ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs

Vladimir V. Kapitonov, Kira S. Makarova, Eugene V. Koonin

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, USA

ABSTRACT

Bacterial genomes encode numerous homologs of Cas9, the effector protein of the type II CRISPR-Cas systems. The homology region includes the arginine-rich helix and the HNH nuclease domain that is inserted into the RuvC-like nuclease domain. These genes, however, are not linked to cas genes or CRISPR. Here, we show that Cas9 homologs represent a distinct group of nonautonomous transposons, which we denote ISC (insertion sequences Cas9-like). We identify many diverse families of full-length ISC transposons and demonstrate that their terminal sequences (particularly 3′ termini) are similar to those of IS605 superfamily transposons that are mobilized by the Y1 tyrosine transposase encoded by the TnpA gene and often also encode the TnpB protein containing the RuvC-like endonuclease domain. The terminal regions of the ISC and IS605 transposons contain palindromic structures that are likely recognized by the Y1 transposase. The transposons from these two groups are inserted either exactly in the middle or upstream of specific 4-bp target sites, without target site duplication. We also identify autonomous ISC transposons that encode TnpA-like Y1 transposases. Thus, the nonautonomous ISC transposons could be mobilized in trans either by Y1 transposases of other, autonomous ISC transposons or by Y1 transposases of the more abundant IS605 transposons. These findings imply an evolutionary scenario in which the ISC transposons evolved from IS605 family transposons, possibly via insertion of a mobile group II intron encoding the HNH domain, and Cas9 subsequently evolved via immobilization of an ISC transposon.

IMPORTANCE

Cas9 endonucleases, the effectors of type II CRISPR-Cas systems, represent the new generation of genome-engineering tools. Here, we describe in detail a novel family of transposable elements that encode the likely ancestors of Cas9 and outline the evolutionary scenario connecting different varieties of these transposons and Cas9.

Cas9 (CRISPR-associated protein 9) is the effector protein of the type II CRISPR-Cas systems, which represent about 10% of the CRISPR-Cas adaptive-immunity systems and, unlike other CRISPR-Cas types, are present only in bacteria (1). The Cas9 protein has a complex domain architecture, with a RuvC-like endonuclease domain; an HNH endonuclease domain that is inserted within the RuvC-like domain; and a large, α-helical recognition lobe for which no homologs have been identified in other proteins (1-4). Cas9 is an RNA-guided dual nuclease that recognizes and cleaves DNA of invading phages and plasmids. More specifically, Cas9 binds mature crRNAs, the products of CRISPR array expression followed by transcript processing, and employs the approximately 20-nucleotide unique portions of these RNAs (corresponding to the spacers in the CRISPR array) as guides to recognize and cleave DNA molecules that contain a sequence complementary to the guide adjacent to a short protospacer-associated motif (PAM). When the crRNA complexed with Cas9 forms a heteroduplex with the cDNA, the HNH domain of Cas9 cleaves the heteroduplex DNA strand, whereas the RuvC-like domain cleaves the second, unpaired DNA strand (3, 4). In addition to the recognition and cleavage of the target DNA at the interference stage of the CRISPR response, Cas9 is involved in PAM recognition during the adaptation stage, where short foreign DNA sequences are excised and inserted as spacers between the type II CRISPR repeats (5). The spacers, which are transcribed and incorporated into crRNAs, embody the bacterial immune memory, the key feature of CRISPR-Cas adaptive immunity.

In addition to the bona fide Cas9 genes that are lodged within type II CRISPR-Cas loci, numerous more distant, standalone homologs of Cas9 have been identified outside the CRISPR-Cas systems (1, 6). These proteins possess sequence features that are shared with Cas9, including the RuvC-like and HNH nuclease domains and a characteristic arginine-rich region (1, 6). The RuvC-like nuclease domain is also present in the widespread family of TnpB proteins that are often encoded by transposons of the IS605 superfamily (1). TnpB, often encoded as a standalone gene, is not required for transposition and, moreover, has been shown to downregulate transposition (7-9). Thus, the insertion elements that encode TnpB proteins alone are considered to be nonautonomous transposons.

It has been noticed that some bacterial genomes encode mul-
Multiple Cas9 homologs (1, 6), but none of these proteins has been experimentally studied, and no specific evidence that the proteins are encoded in transposable elements (TEs) has been presented. Here, we show that these Cas9 homologs are encoded by a novel group of mostly nonautonomous transposons that we denote ISC (insertion sequences Cas9-like). Based on their structural features and the predicted target site specificity, the ISC elements form a distinct group within the IS605/IS200 superfamily of bacterial and archaeal transposons that are mobilized by the Y1 tyrosine transposase (7, 9). The ISC transposon-encoded two nucleotide domain-containing proteins are the likely ancestors of the CRISPR-associated Cas9.

MATERIALS AND METHODS

A set of 2,751 complete bacterial and archaeal genomes was retrieved from the RefSeq database (February 2014) (10), and additional draft genomes were retrieved from GenBank. For detection of homologous proteins, the PSI-BLAST program (11) was used with default parameters to search the NCBI NR database. The same program was used to assign proteins from complete genomes to profiles of known protein families from the CDD database (12), with an E value threshold of $10^{-4}$ and low-complexity filtering turned off. Protein secondary structure was predicted using Jpred 4 (13). The HHpred program was used with default parameters for detection of remote sequence similarity (14). The BLASTCLUST program was used to cluster sequences with at least 90% length coverage and 90% sequence identity (15). Multiple-protein-sequence alignments for phylogenetic analysis were constructed using the MUSCLE program (16). Phylogenetic analysis was performed using the FastTree program with the WAG evolutionary model and the discrete gamma model with 20 rate categories (17) or using the PhyML program with automatic model selection (18).

The Censor program was used for the identification of similar nucleotide sequences and mapping the detected sequences on the genome (19). Each family of transposons was characterized either by its consensus sequences or by a single, apparently intact copy (when the family consisted of fewer than 3 copies >90% identical to each other). Transposon sequences with >75% pairwise nucleotide sequence identity were classified as members of the same family. Multiple alignments of DNA sequences were constructed using MAFFT (20). The consensus sequences were derived from multiple alignments using the "cons" routine of the EMBOSS package (21). For local sequence manipulations, programs from the EMBOSS and ALNPACK packages were used (19, 20).

In order to precisely identify the transposon termini, positions of insertions, and putative modifications of the target sites, two approaches were applied. Under the first strategy, transposons inserted into copies of some other TEs or unclassified repeats were identified. Based on pairwise alignments of these inserted copies of transposons with the consensus sequences, the exact termini of the inserted elements and target sites were identified.

Given that only a small fraction of the analyzed transposons are inserted into copies of other transposons or repetitive elements, we found useful the second strategy, which was based on the detection of polymorphic transposon insertions where one of the compared genomes contained an insert whereas the second genome lacked it. In the genome of a particular species, each transposon copy with intact ends was extended by 70 to 210 bp in each direction. Then, the transposon sequence was excised, and the 3’ and 5’ flanking sequences were concatenated and used as queries in BLASTN searches against the bacterial and archaeal NR, Genomes, and WGS sections of GenBank. Among all the significant hits with >85% identity to the query, those that did not contain insertions longer than 5 bp were selected. These hits corresponded to “empty target sites” in syntenic regions in the genomes of other species or species closely related to the query species. Each empty syntenic region was aligned with the query sequence containing the transposon copy using the aaln2 program (“mon” alignment option for logarithmic gap penalty) from Censor-ALNPACK. Based on manual inspection of these pairwise alignments, the exact termini of transposons and the positions of their insertions into the target sites were reevaluated.

RESULTS

Cas9 homologs in bacteria and archaea. It has been shown previously that Cas9 homologs belong to two major families. One family contains the HNH nuclease domain insertion within the RuvC-like domain, whereas in the other, the RuvC-like domain is continuous (1, 6). The two families are jointly represented in the Pfam database as pfam14239, RRXRR proteins; this Pfam entry includes the N-terminal portion of the proteins, which encompasses the first RuvC motif and the arginine-rich region. In order to avoid multiple false-positive hits to numerous HNH nucleases, we used representatives of the second, RuvC-only family as a query for a PSI-BLAST search. A PSI-BLAST search against the NCBI NR database initiated with the WP_044523571.1 protein from Nostoc sp. strain PCC 7120 identified 1,227 Cas9 homologs after 6 iterations (see Table S1 in the supplemental material). We then generated clusters of highly similar sequences using the BLASTCLUST program in order to identify potential recent expansions of the family (see Table S1 in the supplemental material).

As expected from previous observations (1), we identified many cases in which multiple, highly similar copies of Cas9 homologs were present in the same organism, e.g., 30 copies in the euryarchaeon Methanosarcina lacuistris, 15 in the cyanobacterium Microcystis aeruginosa, 13 in the bacterium Lactobacillus delbrueckii, and 8 in the bacterium Ktedonobacter racemifer (see Table S1 in the supplemental material). We then compared the distributions and abundances of this family of Cas9 homologs, Cas9 and TnpB, within the set of 2,751 complete bacterial and archaeal genomes (see Table S2 in the supplemental material). The TnpB proteins are extremely widespread in both bacteria and archaea (see Table S2 in the supplemental material). In contrast, Cas9 (the component of type II CRISPR-Cas loci) and Cas9 homologs show relatively narrow distributions that are limited to several bacterial phyla (see Table S2 in the supplemental material). The Cas9 homologs are especially abundant in cyanobacteria (31% of the available genomes encode at least one protein of the family). They are rarely present in the same genome with Cas9. Both Cas9 and Cas9 homologs are extremely rare in archaea, especially thermophiles (see Table S2 in the supplemental material).

The results of HHpred searches show that Cas9 homologs are much more similar to Cas9 than to the TnpB family, which, despite the presence of the readily detectable RuvC-like domain, is not even detected in HHpred searches starting from either Cas9 or Cas9 homologs, at least not with a probability of >50% (see Table S3 in the supplemental material). The multiple-sequence alignment of the RuvC-like domain and the arginine-rich helix region for the selected representatives of Cas9, Cas9 homologs, and TnpB proteins also clearly demonstrates the pronounced similarity between Cas9 and its homologs, especially in the arginine-rich helix and the RuvC-like motif III (Fig. 1). Taken together with the presence of the HNH domain inserted into the RuvC domain, a unique derived shared character, these findings are indicative of a direct evolutionary relationship between Cas9 and the family of Cas9 homologs.

Cas9 homologs are encoded by transposable elements. We sought to characterize the genes encoding Cas9 homologs and in particular, given the distant relationship with TnpB, the possibil-
The poorly conserved spacers between the alignment blocks are shown by numbers. The bottom sequence shows the RuvC nuclease from March 2016 Volume 198 Number 5 jb.asm.org

Moreover, as shown below, some copies of the transposon are -GTCA and ATGG-3

FIG 1

TnpB
IscB

Motifs
RuvC I
bridge helix
RuvC II
RuvC III

TnpB
15898041 Sulfolobus solfataricus 188 AAGTCTGGGACCCATGCGT
17232938 Neococcius sp. PCC 7120 AAGTCTGGGACCCATGCGT
165402709 Nitrosopelagia gessenia 166 AAGTCTGGGACCCATGCGT
43068108 Helicobacter pylori 106695 179 AAGTCTGGGACCCATGCGT
39295860 Flexibacter litoralis 130 AAGTCTGGGACCCATGCGT
28851665 Kerschera coli 157 AAGTCTGGGACCCATGCGT
69128642 Clostridium botulinum 177 AAGTCTGGGACCCATGCGT

IscB
29754772 Ktedonobacter racemifer DSM 44963 34 DQ174927 EX12
27612920 Collefactory sinus oblongus PCC 7420 64 DQ174927 EX12
16860323 Mirocystic aeruginosa NIES-843 71 DQ174927 EX12
433662072 Anthocystis albicans DSM 5150 59 DQ174927 EX12
295180844 Nitrosomonas europaea ABE-2 49 DQ174927 EX12
432001730 Xanthobacter flavithermus DMF-90-006 56 DQ174927 EX12
805378795 Methanotherma luxia 2-7290 58 DQ174927 EX12
12626716 Marinobacter sp. EL179 52 EF608927 EX12
528102090 Salipiger mucronatus DSM 16094 55 EF608927 EX12
512576587 Geitlerimana sp. PCC 7105 54 EF608927 EX12
297510252 Ktedonobacter racemifer DSM 44963 52 EF608927 EX12
799146833 Anaerostipes hadrus 52 EF608927 EX12
564332996 Youngiibacter fragilis 232-1 52 EF608927 EX12
654101052 Clostridium haemolyticum NCTC 8963 53 EF608927 EX12

Cas9
227842993 Acidaminococcus sp. D1 I 4 RVRCCNCR TTVDCR
291570705 Coprothermobacter catus GD-7 4 RVRCCNCR TTVDCR
425258457 Tracheophyta denticula ATCC 35405 7 TTVDRKTVV SR
47508688 Nycoplasma mobile 16K 14 RVRCCNCR TTVDCR
55822672 Streptococcus thermophilus CNRZ166 3 DTVDRKTVV SR
218532111 Campylobacter jejuni ATCC 11168 2 RVRCCNCR TTVDCR
433662072 Anthocystis albicans DSM 5150 59 DTVDRKTVV SR
178736489 Akkermansia muciniphila ATCC BAA-835 4 RVRCCNCR TTVDCR
189440764 Bifidobacterium longum SUBLOLA 4 RVRCCNCR TTVDCR
425798735 Wolinella succinogenes ATCC 17400 5 DTVDRKTVV SR
54266138 Legionella pneumophila str. Paris 7 DTVDRKTVV SR
184977355 Fractibacter novelliseda T112 5 DTVDRKTVV SR
79101852 Streptococcus pyogenes MNAS005 4 RVRCCNCR TTVDCR

RuvC (PDB 4EPS) Thermus thermophilus
1 ILE
11 PRO
17 LEU
21 ARG
19 ILE
23 LEU
22 ARG
26 LEU
25 ARG
27 LEU
29 LEU
30 LEU
32 LEU
34 LEU
36 LEU
38 LEU
40 LEU
1 PRO
5 ARG
6 ILE
10 LEU
12 LEU
14 LEU
16 LEU
18 LEU
20 LEU
22 LEU
24 LEU
26 LEU
28 LEU
30 LEU
32 LEU
34 LEU
36 LEU
38 LEU
40 LEU
42 LEU
44 LEU
46 LEU
48 LEU
50 LEU
52 LEU
54 LEU
56 LEU
58 LEU
60 LEU
62 LEU
64 LEU
66 LEU
68 LEU
70 LEU
72 LEU
74 LEU
76 LEU
78 LEU
80 LEU
82 LEU
84 LEU
86 LEU
88 LEU
90 LEU
92 LEU
94 LEU
96 LEU
98 LEU
100 LEU

ity that these proteins are encoded by a distinct family of TEs. The key feature of TEs is the presence of unique terminal structures, often terminal inverted repeats (TIR), and/or target site duplication. For the identification of these diagnostic features, it is essential to compare multiple copies of closely related TEs. For initial analysis, we selected one of the largest tight clusters of closely encoding Cas9 homologs, which was detected in the genome of K. racemifer DSM 493 (9 sequences altogether, 2 of which are identical). Using the multiple alignment of the corresponding DNA sequences that were expanded by 2 kb in each direction, we derived a 1,594-bp consensus sequence that showed, on average, 96% identity to the individual copies (the DNA and protein sequences of transposons are shown in Fig. S1 in the supplemental material). All the copies contained the perfectly conserved terminal sequences 5’-GTCA and ATGG-3’. The presence of these specific termini suggests that the genes encoding Cas9 homologs are copies of a TE rather than products of segmental duplications. Moreover, as shown below, some copies of the transposon are inserted into different unclassified repetitive elements.

The consensus sequence of this transposon, here denoted ISC1-2_KR (insertion sequences encoding Cas9 homolog family 2, subfamily 1, from Ktedonobacter racemifer), encodes a single protein, a Cas9 homolog, which we here denote IscB by analogy with TnpB. The subfamily of IscB proteins that contain the HNH insertion into the RuvC-like domain is referred to here as IscB2 (and the respective TE as ISC2), and the family with the RuvC-like domain only is referred to as IscB1 (and the respective TE as ISC1). To identify the target sites of ISC2-1_KHR, we searched for insertions of the transposon into other repetitive elements and identified three copies of ISC2-1_KHR inserted into unclassified repeats Rep1, Rep2, and Rep3 (most likely, these repeats, identified by Censor, belong to unclassified TEs). Using pairwise alignments of the repeat copies containing insertions of ISC2-1_KHR with the other copies of these repeats that were free of the insertions, we found that the actual termini of ISC2-1_KHR are 5’-CA and AT-3’ and that the ISC2-1_KHR copies are specifically inserted into GTGG target sites. Apparently, the insertions occurred precisely between GT and GG dinucleotides (GTGG), without generating any target site duplications (see Fig. S2 in the supplemental material).

By similar analysis of other IscB clusters in the K. racemifer genome consisting of 2 or 3 elements each, we identified three additional families of TEs related to ISC2-1_KHR and one to ISC1-1_KHR. All these transposons are short (1,341 to 1,707 bp) and encode only the IscB2 or IscB1 protein, respectively (Fig. 2A). Even from this analysis of several TE families in a single species, it was clear that these transposons are highly diverse, with the pairwise protein identity in a range from 30% to 63%. Despite the high interfamily sequence divergence, transposes from different families of ISC2_KHR repeat elements (and the single ISC1 element) show notable sequence conservation at the 5’ terminal and are inserted.
of the ISC2 sites without generating target site duplications (Fig. 2B). Specifically, between the GT and GR dinucleotides of the target sites, the key features of the IscB-encoding transposons initially identified in the K. racemifer genome are shared by transposons populating the genomes of diverse bacteria and some archaea (Fig. 2 and Table 1).

The ISC transposons share termini with IS605-like transposons encoding TnpB proteins. During the analysis of the consensus sequence of the ISC transposons in the K. racemifer genome, we identified two ~1,360-bp elements that were ~95% identical to each other, contained termini similar to those of ISC2-1_KR, and encoded the TnpB protein in their antisense strands (Fig. 3). Here, we refer to the "top" 5' to 3' DNA strand of a transposon as the sense strand, whereas the complementary strand, also from 5' to 3', is denoted antisense.

The ISC transposons share termini with IS605-like transposons encoding TnpB proteins. During the analysis of the consensus sequence of the ISC transposons in the K. racemifer genome, we identified two ~1,360-bp elements that were ~95% identical to each other, contained termini similar to those of ISC2-1_KR, and encoded the TnpB protein in their antisense strands (Fig. 3). Here, we refer to the "top" 5'-to-3' DNA strand of a transposon as the sense strand, whereas the complementary strand, also from 5' to 3', is denoted antisense. The elements that encode TnpB protein in their antisense strands are marked by arrows above the corresponding sequences.

FIG 2 Structures of nonautonomous ISC transposons. (A) Organizations of ISC transposons. Green rectangles, ORFs encoding IscB; RuvC and HNH, nuclease domains; RR, arginine-rich region. The bacterial hosts and the length of each transposon are indicated on the left and right of the transposon schematics, respectively. In the names of transposons, the KR, GS, CC, YF, and SM suffixes denote domains; RR, arginine-rich region. The bacterial hosts and the length of each transposon are indicated on the left and right of the transposon schematics, respectively. In the names of transposons, the KR, GS, CC, YF, and SM suffixes denote domains; RR, arginine-rich region. The bacterial hosts and the length of each transposon are indicated on the left and right of the transposon schematics, respectively. In the names of transposons, the KR, GS, CC, YF, and SM suffixes denote domains; RR, arginine-rich region. The bacterial hosts and the length of each transposon are indicated on the left and right of the transposon schematics, respectively. In the names of transposons, the KR, GS, CC, YF, and SM suffixes denote domains; RR, arginine-rich region.

We then sought to determine whether the identified features of the ISC2 and ISC1 transposons from K. racemifer (Fig. 2) are shared by transposons from other prokaryotes. For comparison, we selected several additional families of putative ISC elements that were detected in other bacteria, namely, Firmicutes (Youngiibacter fragilis 232.1, Anaobacillus flavidthermus TNO-09-006, Mahella australiensis, Anaerostipes hadrus, and Halobacteroides halobius), Cyanobacteria (Geitlerinema sp. strain PCC 7105, Microcystis aeruginosa NIES-843, and Coleofasciculus chthonoplastes PCC 7420), Proteobacteria (Salipiger mucus DSM 16094, Marinobacter sp. strain ELB17, and Halomonas), and a methanogenic archaeon (M. latus Z-7289). Most of the TE families belong to the ISC2 group, with the exception of two ISC2 transposons from the cyanobacterium C. chthonoplastes and M. aeruginosa. Analysis of these TE families using the approach described above confirmed that the key features of the IscB-encoding transposons initially identified in the K. racemifer genome are shared by transposons populating the genomes of diverse bacteria and some archaea (Fig. 2 and Table 1).

The ISC transposons share termini with IS605-like transposons encoding TnpB proteins. During the analysis of the consensus sequence of the ISC transposons in the K. racemifer genome, we identified two ~1,360-bp elements that were ~95% identical to each other, contained termini similar to those of ISC2-1_KR, and encoded the TnpB protein in their antisense strands (Fig. 3). Here, we refer to the "top" 5'-to-3' DNA strand of a transposon as the sense strand, whereas the complementary strand, also from 5' to 3', is denoted antisense. Thus, the elements that encode TnpB alone appear to be copies of a nonautonomous transposon, here called IS605B-1_KR, which most likely is mobilized by the Y1 tyrosine transposase (TnpA) of autonomous transposons. These autonomous transposons encode either both TnpA
and TnPB (IS605 family) or TnpA alone (IS200 family) (7–9). Here, we use “IS605 superfamily” as an encompassing term that also includes the nonautonomous IS1341 and PATE (palindrome-associated transposable elements) (7, 22, 23). For simplicity, we refer to IS1341-like, TnpB-only transposons as IS605B. The same mode of transposition dependent on Y1 transposases supplied in trans can be predicted for the ISC transposons. Indeed, the vast majority of the genomes that contain ISC transposons carry at least one Y1 family transposase, whereas serine transposases are much less common (Table 2).

In both examples described above, an ISC transposon was accompanied by an IS605 or IS605B transposon, with its 3′-terminal sequence similar to that of the 3′ terminus of the ISC transposon. Compared to the iscB genes, the tnpB genes in both IS605B transposons had the opposite orientation, i.e., were encoded by the antisense DNA strands. Surprisingly, this pattern of gene orientation holds among the other 9 families of ISC transposons and their IS605B counterparts (Table 1; see Fig. S1, S3, and S4 in the supplemental material).

**Flexibility among the target site cutting positions.** In addition to the several families of ISC transposons that are inserted precisely between the dinucleotides of the GTGG and GTGA target sites (see Fig. S2 and S3 in the supplemental material), transposons from several other ISC families from Cyanobacteria and Proteobacteria are inserted upstream of the ATG target sites. For example, in the genome of *M. aeruginosa* NIES-843, we identified 4 copies of the ISC1-1_MA transposon that are ~97% identical to their consensus sequence (see Fig. S4 in the supplemental material). The consensus sequence contained the 5′-GTCA and CGAT GA-3′ termini. Therefore, based on the previously described target site specificity, we assumed that the terminal GT and GA dinucleotides belonged to the GT|GA target site (13) shows the

---

**TABLE 1 ISC transposons in bacteria and archaea**

| Transposon family | Species | No. of copies (% identity)* | IS605 transposon(s) with similar terminus (no. of copies, % identity)* | Target site** | Taxonomy |
|-------------------|---------|---------------------------|----------------------------------------------------------|---------------|----------|
| ISC2-1_KR         | Ktedonobacter racenifer DSM 493 | 7 (99) | IS605B-1_KR (2, 99) | GT|GG\(^d\) (GT|GG) | Chloroflexi |
| ISC2-2_KR         | *Deinococcus radiodurans* JR1 | 3 (99) | IS605B-2_KR (4, 99) | GT|GA (GT|AA) | Chloroflexi |
| ISC2-3_KR         | *Deinococcus radiodurans* MJ1000 | 2 (91) | IS605B-3_KR (6, 99) | GT|GA (AGT|GA) | Chloroflexi |
| ISC2-4_KR         | *Deinococcus radiodurans* RY327 | 2 (97) | IS605-4_KR | GT|GA | Chloroflexi |
| ISC1-1_KR         | *Deinococcus radiodurans* RY327 | 3 (97) | | GT|GT | Chloroflexi |
| ISCY1-1_KR        | *Deinococcus radiodurans* RY327 | 2 (99) | IS605B-1_ME (6, 94) | ? | Chloroflexi |
| ISC1-1_CC         | Coleofasciculus chthonoplastes PCC 7420 | 4 (100) | | TT|GG | Cyanobacteria |
| ISC2-2_CC         | *Micrococcus australis* H105 | 2 (99) | IS605-1_CC | TT|GA | Cyanobacteria |
| ISC1-1_MA         | *Geitlerinema* NIES-843 | 4 (97) | IS605-1_MA (46, 96) | N|ATGA\(^d\) (N|ATGA)\(^d\) | Cyanobacteria |
| ISC2-1_GS         | *Geitlerinema* sp. PCC 7105 | 10 (98) | | GT|GG | Cyanobacteria |
| ISC2-2_GS         | *Geitlerinema* sp. PCC 7105 | 2 (91) | | GT|GG | Cyanobacteria |
| ISC1-1_HH         | Anaerostipes hadrus strain PEL 85 | 2 (97) | IS605B-1_AH, IS605B-2_AH | GT|AA (GT|AA) | Firmicutes |
| ISC2-1_HM         | Halobacteroides halobius DSM 5150 | 2 (98) | | AA|GG | Firmicutes |
| ISC2-1_AF         | Anoxybacillus flavithermus TNO-09.006 | 6 (97) | IS605B-1_AF (3, 99) | GT|GA\(^d\) (AT|GA)\(^d\) | Firmicutes |
| ISC2-1_YF         | Youngibacter fragilis 232.1 | 6 (96) | | GT|GA | Firmicutes |
| ISC2-1_CH         | Clostridium haemolyticum NCTC 9963 | 1 | | GT|AA\(^d\) | Firmicutes |
| ISC2-1_SM         | Sulphifer musovus DSM 16094 | 4 (98) | | CC|CG | Proteobacteria |
| ISC2-1_HM         | Halomonas mediterrana strain R113 | 1 | IS605B-1_OS (3, 97), IS605-1_OS, IS605B-1_HM | N|ATGA\(^d\) (N|ATGA) | Proteobacteria |
| ISC2-1_MS         | Marinobacter sp. ELB17 | 5 (94) | IS605B-1_MS | N|ATGA\(^d\) (N|ATNA)\(^d\) | Proteobacteria |
| ISC2-1_NH         | Nitronococcus halophilus Nc 4 | 4 (99) | | ? | Proteobacteria |
| ISC2-1_ML         | Methanococcoides lacuustris Z-7289 | 12 (96) | | GT|GA\(^d\) | Archaea |
| ISC2-2_ML         | Methanococcoides lacuustris Z-7289 | 10 (97) | IS605B-1_ML (7, 99) | GT|GA | Archaea |

* For each transposon, the number of copies is measured by the percent identity of the copies to the respective consensus sequence (the pairwise identity for transposons represented by two copies).
* IS605 and IS605B transposons that share common terminal regions with the corresponding ISC transposons.
* Target sites of the corresponding ISC and IS605 (in parentheses) transposons. ? Target site is not defined.
* Target site verified based on the identification of “empty” orthologous loci or repeats including empty copies and a copy harboring an ISC transposon (see Fig. S2, S4, and S6 in the supplemental material).
I302 genome contains one such locus (see Fig. S4C in A. aeruginosa counterpart loci free of the transposon copies and found that the
ined the cross-genome comparison aiming to identify the coun-
termini of the transposon. To verify this assumption, we exam-
ined the position of the transposon insertions) rather than to the
termini of the transposon. The two terminal hairpins are rendered in black and blue. (B) Pairwise
alignment of the 5’ and 3’ termini in both transposons; the terminal hairpins are formed by inverted repeats (marked by arrows). The identical GTGG target sites
are shown in green.

FIG 3 IS2C-1_KR and IS605B-1_KR transposons share similar termini. (A) Map of IS2C-1_KR copies in the K. racenifer DSM 49463 genome obtained by using Censor. The first three columns from the left contain the GenBank accession numbers and coordinates of GenBank DNA sequences similar to the query sequence (column 4); columns 5 and 6 give the first and last positions of a region in the query sequence that is similar to the corresponding GenBank sequence; column 7 shows the orientation of the GenBank sequence (d, direct; c, complementary); column 8 shows DNA identity (0 to 1). Two 34-bp sequences similar to the 5’ (column 4); columns 5 and 6 give the first and last positions of a region in the query sequence that is similar to the corresponding GenBank sequence; column 7 shows the orientation of the GenBank sequence (d, direct; c, complementary); column 8 shows DNA identity (0 to 1). Two 34-bp sequences similar to the 5’ terminus of IS2C-1_KR are shaded in light brown. These two sequences are the 5’ termini of two 95% identical copies of the IS605B-1_KR transposon. The two terminal hairpins are rendered in black and blue. (B) Pairwise alignment of the 5’ and 3’ termini in both transposons; the terminal hairpins are formed by inverted repeats (marked by arrows). The identical GTGG target sites
are shown in green.

exact position of the transposon insertions) rather than to the
termini of the transposon. To verify this assumption, we exam-
ined the cross-genome comparison aiming to identify the counter-
part loci free of the transposon copies and found that the M.
aeruginosa 9717 genome contains one such locus (see Fig. S4C in
the supplemental material). Pairwise alignment of this locus with
the corresponding ISCI-1_MA-containing locus of M. aeruginosa
NI8-843 showed that the 5’-GT dinucleotide is the end of the
transposon, whereas ATGA-3’ is the target site. In agreement
with this observation, we found that two full-length copies of ISCI-
In the M. aeruginosa 9717 genome were inserted at the loci syntenic to those in M. aeruginosa NIES-843 that did not contain the transposon insertion (see Fig. S4D in the supplemental material). For both loci, pairwise alignment of their DNA sequences with the sequences of their transposon-free orthologs confirmed that ISCI-1_MA has the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As mentioned above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site. As detailed above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717.

The IS605B-1_MS transposon is inserted precisely upstream of the ATCA and ATAA target sites, resembling the ATGA target site of IS2C-2_MS and IS2C-2_MS (see Fig. S6G in the supplemental material). Taken together, these findings strongly suggest that the transposition of IS605 transposons is catalyzed by the same or related Y1 transposases.

**IS605 transposons encoding TnpA-like tyrosine transposases.**

**TABLE 2 Cooccurrence of IscB and TnpB with Y1 tyrosine and serine transposases (S_T) in complete genomes**

| Y1/S_T presence | ISC (−)/ TnpB (−) | ISC (−)/ TnpB (+) | ISC (+)/ TnpB (−) | ISC (+)/ TnpB (+) |
|-----------------|-------------------|-------------------|-------------------|-------------------|
| Y1 (−)/S_T (−) | 1,296             | 235               | 0                 | 4                 |
| Y1 (−)/S_T (+) | 81                | 133               | 0                 | 4                 |
| Y1 (+)/S_T (−) | 532               | 285               | 1                 | 24                |
| Y1 (+)/S_T (+) | 34                | 88                | 0                 | 34                |

*−, absent; +, present.

As mentioned above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As mentioned above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As detailed above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As detailed above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As described above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As pictured above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As detailed above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.

As defined above, ISCI-1_MA and IS605B-1_MA transposons share a 3’-terminus (see Fig. S4A and B in the supplemental material), which in IS605 family transposons is recognized by the Y1 transposase (25). Importantly, these two transposons share not only the Y1-binding sites but also the same target site specificity. In contrast to ISCI-1_MA (4 copies), IS605B-1_MA is much more prolific and is represented in M. aeruginosa NIES-843 by 46 full-length copies. For 12 loci harboring these transposons, we identified transposon-free counterparts (empty target sites) in M. aeruginosa 9717. Examination of all 12 pairwise alignments of the transposon-carrying and transposon-free loci indicated that the IS605B-1_MA transposon was inserted precisely in the N|ATGA target sites and contained the 5’-GT and CG-3’-termini and is inserted in the N|ATGA target site.
Each of these 10 transposases is encoded by an IS protein sequences highly similar to the queries. After elimination of sequences Roseburia inulinivorans \(RI\), Eubacterium hallii \(NCTC 9693\); MA, estimated gamma shape parameter. The support for internal branches is indicated by Bayes approximation values above 95%. Species abbreviations: KR, \(K.\) racemifer \(DSM 44963\); CS, \(C.\) succiniciproducens; RP, \(R.\) propionica \(DSM 11124\); HP, \(H.\) pylori \(16\); BMT2, \(B.\) thetaiotaomicron \(VSB\); EH, \(E.\) helixonae; RP, \(R.\) pitchieri; MM, \(M.\) mazei; HP, \(H.\) pylori \(HP\); RI, Roseburia inulinivorans.

FIG 5 Phylogenetic tree of Y1 transposases encoded by IS\(C2Y\) and IS\(605\) transposons (IscA and TnpA, respectively). In IS\(C2Y\) transposons, IscA (red arrows) and IscB (blue arrows) ORFs overlap by 12 to 82 nucleotides. In IS\(605\) transposons, TnpA and TnpB are depicted as red and black arrows. The right and left arrows indicate ORFs encoded by the sense and antisense strands, respectively. The light-green and brown shading highlights the two distinct clades, each of which combines IscA and TnpA. The tree was obtained using the PhyML program (automatic model selection): LG model, discrete gamma model with 6 categories, and indicate ORFs encoded by the sense and antisense strands, respectively. The light-green and brown shading highlights the two distinct clades, each of which combines IscA and TnpA. The tree was obtained using the PhyML program (automatic model selection): LG model, discrete gamma model with 6 categories, and estimated gamma shape parameter. The support for internal branches is indicated by Bayes approximation values above 95%. Species abbreviations: KR, \(K.\) racemifer \(DSM 44963\); CS, \(C.\) succiniciproducens; RP, \(R.\) propionica \(DSM 11124\); HP, \(H.\) pylori \(16\); BMT2, \(B.\) thetaiotaomicron \(VSB\); EH, \(E.\) helixonae; RP, \(R.\) pitchieri; MM, \(M.\) mazei; HP, \(H.\) pylori \(HP\); RI, Roseburia inulinivorans.

implies that IS\(C2Y\) transposons are an old TE family. However, IS\(C2Y\) transposons are quite rare, with full-length elements detected in only 14 bacterial and archaeal genomes, typically as a single copy (Table 1). Only one genome (\(K.\) racemifer) harbors both ISC and IS\(C2Y\) transposons. Thus, in contrast to autonomous transposons of the IS\(605\) superfamily, IS\(C2Y\) transposons might be deleterious to the hosts and are not often fixed in the course of evolution. Apparently, the great majority of the ISC transposons rely on Y1 transposases encoded in TEs of other families for their transposition.

Using the IscA protein sequences encoded by IS\(C2Y\)-1\_KR, IS\(C2Y\)-1\_CH, and IS\(C2Y\)-1\_EC as queries in BLASTP searches against the NR database, we extracted \(\sim 150\) sequences that were highly similar to the queries. After elimination of sequences >65% identical to each other, they were reduced to a group of 10 protein sequences >40% identical to the IS\(C2Y\) transposases. Each of these 10 transposases is encoded by an IS\(605\) transposon that also carries a tnpB gene separated from the transposase gene by less than 1 kb.

Unexpectedly, a phylogenetic tree of the 6 identified IS\(C2Y\) transposases and 10 diverse IS\(605\) transposases contains two distinct, strongly supported branches, each of which joins IS\(C2Y\) (IscA) and IS\(605\) (TnpA) transposases (Fig. 5; see Fig. S7 in the supplemental material). Given that in all IS\(C2Y\) transposons the IscB and IscA ORFs are codirectional and overlapping, whereas in all but one of the closest IS\(605\) elements the two genes do not overlap and are divergently transcribed (Fig. 5), it is highly unlikely that these two groups of IS\(C2Y\) transposons evolved independently from different IS\(605\) transposons. Thus, the opposite direction of evolution, namely, that IS\(C2Y\) transposons "gave birth" independently to at least two groups of IS\(605\) elements, in each case yielding a chimeric IscA-TnpB transposon, appears most parsimonious (Fig. 5).

DISCUSSION

The findings described here show that IscB proteins, the closest homologs of Cas9 not linked to CRISPR-Cas systems, are encoded by a distinct group of transposons, which we denote ISC. The extensive mobility of ISC transposons is supported by the discovery of numerous empty target sites in pairs of closely related genomes, one of which contains a transposon.

Virtually all types of DNA transposons are defined by their relatively short terminal regions, which are recognized by the respective transposases. The ISC transposons share similar terminal sequences containing characteristic imperfect palindromes with the transposons of the IS\(605\)B family but not with any known transposons from other (super)families with terminal sequences significantly similar to those of the ISCs. Terminal sequences with this structure are typical of the so-called HUH Y1 transposons, so named after the conserved metal-binding motif (histidine-hydrophobic residue-histidine) in their tyrosine transposases (7, 26). Most of the ISC transposons do not
encode a transposase and, accordingly, are nonautonomous elements that are most likely mobilized in trans by Y1 transposases (TnpA proteins) encoded by other transposons of the IS605 superfamily (23). Indeed, most of the genomes that contain ISC transposons also encompass at least one TnpA-encoding transposon. We identified several transposons that encode both IscB and TnpA (ISC2Y transposons), which is obviously compatible with the involvement of TnpA in ISC transposition. However, ISC2Y transposons are rare and apparently do not typically serve as helpers to nonautonomous ISC elements. The Y1 (TnpA) transposase forms a dimer and contains hybrid active sites, with an HUH motif in one monomer and the catalytic tyrosine in the other monomer (26); these catalytic residues are also conserved in the Y1 transposase (IscA) that is encoded by the ISCY transposons. Overall, the identified ISC elements extend the group of numerous nonautonomous IS200/IS605 transposons, including so-called REP and IStron elements (27).

In contrast to many other transposons, ISC and IS605 elements lack terminal inverted repeats that are typically recognized by diverse DDE or cut-and-paste transposases (7, 23). Instead, they contain subterminal imperfect hairpins that are recognized by the Y1 transposase (25, 26). All transposons that employ Y1 transposases are target site specific and integrate into the target sites without generating target site duplications. Unlike many other bacterial DNA transposons that are often inserted into random sites, the IS605 superfamily transposons are inserted 3′ of a specific 4- or 5-bp target site (7, 23). As shown here, the IS605 target site specificity is much more diverse than previously suspected and can be extended to include GT|GG, GT|RA, and N|ATGA target sites (Fig. 4 and Table 1), with transposons inserting either inside the specific target site or 5′ of it. Considering all these observations, the ISC transposons appear to represent a distinct family within the IS200/IS605 superfamily.

The phylogenetic tree of the IscB family consists of two distinct clades, IscB1 and IscB2 (Fig. 6; see Fig. S8 in the supplemental material). Although the tree could not be rooted by TnpB proteins due to the insufficient number of reliably aligned positions between IscB and TnpB, this topology clearly is consistent with the origin of IscB2 from IscB1 through insertion of the HNH nuclease-coding sequence into the ancestral iscB1 gene. The HNH nuclease could originate from a group II intron. Indeed, we detected insertion of group II introns encoding a reverse transcriptase-HNH fusion protein into IS605 transposons in several bacterial genomes (e.g., *Anabaena cylindrica* PCC 7122 and *Nostoc* sp. strain PCC 7107) (see Fig. S9 in the supplemental material).

In more than 15 identified examples of highly diverse ISC
transposons with terminal sequences similar to those of IS605B or IS605 transposons (Table 1; see Fig. S1, S3, and S4A and B in the supplemental material), the IscB and TnpB proteins are always encoded in the sense and antisense strands, respectively. The simplest explanation of this unusually stable orientation of the tnpB and iscB genes could be a tight association of the promoters of these genes with the 3′-subterminal regions of the respective transposons.

Remarkable features of TnpB are its high abundance and wide spread among bacteria and archaea, which are surprising because, as shown in several experimental studies on IS605 and IS607 transposons, TnpB is not essential for transposition and is thought to be involved in the regulation of transposase expression and activity (7, 8, 28, 29). The major function of the RuvC nuclease in bacteria is resolution of the Holliday junctions during recombination and repair of DNA breaks (30, 31). Transposition can lead to a substantial increase in the number of DNA breaks, which could be deleterious to both the host and the transposons. The TnpB proteins could be directly involved in DNA break repair via the activity of the RuvC-like endonuclease domain. As a result, the TnpB-encoding transposons could be more successful than transposons devoid of TnpB in terms of long-term coevolution with the hosts due to minimization of the damage to the host caused by transposition. However, more complicated mechanisms of action of TnpB and IscB proteins cannot be ruled out, in particular those that would involve RNA binding via the arginine-rich helices of these proteins. Such a possibility seems particularly plausible for the IscB2 proteins, which might function analogously to Cas9 proteins, which employ the HNH nuclease domain to cleave the DNA strand paired with the CRISPR RNA and the RuvC-like domain to cleave the opposite strand (3, 32).

The wide spread of TnpB in autonomous and nonautonomous transposons, which contrasts with the relatively narrow occurrence of both Cas9 (within type II CRISPR-Cas loci) and IscB (within ISC transposons), together with the simple domain organization of these proteins, suggests that TnpB is the ancestral form to both IscB and Cas9. The relatively high sequence similarity between Cas9 and IscB2 and, even more important, the shared presence of two nuclease domains indicate that these families have a common ancestor. The direction of evolution between IscB and Cas9 is difficult to infer with confidence, especially given that the recent discovery of additional class 2 CRISPR-Cas systems, which employ a common ancestor. The direction of evolution between IscB and Cas9, with subsequent immobilization of the RuvC-like nuclease domain (34, 35), is fully compatible with this conclusion.

ACKNOWLEDGMENTS

We thank Yuri Wolf for invaluable technical help and Michael Chandler for useful discussions, particularly on TE nomenclature.

The research is supported by the intramural funds of the U.S. Department of Health and Human Services (to the National Library of Medicine).

V.V.K., K.S.M., and E.V.K. designed the research; V.V.K. and K.S.M. performed the research; V.V.K., K.S.M., and E.V.K. analyzed the results; V.V.K. and E.V.K. wrote the manuscript, which was edited and approved by all of us.

FUNDING INFORMATION

The U.S. Department of Health and Human Services provided funding to Vladimir V. Kapitonov, Kira S. Makarova, and Eugene V. Koonin through the intramural funding program.

REFERENCES

1. Chylinski K, Makarova KS, Charpentier E, Koonin EV. 2014. Classification and evolution of type II CRISPR-Cas systems. Nucleic Acids Res 42:6091–6105. http://dx.doi.org/10.1093/nar/gku241.
2. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV. 2006. A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1:7. http://dx.doi.org/10.1186/1745-6150-1-7.
3. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. 2012. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821. http://dx.doi.org/10.1126/science.1225829.
4. Gasanovas G, Barrangou R, Horvath P, Siksnys V. 2012. Cas9-crRNA ribonuclease-protein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A 109:E2579–E2586. http://dx.doi.org/10.1073/pnas.1208570109.
5. Heler R, Samai P, Modell JW, Weiner C, Goldberg GW, Bikard D, Marraffini LA. 2015. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature 519:199–202. http://dx.doi.org/10.1038/nature14245.
6. Makarova KS, Aravind L, Wolf YI, Koonin EV. 2011. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol Direct 6:38. http://dx.doi.org/10.1186/1745-6160-6-38.
7. Sigur` e P, Gourbeyre E, Chandler M. 2014. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev 38:865–891. http://dx.doi.org/10.1111/1574-6976.12067.
8. Pasternak C, Dulermo R, Ton-Hoang B, Debuchy R, Sigur` e P, Coste G, Chandler M, Sommer S. 2013. ISDrα2 transposition in Deinococcus radiodurans is downregulated by TnpB. Mol Microbiol 88:443–455. http://dx.doi.org/10.1111/mmi.12194.
9. He S, Guyenet C, Sigurr e P, Hickman AB, Dyda F, Chandler M, Ton-Hoang B. 2013. IS200/IS605 family single-strand transposition: mechanism of IS608 strand transfer. Nucleic Acids Res 41:3302–3313. http://dx.doi.org/10.1093/nar/gkt014.
10. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65. http://dx.doi.org/10.1093/nar/gkl482.
11. Altshul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. http://dx.doi.org/10.1093/nar/25.17.3389.
12. Marchler-Bauer A, Zheng C, Chitsaz F, Derbyshire MK, Geer LY, Geer RC, Gonzales NR, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Lu S, Marchler GH, Song JS, Thanki N, Yamashita RA, Zhang D, Bryant SH. 2013. CDJ: conserved domains and protein three-dimensional structure. Nucleic Acids Res 41:D348–D352. http://dx.doi.org/10.1093/nar/gks1243.
13. Drozdetskiy A, Cole C, Procter J, Barton GI. 2015. JPre4t: a protein secondary structure prediction server. Nucleic Acids Res 43(W1):W389–W394. http://dx.doi.org/10.1093/nar/gkv332.
14. Soding J, Biegert A, Lupas AN. 2005. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33:W244–W248. http://dx.doi.org/10.1093/nar/gki408.
15. Wheeler D, Bhagwat M. 2007. BLAST QuickStart: example-driven web-based BLAST tutorial. Methods Mol Biol 395:149–176. http://dx.doi.org/10.1007/978-1-59745-514-5_9.
16. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. http://dx.doi.org/10.1093/nar/gkh340.

806 Journal of Bacteriology March 2016 Volume 198 Number 5
17. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. http://dx.doi.org/10.1371/journal.pone.0009490.

18. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704. http://dx.doi.org/10.1080/10635150390235520.

19. Kothy O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474. http://dx.doi.org/10.1186/1471-2105-7-474.

20. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. http://dx.doi.org/10.1093/molbev/mst010.

21. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. http://dx.doi.org/10.1016/S0168-9525(00)02024-2.

22. Dyall-Smith ML, Pfeiffer F, Klee K, Palm P, Gross K, Schuster SC, Rampp M, Oesterhelt D. 2011. Haloquadratum walsbyi: limited diversity in a global pond. PLoS One 6:e20968. http://dx.doi.org/10.1371/journal.pone.0020968.

23. Siguier P, Gourbeyre E, Varani A, Ton-Hoang B, Chandler M. 2015. Everyman’s guide to bacterial insertion sequences. Microbiol Spectr 3:MDNA3-0030-2014. http://dx.doi.org/10.1128/microbiolspec.MDNA3-0030-2014.

24. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34:D32–D36. http://dx.doi.org/10.1093/nar/gkj014.

25. Ronning DR, Guynet C, Ton-Hoang B, Perez ZN, Ghirlando R, Chandler M, Dyda F. 2005. Active site sharing and subterminal hairpin recognition in a new class of DNA transposases. Mol Cell 20:143–154. http://dx.doi.org/10.1016/j.molcel.2005.07.026.

26. Chandler M, de la Cruz F, Dyda F, Hickman AB, Moncalian G, Ton-Hoang B. 2013. Breaking and joining single-stranded DNA: the HUH endonuclease superfamily. Nat Rev Microbiol 11:525–538. http://dx.doi.org/10.1038/nrmicro3067.

27. He S, Corneloup A, Guynet C, Lavatine L, Caumont-Sarcos A, Siguier P, Marty B, Dyda F, Chandler M, Ton Hoang B. 2015. The IS200/IS605 family and “peel and paste” single-strand transposition mechanism. Microbiol Spectr 3. http://dx.doi.org/10.1128/microbiolspec.MDNA3-0039-2014.

28. Boocock MR, Rice PA. 2013. A proposed mechanism for IS607-family serine transposases. Mob DNA 4:24. http://dx.doi.org/10.1186/1759-7753-4-24.

29. Kersulyte D, Mukhopadhyay AK, Shirai M, Nakazawa T, Berg DE. 2000. Functional organization and insertion specificity of IS607, a chimeric element of Helicobacter pylori. J Bacteriol 182:5300–5308. http://dx.doi.org/10.1128/JB.182.19.5300-5308.2000.

30. Aiyoshi M, Vassylyev DG, Iwasaki H, Nakamura H, Shinagawa H, Morikawa K. 1994. Atomic structure of the RuvC resolvase: a Holliday junction-specific endonuclease from E. coli. Cell 78:1063–1072. http://dx.doi.org/10.1016/0092-8674(94)90280-1.

31. Lilley DM, White MF. 2001. The junction-resolving enzymes. Nat Rev Mol Cell Biol 2:433–443. http://dx.doi.org/10.1038/35073057.

32. Jinek M, Jiang F, Taylor DW, Sternberg SH, Kaya E, Ma E, Anders C, Hauer M, Zhou K, Lin S, Kaplan M, Iavarone AT, Charpentier E, Nogales E, Doudna JA. 2014. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science 343:1247997. http://dx.doi.org/10.1126/science.1247997.

33. Koozin EV, Krupovic M. 2015. Evolution of adaptive immunity from transposable elements combined with innate immune systems. Nat Rev Genet 16:184–192. http://dx.doi.org/10.1038/nrg3859.

34. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, I van der Oost Regev A, Konnorn EV, Zhang F. 2015. Cpf1 Is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell 163:759–771. http://dx.doi.org/10.1016/j.cell.2015.09.038.

35. Shmakov S, Abudayeye OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Koonin M, Severinov K, Zhang F, Koonin EV. 2015. Discovery and functional characterization of diverse class 2 CRISPR-Cas systems. Mol Cell 60:385–397. http://dx.doi.org/10.1016/j.molcel.2015.10.008.