both publicly available genomes (67–69) and environmental single amplified genomes (40). From the *Trichodesmium* cultures and metagenome, the pipeline predicted 4,033 nonredundant, noncoding regulatory elements (Fig. 3*A*), of which 3,027 (75%) were expressed in the IMS101 transcriptome (Dataset S6) (see below). When putative regulatory elements were normalized per intergenic kilobase (Kb), *Trichodesmium* and *Synechococcus* yielded ~1.5 per intergenic Kb (Fig. 3*A*, line), whereas *Prochlorococcus* yielded <1 regulatory element per intergenic Kb. Although these predictions also include possible regulatory untranslated regions (UTRs), riboswitches, and terminator sequences, the results show evidence for generally conserved intergenic composition among all three sympatric cyanobacteria, with *Trichodesmium* and *Synechococcus* possibly possessing slightly more regulatory elements per intergenic Kb.

Expression of Transposases and Intergenic Sequence. To determine global, annotated transposase, and intergenic expression patterns, transcriptomes of biological duplicate IMS101 cultures growing semicontinuously in Aquil medium were sampled near the middle of the photoperiod and sequenced by using the Illumina Hi-Seq platform. This analysis showed that ~86% of the intergenic regions in *Trichodesmium*, ~91% of the IMG annotated transposases, and 75% of the 4,033 predicted regulatory elements from the above pipeline were expressed (Datasets S6 and S7). In a laboratory-based study using a different approach, directed toward the identification of transcriptional start sites, it was determined that at least 18.2% of the intergenic space was transcribed as either ncRNAs or 5′ UTRs of protein-coding mRNAs (70). This analysis further revealed that, of all bacteria examined to date, *T. erythraeum* has the highest percentage of transcriptional start sites from which ncRNAs originate. The combination of such widespread noncoding and transposase expression as well as conservation of these sequences across isolates

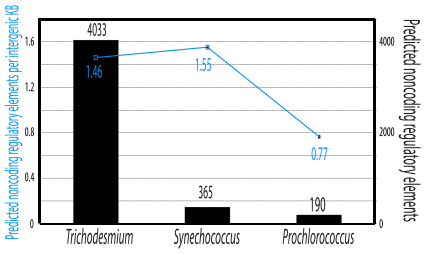

Fig. 3. Predicted noncoding structurally conserved elements via comparative genomics pipeline. The bar graph shows the increased amounts of in situ putative conserved elements in *Trichodesmium* relative to sympatric *Synechococcus* and *Prochlorococcus*. The blue line indicates the relative amount of predicted elements per intergenic Kb of sequence among the three cyanobacteria. Although many more structurally conserved elements are predicted in *Trichodesmium* relative to *Prochlorococcus* and *Synechococcus*, *Trichodesmium* and *Synechococcus* retain similar frequencies of structurally conserved elements relative to intergenic kilobase (KB), whereas *Prochlorococcus* retains approximately half.

and natural populations suggest that possible widespread RNA-based regulation along with active transposition may be commonplace in the genus.

Although the combination of structural predictions and intergenic RNA sequencing lends strong support to active, widely distributed ncRNAs in *Trichodesmium*, it is still difficult to determine whether expressed portions of intergenic regions are discrete ncRNAs or part of expressed UTRs in mRNA transcripts, or both. To corroborate both our sequencing data and informatic predictions, 18 noncoding elements with consistently strong Illumina expression profiles (Dataset S6) between the biological replicates were chosen for Northern blot analysis, and all yielded positive hybridizations (Fig. S3 and Tables S3 and S4). To determine the degree of secondary structural conservation between culture-derived ncRNA sequences and their counterparts in the metagenome and metatranscriptome, we used RNAfold to compare computationally predicted structures (71–74). RNAfold predicted very similar core secondary structures between the culture and environmental sequences for most ncRNAs among the top matches (Fig. S4), with several variations due to shorter/larger loops and hairpins. These conserved features may be selectable traits that drive the conservation of the long intergenic regions in globally distributed *Trichodesmium* populations.

Population Level Processes vs. Natural Selection. Lynch and Conery propose that microbial genomes are streamlined primarily because their effective population sizes are generally large enough to prevent significant colonization of mobile elements and noncoding sequences, whereas effective population sizes in multicellular eukaryotes are low enough to allow a permissive environment for the expansion of noncoding DNA (51). The abundant noncoding sequences in *Trichodesmium* relative to most other free-living bacteria and marine oligotrophs, along with its general genome architecture, may be due to a combination of small effective population size derived from differing morphological genetic subpopulations with varying associated epibionts (e.g., ref. 28), as well as potential rampant active mobile elements via transposase activity. This feature is noteworthy because, unlike *Trichodesmium*, other bloom-forming cyanobacteria with many repetitive sequences such as *Microcystis* have an ∼80% coding average, which suggests that other prominent forces are influencing *Trichodesmium* genome evolution in combination with reductions in effective population size (75). Hence, the absolute causes of the large, intergenic-rich genomes observed relative to other free-living prokaryotes (76) and marine oligotrophs within the same habitat (40) remain obscure.

Because it is thought that many bacteria are deletion-biased (47, 77), stable maintenance of these elements from laboratory isolates to the natural samples suggest that they may be required in some fashion for growth both in culture and in situ. It has been shown in numerous systems that repeating elements (repeats and/or IS; Dataset S8) can be mediators of genomic plasticity (61, 62, 78); however, the direct impacts of these repeats are not always so clear. For example, high IS density in the genome of *Lactobacillus acidophilus* has been described (79), and despite the propensity of these elements to inactivate genes and facilitate recombination of genomic structure (61), the genome of this isolate still displays high levels of synteny with other sequenced Lactobacilli. Because it has also been shown that partial IS sequences can inhibit transposition (78, 80), it is possible that these repeats/pseudogenes have not been deleted because they are controlling transposition in the transposase-heavy IMS101. Others have hypothesized that the conserved repeat structures observed in some bacteria could function as recombination-dependent "promoter banks" for adaptation to new conditions, thereby allowing relatively quick "rewiring" of metabolism in subpopulations (59, 62, 81).

Summary

This study highlights a previously unidentified, environmentally conserved genomic architecture of a successful oligotrophic, free-living cyanobacterial diazotroph that is biogeochemically important across global oceanic regimes (3, 5, 24). Free-living, cyanobacterial diazotrophs such as *Crocosphaera* and *Trichodesmium* contain a wealth of transposases, chemotaxis, signal transduction, and pseudogenes that directly contradict the genome streamlining observed in other oligotrophic prokaryotic genomes. Hence, because of these gene content commonalities among some cyanobacterial diazotrophs, it is very peculiar that such low coding percentage and gene density has persisted genus-wide in *Trichodesmium* populations, both in culture and in situ. One possible explanation is that the intergenic regions experience gradual inflation during certain evolutionary intervals characterized by bloom-driven selective sweeps. Additionally, a central difference in *Trichodesmium* spp. oligotrophic ecology includes periodic aggregate formation, with possibly varying physical interaction with epibiotic prokaryotes and eukaryotes. Although specific causal factors contributing to the unusual IMS101 genome still remain unclear, these data do confirm the environmental relevance of the *Trichodesmium* genome architecture, as well as a nonstreamlined, alternative route to a free-living oligotrophic lifestyle.

Materials and Methods

See *SI Materials and Methods* for logistical protocols. In brief, DNA from batch *Trichodesmium* cultures was isolated, frozen at −20 °C, and processed for sequencing at the Joint Genome Institute (JGI) (IMS101) or at the University of Southern California (USC) (2175 and H94), and annotation was performed using the JGI genome annotation pipeline. *Trichodesmium* metagenome samples were collected under nonbloom conditions in October 2010 on the R/V *Oceanus* cruise number OC469-1 near the Bermuda Atlantic Time Series (BATS) station (28°37.474 N, 66°0.606 W). Colonies and trichomes were gently picked, and colony DNA was extracted immediately, stored at −20 °C, and shipped to the JGI for pyrosequencing. IMS101 genes and intergenic regions were downloaded from https://img.jgi.doe.gov and were used for fragment recruitment plots, comparative genomics, and principal component analyses. RNA was isolated from flash-frozen biological duplicates of IMS101 cultures growing semi-continuously and sequenced at the USC Epigenome Center. Northern blots and structural sequence predictions were conducted as previously described (see *SI Materials and Methods*).

ACKNOWLEDGMENTS. We thank Frank Larimer, Jill Sohm, Suzanne Edmands, Michael Lee, Christopher Dupont, and Andrew Allen for insightful discussions. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the Department of Energy under Contracts DE-AC02-05CH11231 and DE-AC03-76SF00098. Other portions of this work were supported by National Science Foundation Grant OCE-1260490 and the University of Southern California.

1. Falkowski PG, Barber RT, Smetacek V (1998) Biogeochemical controls and feedbacks on ocean primary production. *Science* 281(5374):200–207.
2. Moore CM, et al. (2013) Processes and patterns of oceanic nutrient limitation. *Nat Geosci* 6(9):701–710.
3. Sohm JA, Webb EA, Capone DG (2011) Emerging patterns of marine nitrogen fixation. *Nat Rev Microbiol* 9(7):499–508.
4. Sohm JA, et al. (2011) Nitrogen fixation in the South Atlantic Gyre and the Benguela Upwelling System. *Geophys Res Lett* 38(16):L16608.
5. Zehr JP (2011) Nitrogen fixation by marine cyanobacteria. *Trends Microbiol* 19(4):162–173.
6. Foster RA, et al. (2007) Influence of the Amazon River plume on distributions of free-living and symbiotic cyanobacteria in the western tropical north Atlantic Ocean. *Limnol Oceanogr* 52(2):517–532.
7. Foster RA, Subramaniam A, Zehr JP (2009) Distribution and activity of diazotrophs in the Eastern Equatorial Atlantic. *Environ Microbiol* 11(4):741–750.
8. Foster RA, Zehr JP (2006) Characterization of diatom-cyanobacteria symbioses on the basis of *nifH, hetR* and 16S rRNA sequences. *Environ Microbiol* 8(11):1913–1925.
9. Moisander PH, et al. (2010) Unicellular cyanobacterial distributions broaden the oceanic N_2 fixation domain. *Science* 327(5972):1512–1514.
10. Church MJ, et al. (2009) Physical forcing of nitrogen fixation and diazotroph community structure in the North Pacific subtropical gyre. *Global Biogeochem Cycles* 23(2):GB2020.
11. Church M, Björkman K, Karl D, Saito M, Zehr J (2008) Regional distributions of nitrogen-fixing bacteria in the Pacific Ocean. *Limnol Oceanogr* 53(1):63–77.
12. Langlois RJ, Hümmer D, LaRoche J (2008) Abundances and distributions of the dominant *nifH* phylotypes in the Northern Atlantic Ocean. *Appl Environ Microbiol* 74(6):1922–1931.
13. Montoya JP, et al. (2004) High rates of N_2 fixation by unicellular diazotrophs in the oligotrophic Pacific Ocean. *Nature* 430(7003):1027–1032.

14. Karl DM, Church MJ, Dore JE, Letelier RM, Mahaffey C (2012) Predictable and efficient carbon sequestration in the North Pacific Ocean supported by symbiotic nitrogen fixation. *Proc Natl Acad Sci USA* 109(6):1842–1849.
15. Subramaniam A, et al. (2008) Amazon River enhances diazotrophy and carbon sequestration in the tropical North Atlantic Ocean. *Proc Natl Acad Sci USA* 105(30): 10460–10465.
16. Großkopf T, et al. (2012) Doubling of marine dinitrogen-fixation rates based on direct measurements. *Nature* 488(7411):361–364.
17. Thompson AW, et al. (2012) Unicellular cyanobacterium symbiotic with a single-celled eukaryotic alga. *Science* 337(6101):1546–1550.
18. Webb EA, Jakuba RW, Moffett JW, Dyhrman ST (2007) Molecular assessment of phosphorus and iron physiology in *Trichodesmium* populations from the western Central and western South Atlantic. *Limnol Oceanogr* 52:2221–2232.
19. Post AF, et al. (2002) Spatial and temporal distribution of *Trichodesmium* spp. in the stratified Gulf of Aqaba, Red Sea. *Mar Ecol Prog Ser* 239:241–250.
20. Hynes AM, Webb EA, Doney SC, Waterbury JB (2012) Comparison of cultured *Trichodesmium* (Cyanophyceae) with species characterized from the field. *J Phycol* 48: 196–210.
21. Letelier R, Karl D (1996) Role of Trichodesmium spp in the productivity of the subtropical North Pacific Ocean. *Mar Ecol Prog Ser* 133:263–273.
22. Capone DG, et al. (1998) An extensive bloom of the N_2-fixing cyanobacterium *Trichodesmium erythraeum* in the central Arabian Sea. *Mar Ecol Prog Ser* 172:281–292.
23. Dore JE, Letelier RM, Church MJ, Lukas R (2008) Summer phytoplankton blooms in the oligotrophic North Pacific Subtropical Gyre: Historical perspective and recent observations. *Prog Oceanogr* 76(6):2–38.
24. Capone D, Zehr J, Paerl H, Bergman B, Carpenter E (1997) *Trichodesmium*, a globally significant marine cyanobacterium. *Science* 276(5316):1221.
25. Mulholland M (2007) The fate of nitrogen fixed by diazotrophs in the ocean. *Biogeosciences* 4:37–51.
26. Sheridan C, Steinberg D, Kling G (2002) The microbial and metazoan community associated with colonies of Trichodesmium spp: A quantitative survey. *J Plankton Res* 24: 913–922.
27. Hewson I, et al. (2009) Microbial community gene expression within colonies of the diazotroph, *Trichodesmium*, from the Southwest Pacific Ocean. *ISME J* 3(11):1286–1300.
28. Hmelo LR, Van Mooy B, Mincer TJ (2012) Characterization of bacterial epibionts on the cyanobacterium *Trichodesmium*. *Aquat Microb Ecol* 67:1–14.
29. Wyman M, Hodgson S, Bird C (2013) Denitrifying alphaproteobacteria from the Arabian Sea that express nosZ, the gene encoding nitrous oxide reductase, in oxic and suboxic waters. *Appl Environ Microbiol* 79(8):2670–2681.
30. Paerl HW, Bebout BM, Prufert LE (1989) Bacterial associations with marine oscillatoria Sp. (*Trichodesmium* sp.) populations: Ecophysiological implications. *J Phycol* 25:773–784.
31. Hynes AM, Chappell PD, Dyhrman ST, Doney SC, Webb EA (2009) Cross-basin comparison of phosphorus stress and nitrogen fixation in *Trichodesmium*. *Limnol Oceanogr* 54:1438–1448.
32. Orcutt KM, et al. (2001) A seasonal study of the significance of N2 fixation by Trichodesmium spp. at the Bermuda Atlantic Time-series Study (BATS) site. *Deep Sea Res Part II Top Stud Oceanogr* 48:1583–1608.
33. Prufert-Bebout L, Paerl HW, Lassen C (1993) Growth, nitrogen fixation, and spectral attenuation in cultivated trichodesmium species. *Appl Environ Microbiol* 59(5): 1367–1375.
34. Bell P, et al. (2005) Laboratory culture studies of *Trichodesmium* isolated from the great Barrier Reef Lagoon, Australia. *Hydrobiologia* 532:9–21.
35. Orcutt KM, et al. (2002) Characterization of *Trichodesmium* spp. by genetic techniques. *Appl Environ Microbiol* 68(5):2236–2245.
36. Palenik B, et al. (2003) The genome of a motile marine *Synechococcus*. *Nature* 424(6952):1037–1042.
37. Rocap G, et al. (2003) Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 424(6952):1042–1047.
38. Tripp HJ, et al. (2010) Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 464(7285):90–94.
39. Scanlan DJ, et al. (2009) Ecological genomics of marine picocyanobacteria. *Microbiol Mol Biol Rev* 73(2):249–299.
40. Swan BK, et al. (2013) Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci USA* 110(28):11463–11468.
41. Bench SR, et al. (2013) Whole genome comparison of six *Crocosphaera watsonii* strains with differing phenotypes. *J Phycol* 49:786–801.
42. Giovannoni SJ, et al. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science* 309(5738):1242–1245.
43. Dupont CL, et al. (2012) Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 6(6):1186–1199.
44. Lauro FM, et al. (2009) The genomic basis of trophic strategy in marine bacteria. *Proc Natl Acad Sci USA* 106(37):15527–15533.
45. Lynch M (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci USA* 104(Suppl 1):8597–8604.
46. Lynch M (2006) Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* 60:327–349.
47. Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. *Trends Genet* 17(10):589–596.
48. Welsh EA, et al. (2008) The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. *Proc Natl Acad Sci USA* 105(39):15094–15099.
49. Sinha R, et al. (2014) Comparative genomics of Cylindrospermopsis raciborskii strains with differential toxicities. *BMC Genomics* 15:83.
50. Tripp HJ, et al. (2008) SAR11 marine bacteria require exogenous reduced sulphur for growth. *Nature* 452(7188):741–744.
51. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302(5649): 1401–1404.
52. Morris JJ, Lenski RE, Zinser ER (2012) The Black Queen Hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* 3(2):e00036–12.
53. Bench SR, Ilikchyan IN, Tripp HJ, Zehr JP (2011) Two strains of *Crocosphaera watsonii* with highly conserved genomes are distinguished by strain-specific features. *Front Microbiol* 2:261.
54. Bandyopadhyay A, et al. (2011) Novel metabolic attributes of the genus cyanothece, comprising a group of unicellular nitrogen-fixing Cyanothece. *MBio* 2(5):e00214–11.
55. Kuo C-H, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Res* 19(8):1450–1454.
56. Konstantinidis KT, et al. (2009) Comparative systems biology across an evolutionary gradient within the Shewanella genus. *Proc Natl Acad Sci USA* 106(37):15909–15914.
57. Smith JK, Parry JD, Day JG, Smith RJ (1998) A PCR technique based on the Hip1 interspersed repetitive sequence distinguishes cyanobacterial species and strains. *Microbiology* 144(Pt 10):2791–2801.
58. Treangen TJ, Abraham A-L, Touchon M, Rocha EPC (2009) Genesis, effects and fates of repeats in prokaryotic genomes. *FEMS Microbiol Rev* 33(3):539–571.
59. Matus-Garcia M, Nijveen H, van Passel MWJ (2012) Promoter propagation in prokaryotes. *Nucleic Acids Res* 40(20):10032–10040.
60. Lin S, et al. (2011) Genome-wide comparison of cyanobacterial transposable elements, potential genetic diversity indicators. *Gene* 473(2):139–149.
61. Siguier P, Gourbeyre E, Chandler M (2014) Bacterial insertion sequences: Their genomic impact and diversity. *FEMS Microbiol Rev* 38(5):865–891.
62. Zhou K, Aertsen A, Michiels CW (2014) The role of variable DNA tandem repeats in bacterial adaptation. *FEMS Microbiol Rev* 38(1):119–141.
63. Voss B, Georg J, Schön V, Ude S, Hess WR (2009) Biocomputational prediction of noncoding RNAs in model cyanobacteria. *BMC Genomics* 10:123.
64. Gierga G, Voss B, Hess WR (2012) Non-coding RNAs in marine *Synechococcus* and their regulation under environmentally relevant stress conditions. *ISME J* 6(8):1544–1557.
65. Venter JC, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66–74.
66. Hewson I, et al. (2009) In situ transcriptomic analysis of the globally important keystone N_2-fixing taxon *Crocosphaera watsonii*. *ISME J* 3(5):618–631.
67. Kettler GC, et al. (2007) Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet* 3(12):e231.
68. Steglich C, et al. (2008) The challenge of regulation in a minimal photoautotroph: Non-coding RNAs in *Prochlorococcus*. *PLoS Genet* 4(8):e1000173.
69. García-Fernández JM, de Marsac NT, Diez J (2004) Streamlined regulation and gene loss as adaptive mechanisms in *Prochlorococcus* for optimized nitrogen utilization in oligotrophic environments. *Microbiol Mol Biol Rev* 68(4):630–638.
70. Pfreundt U, Kopf M, Belkin N, Berman-Frank I, Hess WR (2014) The primary transcriptome of the marine diazotroph *Trichodesmium erythraeum* IMS101. *Sci Rep* 4:6187.
71. Lorenz R, et al. (2011) ViennaRNA Package 2.0. *Algorithms Mol Biol* 6:26.
72. Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 319(5):1059–1066.
73. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLOS Comput Biol* 3(4):e65.
74. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF (2008) RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.
75. Kaneko T, et al. (2007) Complete genomic structure of the bloom-forming toxic cyanobacterium *Microcystis aeruginosa* NIES-843. *DNA Res* 14(6):247–256.
76. Ochman H, Davalos LM (2006) The nature and dynamics of bacterial genomes. *Science* 311(5768):1730–1733.
77. Nilsson AI, et al. (2005) Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci USA* 102(34):12112–12116.
78. Delihas N (2011) Impact of small repeat sequences on bacterial genome evolution. *Genome Biol Evol* 3:959–973.
79. Callanan M, et al. (2008) Genome sequence of Lactobacillus helveticus, an organism distinguished by selective gene loss and insertion sequence element expansion. *J Bacteriol* 190(2):727–735.
80. Gueguen E, Rousseau P, Duval-Valentin G, Chandler M (2006) Truncated forms of IS911 transposase downregulate transposition. *Mol Microbiol* 62(4):1102–1116.
81. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* 44: 445–477.