CRISPR

The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases

Han Altae-Tran[1,2,3,4,5]†, Soumya Kannan[1,2,3,4,5]†, F. Esra Demircioglu[1,2,3,4,5], Rachel Oshiro[1,2,3,4,5], Suchita P. Nety[1,2,3,4,5], Luke J. McKay[6,7,8], Mensur Dlakić[9], William P. Inskeep[6,7], Kira S. Makarova[10], Rhiannon K. Macrae[1,2,3,4,5], Eugene V. Koonin[10], Feng Zhang[1,2,3,4,5]*

IscB proteins are putative nucleases encoded in a distinct family of IS200/IS605 transposons and are likely ancestors of the RNA-guided endonuclease Cas9, but the functions of IscB and its interactions with any RNA remain uncharacterized. Using evolutionary analysis, RNA sequencing, and biochemical experiments, we reconstructed the evolution of CRISPR-Cas9 systems from IS200/IS605 transposons. We found that IscB uses a single noncoding RNA for RNA-guided cleavage of double-stranded DNA and can be harnessed for genome editing in human cells. We also demonstrate the RNA-guided nuclease activity of TnpB, another IS200/IS605 transposon-encoded protein and the likely ancestor of Cas12 endonucleases. This work reveals a widespread class of transposon-encoded RNA-guided nucleases, which we name OMEGA (obligate mobile element–guided activity), with strong potential for developing as biotechnologies.

The prokaryotic RNA-guided defense system CRISPR-Cas9 (type II CRISPR-Cas), which has been adopted for genome editing in eukaryotic cells (*1*, *2*), is thought to have evolved from IscB proteins (*3*). Despite its wide distribution across prokaryotes and its shared domain composition and architecture with Cas9, the function of IscB remains unknown (fig. S1). Moreover, given that IscB has not been reported to be associated with noncoding RNA (ncRNA) or CRISPR arrays, the evolutionary origins of the RNA-guided activity in Cas9 systems are unclear. IscB is encoded by a distinct subset of IS200/IS605 superfamily transposons that also include transposons encoding *tnpB*, a putative endonuclease distantly related to *iscB* and thought to be the ancestor of Cas12, the type V CRISPR effector (*3*–*5*). Using phylogenetic analysis, RNA sequencing (RNA-seq), and biochemical experiments, we sought to elucidate the functions of these proteins and the origin of RNA-guided activity in class 2 CRISPR systems.

[1]Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. [3]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [5]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [6]Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717, USA. [7]Thermal Biology Institute, Montana State University, Bozeman, MT 59717, USA. [8]Center for Biofilm Engineering, Montana State University, Bozeman, MT 59717, USA. [9]Department of Microbiology and Cell Biology, Montana State University, Bozeman, MT 59717, USA. [10]National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894, USA.
*Corresponding author. Email: zhang@broadinstitute.org
†These authors contributed equally to this work.

IscB is associated with an evolutionarily conserved noncoding RNA

IscB is ~400 amino acids long and contains a RuvC endonuclease domain split by the insertion of a bridge helix (BH) and an HNH endonuclease domain, an architecture that is shared with Cas9 (Fig. 1A) (*3*). We performed a comprehensive search for proteins containing an HNH or a split RuvC endonuclease domain and found that Cas9 and IscB were the only proteins that contained both domains (data S1). This search also showed that IscB contains a previously unidentified N terminus that lacks clear homology to known domains and is absent in Cas9, which we denoted PLMP after its conserved sequence motifs (Fig. 1A and fig. S2). Clustering and phylogenetic analysis of the combined RuvC, BH, and HNH domains strongly suggests that all extant Cas9s descended from a single ancestral IscB (Fig. 1B and data S2 and S3). We searched for CRISPR arrays adjacent to *iscB* genes from each cluster and found six distinct groups of IscB, containing 16 clusters (of 603 total), that were CRISPR-associated, contrary to previous observations (*3*). CRISPR-associated IscBs were scattered around the IscB phylogenetic tree, which suggests that they evolved independently, with one association event leading to the Cas9 lineage (Fig. 1B). In total we identified 31 unique CRISPR-associated *iscB* loci (of 2811 total).

Given their association with CRISPR arrays, we suspected that the rarely occurring CRISPR-associated IscBs may be RNA-guided nucleases. We first examined a cluster of CRISPR-associated IscBs similar to non–CRISPR-associated IscBs (at ~50% amino acid identity). We heterologously expressed a representative locus from this clade in *Escherichia coli* and performed small RNA-seq, which showed expression of not only the CRISPR array, but also a 329–base pair (bp) intergenic region between the CRISPR array and the IscB open reading frame (ORF) (Fig. 1C). We purified the IscB protein and sequenced the copurified RNA, demonstrating that this protein interacts with a single ncRNA component that encompasses both the CRISPR array and this intergenic region (Fig. 1C).

Given its interaction with a ncRNA that includes the CRISPR direct repeat (DR) and spacer, as well as its similar domain architecture to Cas9, we tested this IscB for RNA-guided endonuclease activity. Using a previously established protospacer adjacent motif (PAM)–discovery assay (table S1) (*6*), we observed depletion of specific PAM sequences (Fig. 1D and fig. S3), indicating that CRISPR-associated IscBs are reprogrammable RNA-guided nucleases. We confirmed this enzymatic activity with an in vitro cleavage assay using recombinant ribonucleoprotein (RNP) complexes (Fig. 1E).

Our finding that IscB functionally associated with CRISPR at least once, and likely on additional occasions, suggested that IscB systems more generally share a core ancestral ncRNA gene that is prone to evolving into a CRISPR array and in some cases a separate trans-activating CRISPR RNA (tracrRNA) (*7*). To test this hypothesis, we aligned 563 nonredundant *iscB* loci and searched for conserved nucleotide (nt) sequences either upstream or downstream of the *iscB* ORF. This analysis revealed a highly conserved intergenic region ~300 bp in length upstream of the ORF with a drop in conservation at the 5′ end, which corresponds to an IS200/IS605 transposon end. Secondary structure predictions for individual sequences revealed the presence of multiple G:U pairs (fig. S4), suggesting that the conserved region encodes an ncRNA containing functionally important hairpins, which we named ωRNA. Small RNA-seq on a sample of *Ktedonobacter racemifer* strain SOSP1-21, a soil bacterium that harbors 46 IscB loci in its genome (*3*), demonstrated expression of the predicted ωRNA in many of these loci (Fig. 1F and figs. S5 and S6A). Moreover, we observed that the transcripts consistently extended beyond the conservation boundary at the 5′ end.

An Rfam search for potential homologs of the ωRNA showed that the conserved region of the ωRNA partially matched the previously reported HEARO RNA, a ncRNA that was found upstream of HNH domain–containing proteins, which at the time were thought to be homing endonucleases (*8*, *9*). However, the Rfam search did not provide any clues about the nature of the 5′-terminal nonconserved

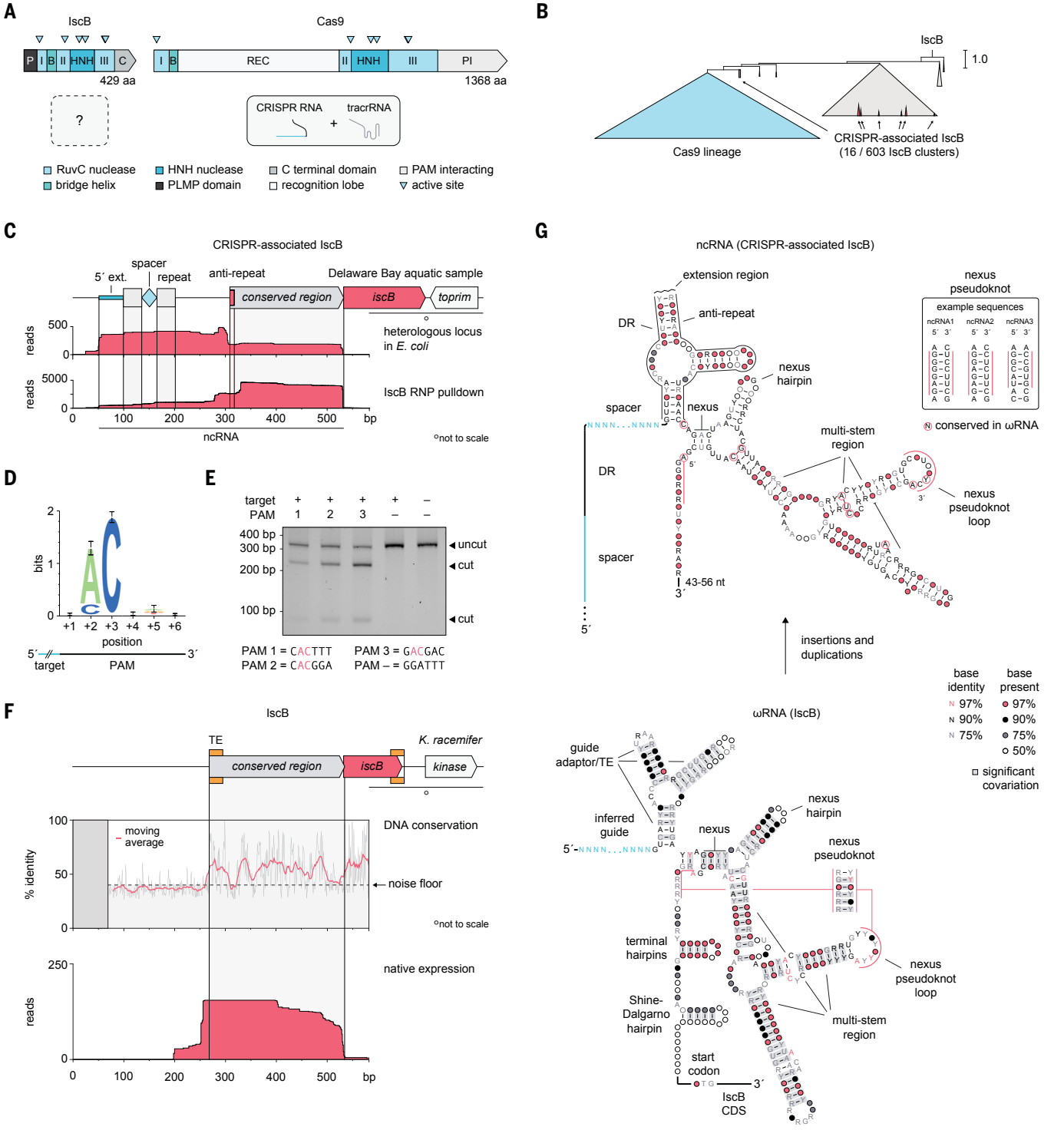

Fig. 1. IscBs are associated with ncRNAs of unknown function. (**A**) Comparison of IscB and Cas9 domains and previously described ncRNAs. (**B**) Phylogenetic analysis of the RuvC, BH, and HNH domains of Cas9 and IscB clusters using IQ-Tree 2. Genomic association shows that 16 of 603 IscB clusters have strong association to CRISPR, occurring independently in multiple clades. (**C**) Small RNA-seq of a heterologously expressed CRISPR-associated IscB locus (top) and RNP pulldown (bottom). (**D**) Sequence logo for the PAM as determined by a plasmid depletion assay. (**E**) In vitro cleavage by IscB single guide RNA–RNP complex. (**F**) Top: Conservation analysis of regions upstream of $N = 563$ nonredundant IscB loci. Bottom: Small RNA-seq of an IscB locus in *K. racemifer* strain SOSP1-21. TE, transposon end. (**G**) Secondary structure predictions of CRISPR-associated IscB ncRNA and IscB ωRNA. Guiding function of ωRNAs was inferred by comparison of the two structures.

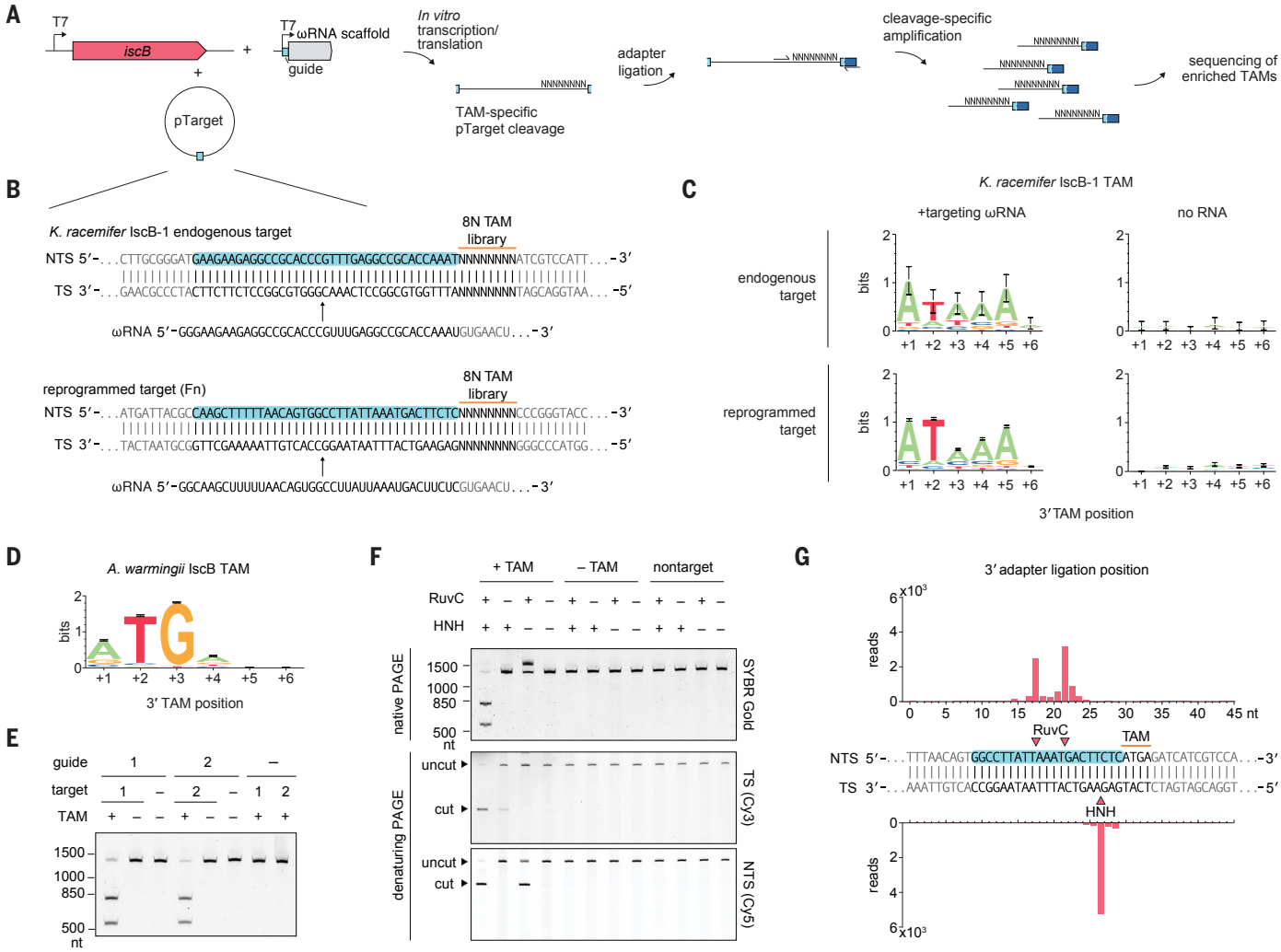

Fig. 2. IscB is an RNA-guided DNA endonuclease. (**A**) Design of an IVTT-based TAM screen. (**B**) KraIscB-1 endogenous target and reprogrammed target sequences used in IVTT TAM screens. (**C**) dsDNA cleavage by KraIscB-1 and ωRNA targeting sequence flanked by ATAAA 3′ TAM. (**D**) dsDNA cleavage by AwaIscB and ωRNA targeting sequence flanked by ATGA 3′ TAM. (**E**) In vitro reconstituted AwaIscB-ωRNA RNP cleavage of dsDNA substrates in the presence or absence of a target and/or TAM. (**F**) In vitro dsDNA cleavage of AwaIscB with selectively inactivated nuclease domains. TS, target strand; NTS, nontarget strand. (**G**) Sequencing of cleavage products generated by AwaIscB.

portion of these transcripts. Comparison of the consensus CRISPR-associated IscB ncRNA and the covariance folded ωRNA secondary structures revealed high degrees of structural and sequence similarity, particularly in shared multistem regions and pseudoknots (Fig. 1G, fig. S7, and supplementary text). We inferred that the 5′-most nonconserved sequence in the ωRNA might function as a guide sequence, because the sequence immediately downstream was predicted to form hairpins that structurally resembled the hairpins formed by the DR/anti-repeat duplex in the CRISPR-associated IscB ncRNA (Fig. 1G).

IscB is a reprogrammable RNA-guided DNA endonuclease

To test whether IscB was capable of cleaving DNA complementary to the putative ωRNA guide, we performed an in vitro plasmid cleavage assay with KraIscB-1 using an in vitro transcription/translation (IVTT) expression system (Fig. 2, A and B). We found that KraIscB-1 cleaved the target in an ωRNA-dependent manner, with an ATAAA 3′ target-adjacent motif (TAM) (Fig. 2C). Retargeting of KraIscB-1 using a different guide (Fn guide) (*6*) also mediated cleavage of the cognate target (Fig. 2C and fig. S6B), implying that IscB is a reprogrammable RNA-guided nuclease.

Next, we biochemically characterized IscB in vitro. We identified activity in 57 of 86 (66%) selected phylogenetically diverse systems (table S2) as determined by the identification of a TAM (fig. S8). Of these 57 functional IscBs, five could be reconstituted with the respective ωRNA in vitro to achieve efficient target cleavage, and from those, we selected AwaIscB (from *Allochromatium warmingii*) for detailed biochemical characterization (Fig. 2, D to G).

We confirmed the ability of recombinant AwaIscB to cleave multiple double-stranded DNA (dsDNA) targets in a programmable manner (Fig. 2E) and showed that the activity of AwaIscB is magnesium-dependent with a temperature optimum from 35° to 40°C (fig. S9, A and B). Appreciable activity was observed in vitro with guide lengths between 15 and 45 nt (fig. S9D). Mutation of the catalytic RuvC-II residue (E157A) abolished the nucleolytic activity on the nontarget DNA strand, whereas the HNH domain catalytic mutant H212A abolished the nucleolytic activity on the target strand (Fig. 2F). Combination of the E157A and H212A mutations (dAwaIscB) abolished all dsDNA nucleolytic activity (Fig. 2F) (*10*, *11*). Sequencing of the cleavage products showed

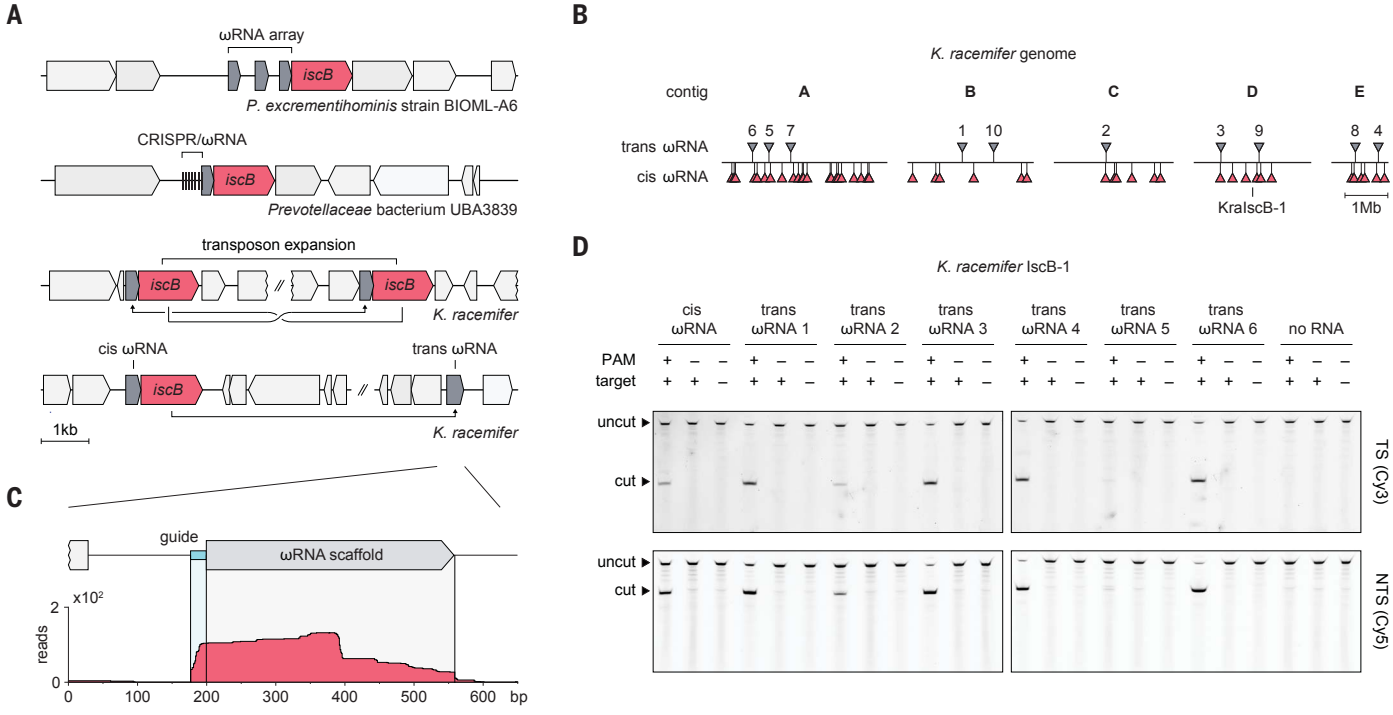

Fig. 3. Guide-encoding mechanisms of IscB. (**A**) Example loci for each major mechanism of encoding multiple guides: Entire ωRNA arrays associate with IscB; ωRNAs duplicate or insert into CRISPRs; transposition expansion results in multiple nearly identical loci that each express different guides; and standalone trans-acting ωRNAs form independently of adjacent IscBs. (**B**) *K. racemifer* encodes 48 IscB loci with cis ωRNAs and 10 standalone trans-acting ωRNAs. (**C**) Small RNA-seq of a standalone ωRNA locus in *K. racemifer*. (**D**) KraIscB-1, in complex with cis or trans ωRNAs with the same guide sequence, mediates cleavage of dsDNA in a TAM- and target-dependent manner. Reactions were performed in IVTT using 5′ strand–specific labeled linear targets. Contig accession and position information for all displayed loci are listed in table S6.

that AwaIscB cleaves the target strand 3 nt upstream of the TAM, similar to Cas9s (*12*). Cleavage of the nontarget strand occurred 8 or 12 nt upstream of the TAM, generating 5′ overhangs 5 or 9 nt in length (Fig. 2G and fig. S10). Exonuclease III mapping of a target substrate engaged by the dAwaIscB-ωRNA RNP showed that the RNP hindered exonuclease III treatment 19 nt upstream of the TAM on the target strand and 6 nt downstream of the targeted sequence on the nontarget strand (fig. S11) (*13*). We also found that truncation of more than four amino acids of the PLMP domain of AwaIscB abolished cleavage activity (fig. S12).

IscB uses multiple guide-encoding mechanisms

A distinct advantage of RNA-guided systems is that they allow an effector to target many substrates by simply reprogramming the RNA guide. One way IscB evolved to use multiple guides is association with CRISPR arrays (Fig. 3A). However, given that *iscB* loci typically encode a single ωRNA, it is unclear how or even whether these systems achieve such modularity in general. By searching for ωRNAs not directly adjacent to *iscB* ORFs, we uncovered three additional potential mechanisms for guide encoding and switching: ωRNA arrays, transposon expansion, and standalone, trans-acting ωRNAs (Fig. 3A). ωRNA arrays consist of multiple ωRNAs, each encompassing a distinct guide, separated by up to 200 bp, and are found in 15 of 3356 unique IscB/IsrB loci (0.4%). Transposon expansion involves the insertion of nearly identical IS200/IS605 superfamily transposons in multiple locations, resulting in multiple loci per genome, each capable of expressing a nearly identical ωRNA scaffold with a unique guide (fig. S13). By contrast, standalone ωRNAs, which show no detectable genomic associations with *iscB*, were more common and were found in multiple copies in some genomes (table S3). Cis ωRNAs from 95 of 3356 unique IscB/IsrB loci (2.8%) were nearly identical (≥95% sequence identity) to distally encoded standalone ωRNAs (fig. S14), implying that these standalone ωRNAs could encode guides used by trans-encoded IscBs.

We tested this possibility by examining 10 standalone ωRNAs in the *K. racemifer* genome (Fig. 3B), nine of which were found to be expressed (Fig. 3 and fig. S15). Of the six standalone ωRNAs tested, we found that five could mediate RNA-guided DNA cleavage with a distally encoded IscB from the same genome (Fig. 3D), demonstrating that a single IscB can use multiple trans-encoded ωRNAs. Guides from many ωRNAs, both IscB-adjacent and trans-encoded, mostly target prokaryotic genomic sequences (61.5% genomic, 0.7% plasmid, 2.0% phage, 35.8% unmatched; N = 36,323), suggesting a nondefense function for IscB systems (fig. S14 and table S3). In particular, we found that more than one-third of the ωRNAs (34.1%) targeted the same locus without the IS200/IS605 transposon insertion (table S3 and fig. S16).

Evolution and diversity of IscB systems

We next investigated the evolutionary relationships among IscB, Cas9, and other homologous proteins to gain a broader insight into the evolution of RNA-guided mechanisms. In our search for proteins containing split RuvC domains, we detected another group of shorter, ~350–amino acid IscB homologs that are also encoded in IS200/IS605 superfamily transposons. These proteins contain a PLMP domain and split RuvC but lack the HNH domain. We renamed these proteins IsrB (insertion sequence RuvC-like OrfB) to emphasize their distinct domain architecture, replacing the previous designation, IscB1 (*3*). In addition to IscB and IsrB, we identified a family of even smaller proteins (~180 amino acids) that only contained the PLMP domain and HNH domain but no RuvC domain, which we named IshB (insertion sequence HNH-like OrfB).

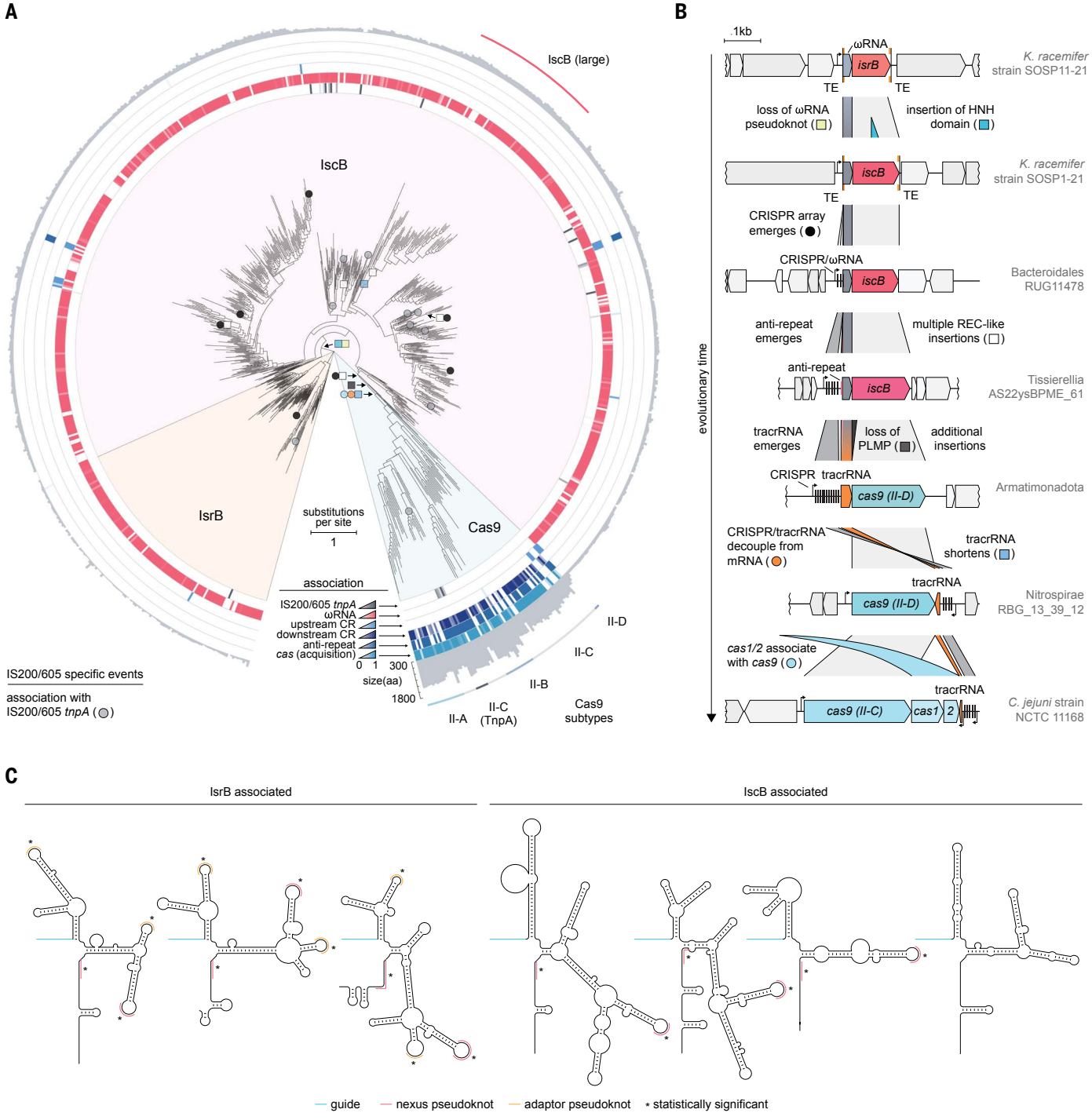

Fig. 4. Diversity and evolution of IscB. (**A**) Phylogenetic tree of IsrB, IscB, and Cas9. Associations with IS200/IS605 TnpA, ωRNA, CRISPR arrays, anti-repeats (where applicable), and *Cas* acquisition genes are shown. ORF size of cluster representative is shown on the second outermost ring. Notable groups are shown as colored arcs on the outermost ring. First occurrences of evolutionary events in each clade are marked by colored circles and squares, as described in (B). CR, CRISPR array. (**B**) Parsimonious evolutionary timeline linking IsrB to Cas9 with exemplifying loci. Colors of protein of interest indicate distinct stages in the evolution of IsrB to Cas9. (**C**) Structural diversity of ωRNAs in IsrB and IscB systems.

To investigate the relationships among these proteins, we built a maximum likelihood (ML) tree from a multiple alignment of the split RuvC nuclease and BH domains using IQ-TREE 2 (Fig. 4A, figs. S17 and S18, data S2 and S3, and table S4) (*14*). The topology of the resulting tree was supported by several additional ML and Bayesian phylogenetic and robustness analyses (figs. S17 to S25 and data S2 and S3; see supplementary text for details). In the resulting tree, IsrB, IscB, and Cas9 formed distinct, strongly supported clades, which suggests that each of these nucleases originated from a unique evolutionary event (Fig. 4A, figs. S20, C and D, S21, S22, A and C, and S23, and supplementary text). We then analyzed the associations between each protein cluster and IS200/IS605 *tnpA* genes (*3*),

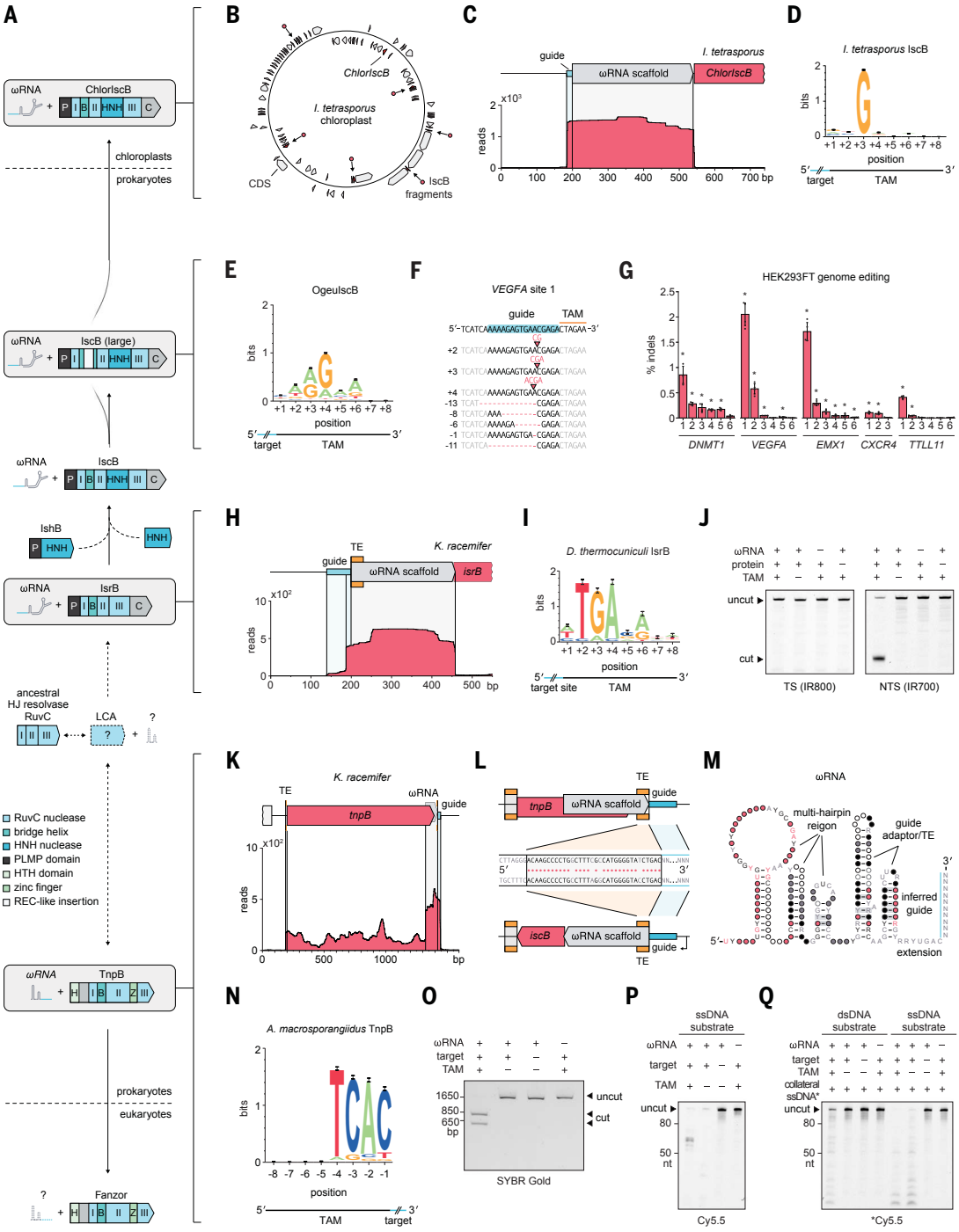

Fig. 5. Exploration of the diversity of IS200/IS605 superfamily nucleases. (**A**) Evolution between IS200/IS605 transposon superfamily–encoded nucleases and associated RNAs. Dashed lines reflect tentative/unknown relationships. LCA, last common ancestor. (**B**) Locations of IscB loci and fragments in the *I. tetrasporus* genome. Intact locus is labeled as "ChlorIscB." (**C**) Small RNA-seq of *I. tetrasporus*. (**D**) WebLogo of ChlorIscB cleavage TAM using a reprogrammed guide in an IVTT TAM screen. (**E**) WebLogo of OgeuIscB TAM using a reprogrammed guide in an IVTT TAM screen. (**F**) Targeted OgeuIscB-mediated indel formation at the *VEGFA* locus in HEK293FT cells ordered by abundance, with indel size at left. (**G**) OgeuIscB-mediated indel formation at multiple sites in HEK293T cells. Error bars denote SD. *P < 0.05. (**H**) Small RNA-seq of ωRNA from IsrB locus in *K. racemifer* strain SOSP1-21. (**I**) WebLogo of *Desulfovigula thermocuniculi* (DthIsrB) TAM using a reprogrammed guide in an IVTT TAM screen. (**J**) DthIsrB mediates ωRNA-guided nontarget strand nicking in a TAM- and target-dependent manner in an IVTT cleavage assay using 5′ strand–specific labeled targets. (**K**) Small RNA-seq of ωRNA from TnpB locus in *K. racemifer* strain SOSP1-21. (**L**) Comparison of ωRNAs from *K. racemifer* IscB and TnpB loci. (**M**) Secondary structure prediction of KraTnpB-associated ωRNA. (**N**) WebLogo of *A. macrosporangiidus* TnpB (AmaTnpB) TAM using a reprogrammed guide in an IVTT TAM screen. (**O**) In vitro reconstituted AmaTnpB cleavage of dsDNA substrates in the presence or absence of ωRNA, target, and/or TAM. (**P**) AmaTnpB performs ωRNA-guided, TAM-independent, target-dependent cleavage of 3′ Cy5.5-labeled ssDNA substrates. (**Q**) AmaTnpB cleaves a 3′ Cy5.5-labeled collateral ssDNA substrate in the presence of TAM- and target-containing dsDNA or target-containing ssDNA substrates. Contig accession and position information for all displayed loci are listed in table S6.

ωRNAs, CRISPR-Cas adaptation genes (*cas1*, *cas2*, *cas4*, and *csn2*), CRISPR arrays upstream and downstream of the respective ORF, and CRISPR anti-repeats (Fig. 4A). As discussed above, *iscB* and *isrB* were rarely associated with CRISPR arrays and were not found to be associated with CRISPR-Cas adaptation genes. The *isrB*s were associated with structurally distinct ωRNAs. The *iscB*s were flanked by transposon ends similar to those mobilized by TnpA (*3*) but were found near *tnpA* in only 56 of 2811 (2.0%) unique IscB loci (Fig. 4A and fig. S26D).

Additionally, we identified two distinct groups of Cas9s. The first is a new subtype, II-D, a group of relatively small *cas9*s (~700 amino acids) that are not associated with any other known *cas* genes (*15*). The second is a distinct clade branching from within the II-C subtype, which includes exceptionally large *cas9*s (>1700 amino acids) that are associated with *tnpA* (Fig. 4A and fig. S26). The *tnpA*-associated II-C loci often encompass unusually long DRs (more than 42 bp in length) and in some cases encode HIRAN domain proteins between the *cas9* and other *cas* genes (Fig. 4A and fig. S27). Predicted transposon ends surround various combinations of the *tnpA*, *cas* acquisition genes, and CRISPR arrays in these loci.

These phylogenetic and association analyses confirm that IS200/IS605 transposon-encoded IscBs and IsrBs share a common evolutionary history with Cas9 (supplementary text). Given the deep position of the IsrB clade in the tree (Fig. 4A) and the lack of the HNH domain, IsrBs likely represent the ancestral state, probably having evolved from the compact RuvC endonuclease (*16*). Almost all *isrB*s are associated with an ωRNA; this suggests that these systems became RNA-guided at an early stage of evolution. IsrB subsequently gained the HNH domain, possibly through insertion of another mobile element or recombination with a gene encoding an IshB-like protein, founding the IscB family (Fig. 4, A and B, turquoise squares, and supplementary text).

CRISPR arrays emerged within IscB systems on multiple, independent occasions (Fig. 4, A and B, black circles). These short arrays consist of repeats that could have evolved by duplication of segments of the ancestral ωRNA. The resulting systems encompass a hybrid CRISPR-ωRNA that consists of a CRISPR array preceding a partial ωRNA. These CRISPR-associated IscB proteins likely also gained REC-like insertions between the RuvC-I and RuvC-II subdomains on a number of occasions, often contemporaneously with or shortly after the CRISPR association (Fig. 4, A and B, white squares, and fig. S28). In particular, one CRISPR-associated IscB cluster (cluster 2089) likely founded the Cas9 family (fig. S23) upon the loss of the hallmark PLMP domain (Fig. 4, A and B, gray square, and fig. S28). Moreover, the tracrRNAs of subtype II-D, a deep branch in the Cas9 subtree (ML branch support, ≥97/100; Bayesian posterior probability, 100%; figs. S20, B to D, and S23), shows significant similarity to IscB ωRNAs (E-value 4.1×10^{-8}), which suggests that the Cas9 tracrRNA originally evolved from ωRNA (fig. S29). The continued evolution of Cas9 likely involved the gain of additional REC-like insertions between the bridge helix and the RuvC-II domains, resulting in increased protein size (fig. S28). Finally, upon the association with the CRISPR adaptation machinery (*cas1*, *cas2*, and possibly *cas4*) (Fig. 4, A and B, light blue circles), a burst of Cas9 diversification and widespread dispersion among bacteria via horizontal gene transfer followed, resulting in the evolution of multiple type II CRISPR subtypes.

We also explored the evolutionary history of ωRNAs. By iteratively building a set of ωRNA profiles that spanned all major groups of ωRNAs associated with *iscB*s and *isrB*s, we found that diverse ωRNAs are associated with almost all *iscB*s and *isrB*s. Moreover, different IsrB and IscB clades are associated with distinct ωRNA structures (Fig. 4, A and C, and figs. S18A, S24A, and S30). The transition from *isrB* to *iscB* was likely accompanied by loss of a second pseudoknot, the adaptor pseudoknot, between the transposon end region and the multi-stem loop in *isrB*-associated ωRNAs (Fig. 4, A to C, yellow square). The inverse relationship between the complexity of the ωRNA structure and the associated protein size is also reflected by the simplified ωRNA structures associated with clades of large IscBs and the even smaller tracrRNAs associated with large Cas9s (Fig. 4C and fig. S30).

IS200/IS605 elements encode diverse RNA-guided nucleases

In addition to the distinct succession of evolutionary events that yielded the abundant and diverse type II CRISPR systems, our phylogenetic analysis revealed several other events in the evolution of IscB and related proteins that led to the extant diversity, which we sought to experimentally explore.

First, we searched for IscB homologs in eukaryotic genomes and identified multiple *iscB* loci in the chloroplast genome of *Ignatius tetrasporus* UTEX B 2012, a terrestrial green alga (Fig. 5, A and B, and fig. S31). Although the ORF is disrupted by multiple stop codons in most of these loci, one locus encodes an intact IscB (~50% amino acid identity to related prokaryotic IscBs) and a transcriptionally active ωRNA (Fig. 5C). This eukaryotic IscB cleaves DNA with a minimal NNG TAM (Fig. 5D), which differs from other characterized IscB TAMs (fig. S8).

Second, we investigated the clade of large IscBs, which contain a BH domain that is split in two by REC domain–like insertions (Figs. 4A and 5A, white squares). We hypothesized that these insertions might enhance DNA unwinding, similarly to the REC lobe of Cas9 (*17*), and would therefore facilitate genome editing in the complex landscape of eukaryotic chromatin structure. We screened six large IscB proteins, using a pool of 12 guides each, for their ability to generate insertions/deletions (indels) in HEK293FT cells (table S5 and supplementary materials); one (OgeuIscB) produced appreciable indels (Fig. 5, E and F, and fig. S32A). To further examine OgeuIscB activity, we tested a range of guide lengths targeting three loci in the human genome and found that OgeuIscB achieved the maximum indel rate with a 16-nt guide (fig. S32B). On a panel of 46 sites in the human genome, we found that OgeuIscB induced indels at 28 of these sites with varying efficiency up to 4.4% (Fig. 5G, fig. S32C, and table S5). Thus, OgeuIscB seems a promising candidate for further development of IscB-based genome editing tools.

Third, we experimentally characterized the putative nuclease activity of IsrB, the apparent ancestor of IscB (Fig. 5A). *K. racemifer* contains five *isrB*s associated with ωRNAs that are natively expressed (Fig. 5H and fig. S33). We found that the IsrB-ωRNA RNP nicks the nontarget strand of a dsDNA substrate in a guide- and TAM-specific manner (Fig. 5, I and J, and fig. S34), which is analogous to the activity of IscB upon inactivation of the HNH domain (Fig. 2F).

Finally, we sought to determine whether IS200/IS605 transposons in general harbor RNA-guided nucleases. In addition to the distinct IscB and IsrB families, most IS200/IS605 transposons encode RuvC-like endonucleases of another family, TnpB, which is thought to be the ancestor of Cas12s, the type V CRISPR effectors (Fig. 5A) (*5*). Additionally, TnpB is the likely ancestor of larger proteins, Fanzors, encoded in diverse eukaryotic transposons (Fig. 5A) (*18*). The TnpB family, including Fanzor, is an order of magnitude more diverse than the IscB family; an HMMER search identified more than 1 million *tnpB* loci in publicly available prokaryotic genomes.

We identified conserved noncoding regions immediately downstream of the coding sequence (CDS) of many *tnpB*s, suggesting the presence of associated ncRNAs that could function as RNA guides (fig. S35). Previous work has identified ncRNAs overlapping the 3′-end of *tnpB* genes in archaea and bacteria (*19*, *20*), but the function of these ncRNAs has not been characterized. Small RNA-seq of *K. racemifer* revealed native expression of a ncRNA overlapping the 3′ end of the associated *tnpB* ORF (Fig. 5K), which we classified as a distinct group of ωRNAs. The reverse complement of the KraTnpB ωRNA 3′ end is nearly identical to the 5′ of the ωRNA associated with

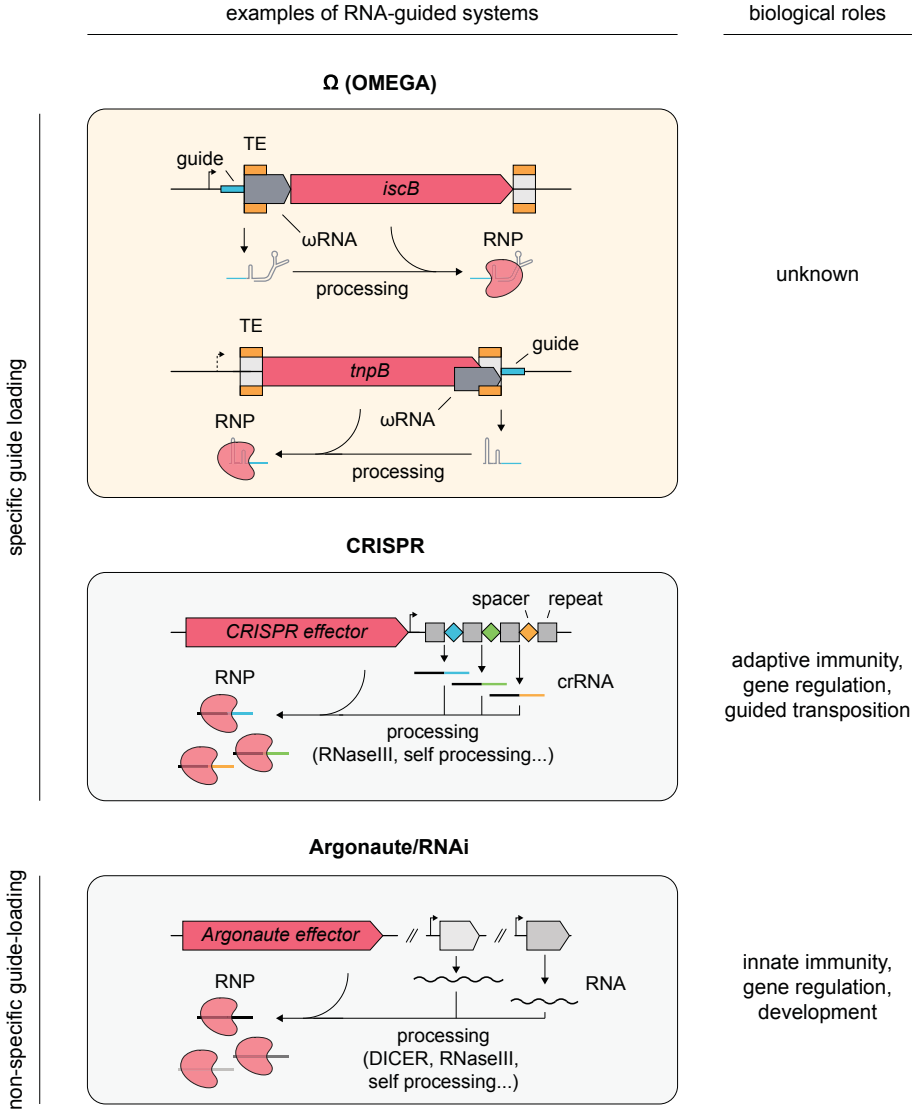

Fig. 6. Naturally occurring RNA-guided DNA-targeting systems. Comparison of OMEGA systems with other known RNA-guided systems. In contrast to CRISPR systems, which capture spacer sequences and store them in the locus within the CRISPR array, OMEGA systems may transpose their loci (or trans-acting loci) into target sequences, converting targets into ωRNA guides.

some KraIscBs, a region that corresponds to the predicted transposon end in each locus (Fig. 5L).

Analysis of nonredundant loci containing *tnpB* genes that clustered with KraTnpB showed a drop of sequence conservation at the 3′ end of the loci (fig. S35), corresponding to the IS200/IS605 transposon end. Comparison to the small RNA-seq trace revealed expression beyond the conservation drop, indicating possible presence of a guide sequence in the transcript (Fig. 5M). In vitro plasmid cleavage assays for multiple TnpB proteins from this cluster using a reprogrammed guide demonstrated RNA-guided cleavage with a 5′ TAM (Fig. 5N and fig. S36). We recombinantly purified a TnpB from *Alicyclobacillus macrosporangiidus* (AmaTnpB) and confirmed its reprogrammable RNA-guided dsDNA endonuclease activity (Fig. 5O and fig. S36). We also observed that AmaTnpB robustly cleaved target-containing single-stranded DNA (ssDNA) substrates (Fig. 5P) and nonspecifically cleaved a collateral substrate upon recognition of dsDNA or ssDNA substrates (Fig. 5Q).

Discussion

Naturally programmable biological systems offer an efficient solution for diverse organisms to achieve scalable complexity via modularity of their components. RNA-guided defense and regulatory systems, which are widespread in prokaryotes and eukaryotes, are a prominent case in point, and have served as the basis of numerous biotechnology applications thanks to the ease with which they can be engineered and reprogrammed (*21*–*23*).

Here, through the exploration of Cas9 evolution, we discovered the programmable RNA-guided mechanism of three highly abundant but previously uncharacterized transposon-encoded nucleases: IscB, IsrB, and TnpB, which we collectively refer to as OMEGA (obligate mobile element–guided activity) (Fig. 6) because the mobile element localization and movement likely determines the identity of their guides. Although the biological functions of OMEGA systems remain unknown, several hypotheses are compatible with the available evidence, including roles in facilitating TnpA-catalyzed, RNA-guided transposition, or acting as a toxin, with the transposon acting as the antitoxin, securing maintenance of IS200/IS605 insertions (supplementary text).

The broad distribution of the OMEGA systems characterized here indicates that RNA-guided mechanisms are more widespread in prokaryotes than previously suspected and suggests that RNA-guided activities are likely ancient and evolved on multiple, independent occasions, of which only the most common ones have likely been identified so far. The TnpB family is far more abundant and diverse than the IscB family; indeed, we identified more than 1 million putative *tnpB* loci in bacterial and archaeal genomes, making it one of the most common prokaryotic genes. These TnpBs might represent an untapped wealth of diverse RNA-guided mechanisms present not only in prokaryotes but also in eukaryotes. Combined with our identification of a chloroplast-encoded IscB, these findings suggest that the expansion of RNA-guided systems into eukaryotic genomes could be a general phenomenon, and more broadly, that RNA-guided systems are functionally diverse and permeate all domains of life.

REFERENCES AND NOTES

1. F. Zhang, *Q. Rev. Biophys.* **52**, e6 (2019).
2. F. Hille *et al.*, *Cell* **172**, 1239–1259 (2018).
3. V. V. Kapitonov, K. S. Makarova, E. V. Koonin, *J. Bacteriol.* **198**, 797–807 (2015).
4. P. Siguier, E. Gourbeyre, M. Chandler, *FEMS Microbiol. Rev.* **38**, 865–891 (2014).
5. S. Shmakov *et al.*, *Nat. Rev. Microbiol.* **15**, 169–182 (2017).
6. B. Zetsche *et al.*, *Cell* **163**, 759–771 (2015).
7. E. Deltcheva *et al.*, *Nature* **471**, 602–607 (2011).
8. I. Kalvari *et al.*, *Nucleic Acids Res.* **49**, D192–D200 (2021).
9. Z. Weinberg, J. Perreault, M. M. Meyer, R. R. Breaker, *Nature* **462**, 656–659 (2009).
10. M. Jinek *et al.*, *Science* **337**, 816–821 (2012).
11. G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2579–E2586 (2012).
12. G. Gasiunas *et al.*, *Nat. Commun.* **11**, 5512 (2020).
13. M. Jinek *et al.*, *Science* **343**, 1247997 (2014).
14. B. Q. Minh *et al.*, *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

15. K. S. Makarova *et al*., *Nat. Rev. Microbiol.* **18**, 67–83 (2020).
16. K. A. Majorek *et al*., *Nucleic Acids Res.* **42**, 4160–4179 (2014).
17. H. Nishimasu *et al*., *Cell* **156**, 935–949 (2014).
18. W. Bao, J. Jurka, *Mob. DNA* **4**, 12 (2013).
19. J. V. Gomes-Filho *et al*., *RNA Biol.* **12**, 490–500 (2015).
20. Z. Weinberg *et al*., *Nucleic Acids Res.* **45**, 10811–10823 (2017).
21. A. Hüttenhofer, P. Schattner, *Nat. Rev. Genet.* **7**, 475–482 (2006).
22. A. Schneider, *EMBO Rep.* **21**, e51918 (2020).
23. E. V. Koonin, *Biol. Direct* **12**, 5 (2017).
24. H. Altae-Tran, S. Kannan, F. Zhang, Code and processed data for: The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases (Version 1.0). Zenodo 10.5281/zenodo.5168777.

ACKNOWLEDGMENTS

We thank J. Strecker, S. Hirano, and D. Strebinger for advice regarding biochemistry experiments, G. Faure for advice regarding computational analyses, and all members of the Zhang lab for helpful discussions. We are grateful to the following individuals for generously providing access to their metagenomic data (IMG accessions in parentheses): B. Campbell (IMG3300025818 and IMG3300007960), A. Buchan (IMG3300017968), and E. Edwards (IMG3300020812 and IMG3300023203). We appreciate assistance in DNA extraction and troubleshooting from M. Forbes for the Yellowstone Lake metagenomes. Yellowstone Lake samples were collected with support from the National Park Service, Yellowstone National Park (Research Permit YELL-2016/17-SCI-7018). **Funding:** Supported by NSF Integrated Earth Systems grant subaward A101357 (L.J.M. and W.P.I.); NSF Division of Environmental Biology grant 1950770 (M.D. and W.P.I.); Department of Energy–Joint Genome Institute grant CSP 1675 (W.P.I.); the National Library of Medicine (K.M.S. and E.V.K.); and NIH grants 1R01-HG009761 and 1DP1-HL141201, the Howard Hughes Medical Institute, the Open Philanthropy Project, the Harold G. and Leila Mathers Foundation, the Edward Mallinckrodt Jr. Foundation, the Poitras Center for Psychiatric Disorders Research at MIT, the Hock E. Tan and K. Lisa Yang Center for Autism Research at MIT, the Yang-Tan Center for Molecular Therapeutics at MIT, and the Phillips family, R. Metcalfe, and J. and P. Poitras (F.Z.). **Author contributions:** H.A.-T., S.K., and F.Z. conceived of the project. H.A.-T., S.K., F.E.D., R.O., S.P.N., K.S.M., E.V.K., and F.Z. designed and performed experiments. L.J.M., M.D., and W.P.I. collected metagenomic data. F.Z. supervised the research and experimental design with support from R.K.M. H.A.-T., S.K., R.K.M., E.V.K., and F.Z. wrote the manuscript with input from all authors. **Competing interests:** H.A.-T., S.K., F.E.D., S.P.N., and F.Z. are co-inventors on U.S. provisional patent applications filed by the Broad Institute related to this work. F.Z. is a cofounder of Editas Medicine, Beam Therapeutics, Pairwise Plants, Arbor Biotechnologies, and Sherlock Biosciences. **Data and materials availability:** Sequences of genes used in the experimental studies are available via online sequence repositories; expression plasmids listed in table S1 are available from Addgene under a material transfer agreement with the Broad Institute. Raw reads from microbial small RNA-seq are available on SRA under BioProject PRJNA744508. Scripts for data analysis and visualization are available at Zenodo (*24*). Additional information is available at the Zhang lab website (https://zhanglab.bio).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.abj6856
Materials and Methods
Supplementary Text
Figs. S1 to S36
Tables S1 to S6
Data S1 to S4
References (*25*–*65*)
MDAR Reproducibility Checklist

View/request a protocol for this paper from *Bio-protocol*.

26 May 2021; accepted 9 August 2021
Published online 9 September 2021
10.1126/science.abj6856

The widespread IS200/IS605 transposon family encodes diverse programmable RNA-guided endonucleases

Han Altae-TranSoumya KannanF. Esra DemircioğluRachel OshiroSuchita P. NetyLuke J. McKayMensur Dlaki#William P. InskeepKira S. MakarovaRhiannon K. MacraeEugene V. KooninFeng Zhang

Science, 374 (6563), • DOI: 10.1126/science.abj6856

Tracing the origin of CRISPR-Cas

CRISPR-Cas systems have transformed genome editing and other biotechnologies; however, the broader origins and diversity of RNA-guided nucleases have largely remained unexplored. Altae-Tran *et al*. show that three distinct transposon-encoded proteins, IscB, IsrB, and TnpB, are naturally occurring, reprogrammable RNA-guided DNA nucleases (see the Perspective by Rousset and Sorek). In addition to identifying diverse guide-encoding mechanisms, the authors elucidate the evolutionary relationship between IsrB, IscB, and CRISPR-Cas9. Overall, these newly characterized systems, called OMEGA (for obligate mobile element–guided activity) systems, are found in all domains of life and may be harnessed for biotechnology development. —DJ