Published online 11 April 2014

Nucleic Acids Research, 2014, Vol. 42, No. 10 ***6091–6105***
doi: 10.1093/nar/gku241

SURVEY AND SUMMARY

Classification and evolution of type II CRISPR-Cas systems

Krzysztof Chylinski[1,2], Kira S. Makarova[3], Emmanuelle Charpentier[1,4,5] and Eugene V. Koonin[3,*]

[1]The Laboratory for Molecular Infection Medicine Sweden (MIMS), Umeå Centre for Microbial Research (UCMR), Department of Molecular Biology, Umeå University, Umeå 90187, Sweden, [2]Max F. Perutz Laboratories, University of Vienna, Vienna 1030, Austria, [3]National Center for Biotechnology Information, NLM, National Institutes of Health, Bethesda, MD 20894, USA, [4]Helmholtz Centre for Infection Research, Department of Regulation in Infection Biology, Braunschweig 38124, Germany and [5]Hannover Medical School, Hannover 30625, Germany

Received January 13, 2014; Revised March 11, 2014; Accepted March 12, 2014

ABSTRACT

The CRISPR-Cas systems of archaeal and bacterial adaptive immunity are classified into three types that differ by the repertoires of CRISPR-associated (*cas*) genes, the organization of *cas* operons and the structure of repeats in the CRISPR arrays. The simplest among the CRISPR-Cas systems is type II in which the endonuclease activities required for the interference with foreign deoxyribonucleic acid (DNA) are concentrated in a single multidomain protein, Cas9, and are guided by a co-processed dual-tracrRNA:crRNA molecule. This compact enzymatic machinery and readily programmable site-specific DNA targeting make type II systems top candidates for a new generation of powerful tools for genomic engineering. Here we report an updated census of CRISPR-Cas systems in bacterial and archaeal genomes. Type II systems are the rarest, missing in archaea, and represented in ∼5% of bacterial genomes, with an over-representation among pathogens and commensals. Phylogenomic analysis suggests that at least three *cas* genes, *cas1, cas2* and *cas4*, and the CRISPR repeats of the type II-B system were acquired via recombination with a type I CRISPR-Cas locus. Distant homologs of Cas9 were identified among proteins encoded by diverse transposons, suggesting that type II CRISPR-Cas evolved via recombination of mobile nuclease genes with type I loci.

INTRODUCTION

The CRISPR (clustered regularly interspaced short palindromic repeats)-Cas (CRISPR-associated genes) system is the first and so far the only adaptive immunity system discovered in archaea and bacteria. Similar to restriction–modification and several other defense systems, it is based on the self–non-self discrimination principle but unlike other defense mechanisms, CRISPR-Cas imprints pieces of genetic material as a memory of previously encountered viruses and plasmids [(1) and references therein]. These short fragments matching the infectious agent deoxyribonucleic acid (DNA) are inserted into an array of CRISPR repeats and are used in the form of guide CRISPR RNA (crRNA) to target the cognate virus or plasmid upon new infections (2–5).

The diversity of the protein components associated with the CRISPR-Cas systems is remarkable (6–9), but identification of several signature elements in their genomic organization has provided for a relatively simple classification of these systems (5). This classification differentiates three major types of the CRISPR-Cas systems, with *cas3*, *cas9* (formerly *csn1*) and *cas10* being the signature genes for the type I, type II and type III systems, respectively. Moreover, recent bioinformatic analysis and structural studies revealed striking similarity in the organization of effector complexes between CRISPR-Cas systems of type I and type III suggesting their origin from a common ancestor (10–12). However, the origin of type II CRISPR-Cas systems remains obscure.

According to the current classification, two subtypes of type II CRISPR-Cas systems are defined on the basis of their characteristic operon organizations (5). Type II-A systems encompass an additional gene, known as *csn2*. The

*To whom correspondence should be addressed. Tel: 301-435-5913; Fax: 301-435-7793; Email: koonin@ncbi.nlm.nih.gov

Csn2 protein is involved in spacer integration but is not required for interference (13,14). Two distantly related Csn2 subfamilies that include, respectively, short (15,16) and long forms (17) of the protein have been identified. For both of these forms, the structures have been solved and the proteins have been biochemically characterized (15–18). The Csn2 structures have been shown to adopt a highly derived P-loop ATPase fold in which the adenosine triphosphate (ATP) binding center appears to be inactivated (15–17). The structurally characterized Csn2 proteins form homotetrameric rings that bind linear double-stranded (ds) DNA through the rings central hole. The major difference between the short and long forms of Csn2 is that the short but not the long forms display Ca^{2+}-dependent DNA binding (15,16,19). The specific function of Csn2 in spacer integration remains unclear, but it has been hypothesized that Csn2 is an accessory component that binds the dsDNA ends and protects them from exonucleolytic degradation, while recruiting DNA-repair proteins to seal the breaks (18). Type II-B systems lack the *csn2* gene but possess another additional gene of the *cas4* family (5). The Cas4 proteins belong to the PD-(D/E)xK family of nucleases (8) and indeed have been shown to possess 5′-single-stranded DNA exonuclease activity (20). Similar to Csn2, the actual role of Cas4 proteins in the CRISPR-Cas systems remains unknown. Unlike *csn2*, which is found only in association with the type II-A system, the *cas4* gene is not tightly linked to a particular CRISPR-Cas defense mechanism and thus has been proposed to play a role in associated immunity (21). Since the original classification was developed, numerous additional type II CRISPR-Cas systems with only three genes (*cas1*, *cas2* and *cas9*) in the operon have been identified and a third subtype, type II-C, has been proposed (11,22).

Type II CRISPR-Cas is considered to be the minimal CRISPR-Cas system that includes the CRISPR repeat–spacer array and only four (but often three) *cas* genes (Figure 1), although additional bacterial factors, in particular trans-activating crRNA (tracrRNA) and RNase III, contribute to the function of this system (23). All type II CRISPR-Cas modules contain the pair of *cas1* and *cas2* genes that are required for spacer acquisition (24–26). These two genes are also present in the vast majority of the genomes that contain at least one CRISPR-Cas system of type I or type III. Similar to type I, type II CRISPR-Cas systems require a well-defined short protospacer adjacent motif (PAM) that is located immediately downstream of the protospacer on the non-target DNA strand (27–30). The PAM sequence is important both for spacer acquisition and for target recognition and cleavage (14,27,30–33). The *cas9* gene, the signature of type II systems (5), is a large multidomain protein that alone is sufficient for targeting and cleaving the invader DNA (23,33). Two readily identifiable, split nuclease domains of Cas9, namely HNH and RuvC-like nucleases, are required for target DNA cleavage (8,12,27,30,32,33).

Surprisingly, the minimal type II CRISPR-Cas system employs an elaborate, unique processing mechanism of precursor crRNA (pre-crRNA) (Figure 1). Unlike most other systems of types I and III that use a dedicated Cas endoribonuclease to cleave pre-crRNA by recognizing the repeat units (12,34–37), type II systems use endogenous RNase III and a specialized ribonucleic acid (RNA) molecule, denoted as tracrRNA. The tracrRNAs have been identified in most genomes encoding type II systems and are now considered to be an integral component of this CRISPR-Cas type (22,23). These RNAs are encoded in the vicinity of or within the *cas* operon and CRISPR array, and all tracrRNA orthologs are characterized by the presence of an anti-repeat sequence homologous to cognate CRISPR repeats. A first processing event involves base-pairing of tracrRNA with the pre-crRNA repeats in the presence of Cas9 to form RNA duplexes that are cleaved by the executioner endogenous RNase III (23). The intermediate crRNA species undergo further maturation resulting in mature individual crRNAs that remain duplexed with tracrRNA in a complex with the Cas9 protein (23,33).

Several type II systems show additional, distinct features (Figure 1). For example, it has been reported that in type II-C of *Neisseria meningitidis* WUE2594, RNase III is dispensable for the DNA targeting function of CRISPR-Cas (38). Certain mature crRNAs have been shown to be transcribed from promoters embedded within each CRISPR repeat and thus would be produced without the requirement of any maturation event (38). Furthermore, an additional RNA component of type II-B in *Francisella novicida* U112, denoted as scaRNA (small, CRISPR-Cas-associated RNA), has recently been discovered. Together with scaRNA and Cas9, tracrRNA is involved in the repression of a messenger RNA (mRNA) encoding a lipoprotein implicated in the virulence of *F. novicida*, a notable case of a non-immune function of the CRISPR-Cas system (39). A link of type II CRISPR-Cas to virulence has also been described in *Campylobacter jejuni* (40), *N. meningitidis* (39) and *Legionella pneumophila* (41).

Given their simplicity and potential for easy RNA programming, type II is the best candidate of all CRISPR-Cas systems for exploitation in genetic engineering. It has been shown that the dual-tracrRNA:crRNA can be replaced with a single guide RNA that combines the two RNA molecules, eliminating the maturation steps required for RNA-programmable Cas9 activation (33). Thus, a variety of guide RNAs can be designed to direct the Cas9 endonuclease for site-specific DNA cleavage and further genetic manipulations such as gene editing, insertion or deletions (32,33). The easy conversion of Cas9 into a nickase was utilized to facilitate homology directed repair in mammalian genomes with reduced mutagenic activity and reported increased specificity (42–45). Furthermore, the DNA-binding capacity of a catalytically inactive Cas9 mutant can be exploited to engineer diverse RNA-programmable devices that can be used to mediate transcriptional silencing or activation, or as DNA modification tools (32,33,46–51). The unprecedented versatility in alternatives of genome engineering and modulation of gene expression makes RNA-programmable Cas9 a unique technology in molecular biology. At the time of this writing, systems for eukaryotic gene targeting using type II CRISPR-Cas systems have been developed for human cells (22,42,45,52,53), monkey (54), pig (55), mice (56,57), zebrafish (58), *Drosophila* (59), yeast (60), plants (61,62) and *Caenorhabditis elegans* (63), as well as bacteria (64). The successful, rapid application of sequence-specific RNA-programmable Cas9 for genome editing in

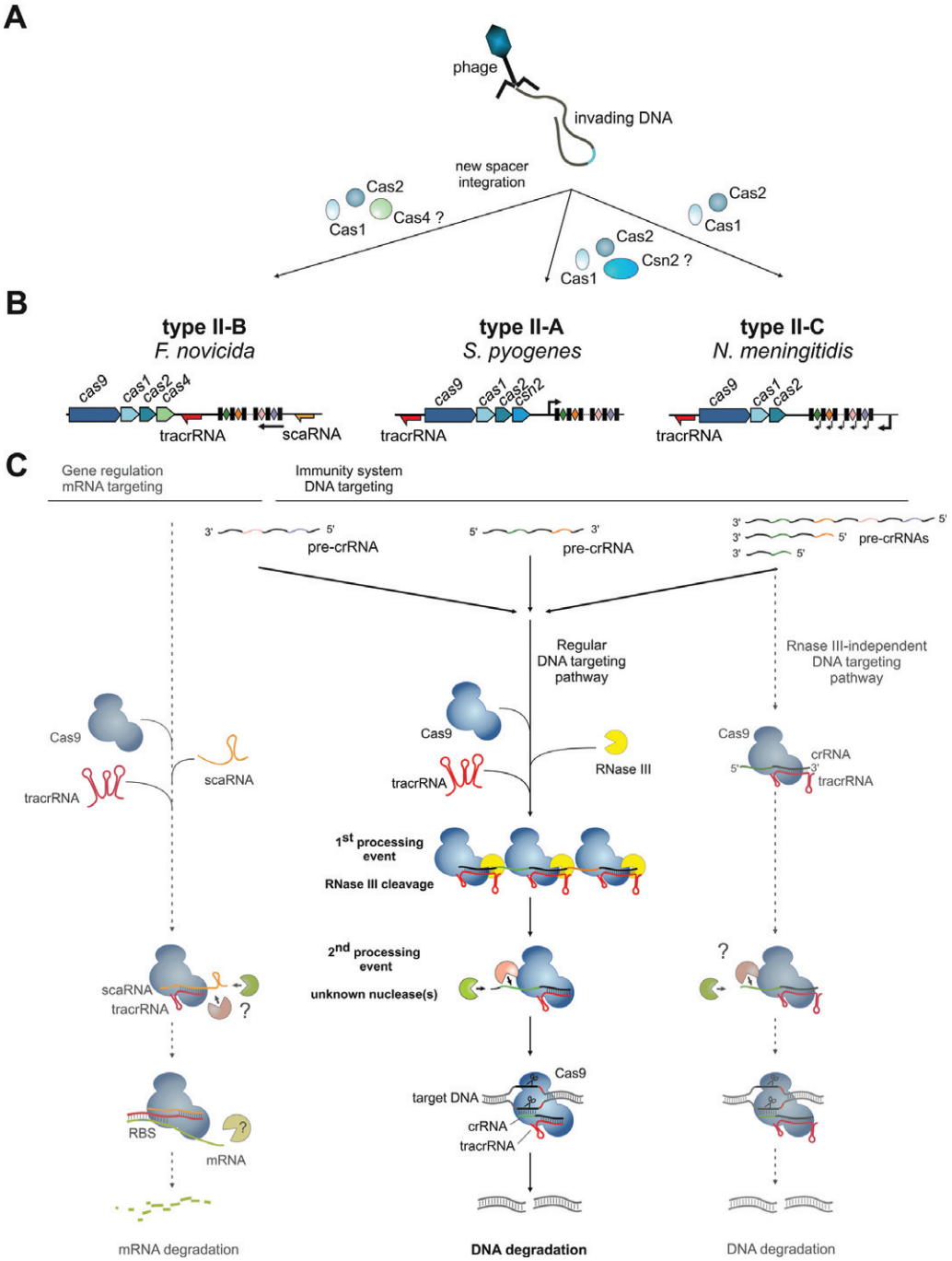

Figure 1. General scheme of the mechanism of type II CRISPR-Cas systems. (**A**) Proteins responsible for new spacer acquisition are shown for different type II subtypes. (**B**) Typical type II CRISPR-Cas locus architecture for three major subtypes shown together with a representative strain locus scheme. Red and orange arrows: tracrRNA and scaRNA with transcription direction indicated, respectively; black rectangles: repeats; diamonds: spacers; red rectangles: degenerated repeats; black arrows: pre-crRNA promoters. In type II-B, the localization of the pre-crRNA promoter in relation to the scaRNA is not known (see the paragraph 'Role of type II CRISPR-Cas in virulence and origin of scaRNA'); the arrow represents only the direction of pre-crRNA transcription. Note the differences in the loci architecture with respect to *cas* gene composition, tracrRNA and repeat–spacer array transcription orientation and tracrRNA position. (**C**) Mechanisms of type II CRISPR-Cas systems. The classical DNA targeting pathway, common to all type II CRISPR-Cas systems (middle), involves co-processing of Cas9-stabilized tracrRNA:pre-crRNA duplexes by RNase III upon binding of tracrRNA anti-repeat to the pre-crRNA repeat, followed by trimming of crRNA by a yet unknown mechanism. The mature tracrRNA:crRNA guides the Cas9 endonuclease to introduce site-specifically dsDNA breaks in the invading DNA. The mechanism shown here for the type II-A of *S. pyogenes* was also shown for the type II-A of *S. thermophilus* (22,51). The alternative DNA targeting mechanism (right), described in type II-C of *N. meningitidis* (38), does not involve RNase III co-processing due to transcription of a short crRNA directly from an upstream repeat-encoded promoter. In type II-B of *F. novicida* (39), the system evolved to possibly target endogenous mRNA expression (left). We hypothesize that similar to tracrRNA:crRNA-Cas9, the tracrRNA:scaRNA-Cas9 complex is first formed. The scaRNA in the complex would undergo trimming by unknown nucleases [the size of most abundant scaRNA forms is shorter than predicted (39) according to RNAseq data (not shown)]. The tracrRNA:scaRNA-Cas9 further recognizes mRNA upon binding of the tracrRNA 3′ region to the target mRNA leading to its degradation by an unknown mechanism.

a broad variety of cells and organisms demonstrates the power of the system that upon further refinement could supersede such popular genome engineering tools as zinc finger nucleases and Transcription Activator-Like Effector Nucleases (TALENs) (65,66).

Given the high potential of RNA-guided Cas9 as a tool for genome manipulation, the diversity of the type II CRISPR-Cas systems across bacterial genomes is of special interest. We recently demonstrated high variability among Cas9, dual-RNA structure and PAM sequences (22,31). In addition, we characterized functional exchangeability among dual-RNA and Cas9 orthologs according to their phylogenetically defined co-evolution (31). Thus, the collection of bacterial dual-RNA–Cas9 complexes associated with diverse specific PAM sequences broadens the functional capabilities of the toolbox for multiplex engineering. Here we present an update on comparative genomics and phylogenetic analysis of type II CRISPR-Cas systems and develop a hypothesis on the origin and evolution of all major components of these systems.

Type II CRISPR-Cas loci in bacterial genomes

As indicated above, *cas9* is the signature gene of type II CRISPR-Cas systems. The typical domain organization of Cas9 proteins is shown in Figure 2. Due to the abundance of the two individual Cas9 nuclease domains in prokaryotic genomes outside of the CRISPR-Cas loci, Cas9 alone is not a suitable probe for computational detection of type II CRISPR-Cas systems, especially given that stand-alone distant Cas9 homologs have been identified (12). To confidently predict the presence of a complete type II CRISPR-Cas system in a genome, other components, such as *cas1* and *cas2* genes, cognate CRISPR repeats and tracrRNA, have to be identified in the same locus, in addition to the presence of the two nuclease domains within Cas9, although functionality of some apparently incomplete loci cannot be ruled out. Here we present comparative genomic data for a curated set of Cas9 sequences (including stand-alone *cas9* genes) identified in currently available complete bacterial genomes (Supplementary Table S1).

To explore statistical trends in the distribution of type II systems in bacteria, we selected a representative set of 661 genomes (the representative with the largest genome was chosen for each genus, to minimize the potential bias toward extensively sequenced groups) and assigned the three major CRISPR-Cas types to each of these genomes (Supplementary Table S2). Of the three types, type II is the least abundant and is present only in ∼5% of the selected genomes (32 in the analyzed representative set), compared with ∼40% for type I and ∼12% for type III. Of the 32 bacterial genomes that encompass type II CRISPR-Cas, 10 possess a type II-A system, 21 have a type II-C system and only one has a type II-B system. Type II is the only major type of CRISPR-Cas systems that has not been detected in archaea. Although type II CRISPR-Cas is relatively rare even in bacteria, statistical analysis for the subset of these genomes that encode at least one CRISPR-Cas system of any type indicates that the presence of type II in bacteria but not in archaea is not a random fluctuation (Chi-square P value = 0.005). The exclusive presence of type II CRISPR-Cas systems in bacteria is compatible with the involvement of RNase III, a primarily bacterial enzyme [except for the presence of this enzyme in several mesophilic archaea (67)], in the function of these variants of CRISPR-Cas. In contrast, random distribution of type II systems across major bacteria phyla cannot be rejected. However, in both the representative set and a larger set of complete and draft genomes (5865) (that was not used for statistical analysis because, unlike the representative set of 661 genus representative genomes, these genomes could not be considered independent), the type II CRISPR-Cas system was not identified in many phyla, namely Cyanobacteria, Chlorobi, Chloroflexi (three major groups of photosynthetic bacteria), Thermotogae, Aquificae, Deinococcus-Thermus and Chlamydia (Supplementary Table S3).

We also detected statistically significant over-representation of type II systems in host-associated (parasitic or commensal) bacteria (P = 4.3E−08) and under-representation in thermophilic bacteria (P = 0.0019). These observations suggest that environmental traits substantially contribute to the distribution of type II CRISPR-Cas systems among bacteria. In particular, horizontal gene transfer (HGT) of this system might have been favored in environments with diverse bacterial communities such as animal-associated microbiomes. Conversely, given that archaea lack type II CRISPR-Cas systems, HGT from archaeal to bacterial thermophiles appears to be out of the question, in accordance with the under-representation of type II among the latter. In agreement with these findings, indications of the presence of type II CRISPR-Cas systems in diverse human- and animal-associated microbiomes have been reported (68–70).

Another notable observation is the narrow spread of type II-B systems (because of its low abundance, it is not currently amenable for statistical analysis). This system is present only in some representatives of several bacterial genera, namely *Francisella*, *Parasutterella*, *Sutterella*, *Legionella*, *Wolinella* (Proteobacteria) and *Leptospira* (Spirochaeta), most of which are pathogens or commensals (Supplementary Table S1).

It is not uncommon for type II systems to be present in a genome along with other CRISPR-Cas systems. In our representative set of genomes, type II co-occurred with type I CRISPR-Cas systems in eight genomes, with type III systems in two genomes, and with both in the genome of the alphaproteobacterium *Tistrella mobilis* that encompasses subtypes III-A, I-C and II-C. The only other example of all three systems present in one genome is *Azospirillum* B510 (III-A, I-C and II-C) from the full genome set. In addition, some *Streptococcus thermophilus* strains that are widely employed as models in CRISPR research encompass two type II-A systems (that belong to two distinct groups; see details below) along with a type III-A system in LMD9 (23,27), and with a type I-E and a type III-A in DGCC7710 (71,72). In *Streptococcus pyogenes* SF370, similar to some other *S. pyogenes* strains, type II-A also is present along with I-C systems (23). In addition, a fusion between type I-C and type II-C systems has been detected in the genome of the uncultured Termite group 1 bacterium phylotype Rs D17 (72).

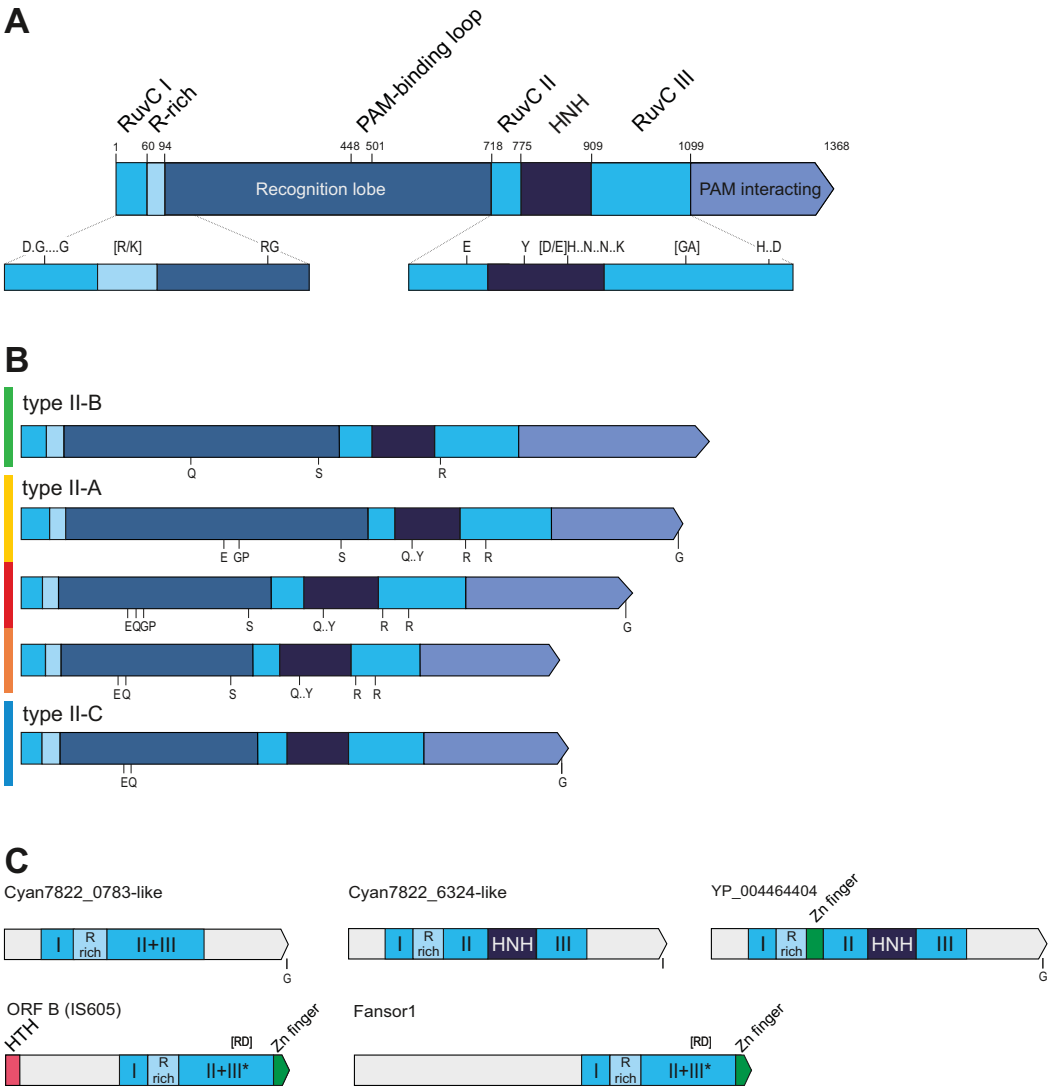

Figure 2. Schematic representation of Cas9 domain organization, motifs and relationships with distant homologs. (**A**) A general view of the domain architecture of Cas9. (**B**) Comparison of the domain organizations and conserved sequences motifs between the major groups of Cas9 proteins. (**C**) Domain architectures of distant homologs of Cas9. Homologous regions are shown by the same color. Compare with Supplementary Figure S8. The *S. pyogenes* Cas9 schematic representation with domains and domain boundaries according to the Cas9 structures (76,77) is shown in (A). See Supplementary Figure S4. Distinct sequence motifs are denoted by the corresponding conserved amino acid residues. The residues indicated in (A) are conserved in all five Cas9 groups and in (B), within the given subtype. Compare with Supplementary Figure S4. The size of a domain or a distinct region is roughly proportional to the length and the motifs are shown in accordance with their approximate position within a respective protein. The scheme was derived from the multiple alignments of each group. The color code to the left of the protein schematics in (B) corresponds to the major branches of the Cas9 phylogenetic tree in Figure 4. HTH: helix turn helix DNA-binding domain; R-rich: arginine-rich region; HNH: nuclease of the corresponding family.

Evolution of type II CRISPR-Cas subtypes

The *cas1* gene phylogeny and *cas* operon organization are pivotal for the classification of subtypes of CRISPR-Cas systems. In a previous analysis, the majority of type II systems formed a clade in the Cas1 tree (5). With many more sequenced genomes now available, we updated this analysis with a focus on the evolution of distinct subtypes of type II CRISPR-Cas.

The first question we addressed was whether the monophyly of all three subtypes of type II was supported by the Cas1 phylogeny. We generated a representative set of Cas1 sequences from all completely sequenced genomes to reconstruct a multiple alignment and a phylogenetic tree (Figure 3A and Supplementary Figures S1 and S2; see Supplementary Materials and Methods for details). The resulting tree strongly supports monophyly of the Cas1 sequences of type II, with the exception of type II-B. The type II-B Cas1 sequences form a clade within a subtree that consists mostly of type I-A but also includes representatives of other groups from type I and type III (Figure 3A). Furthermore, several conserved protein sequence motifs that are characteristic of the Cas1 sequences of the major type II clade are absent in the type II-B group (Supplementary Figure S1). Notably, II-B is the only type II subtype that shares *cas4* genes with several type I subtypes. We further addressed the origin of the *cas4* and *cas2* genes associated with the type II-B systems.

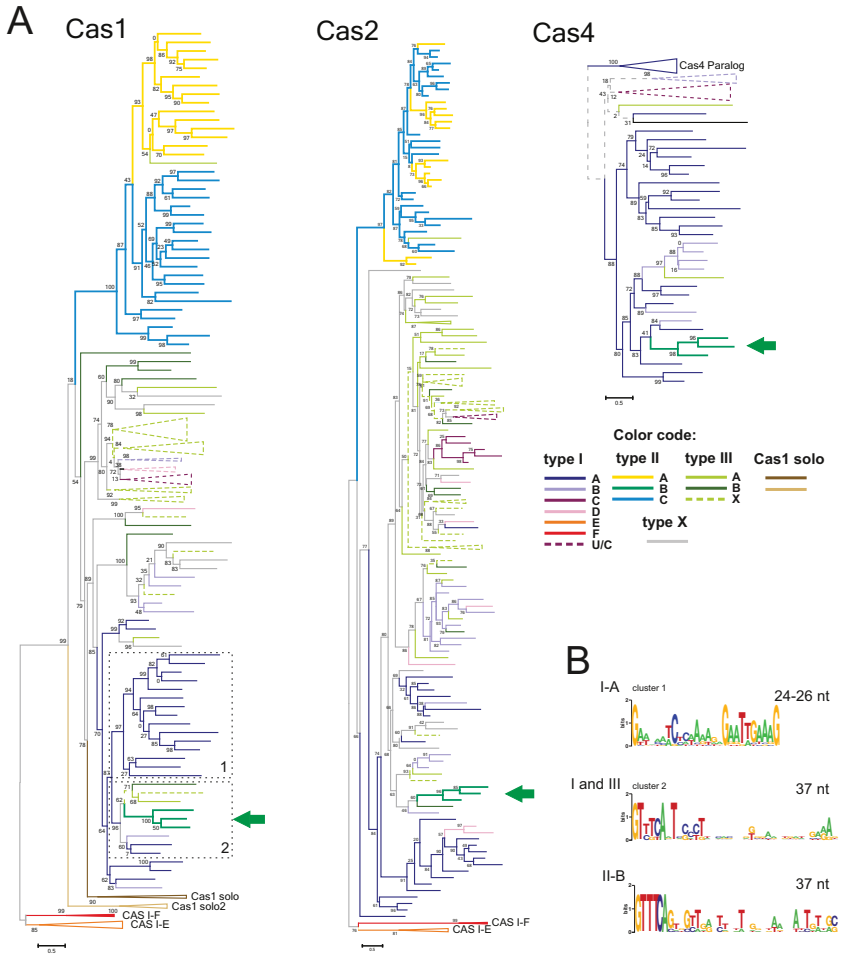

Figure 3. Origin of type II-B CRISPR-Cas system. (**A**) The PSI-BLAST program was used to retrieve Cas1 protein sequences from 2262 complete genomes in the Refseq database. The BLASTCLUST program (length coverage cutoff 0.8; score density threshold 1.0) was used to select 205 representative sequences. The multiple alignment was built using the MUSCLE program (see Supplementary Materials and Methods for details). The FastTree program ([Jones-Taylor-Thornton (JTT) evolutionary model, discrete gamma model for site rates with 20 rate categories; see Supplementary Materials and Methods for details] was used for the tree reconstruction. The Cas2 and Cas4 phylogenetic trees were reconstructed using the FastTree program as indicated for the Cas1 tree above. The sequences of these families were chosen from the same genomic neighborhoods as the selected Cas1 representatives (a few incomplete sequences from both protein families were either omitted or replaced by closely related sequences from other species). Type II-B branches are indicated by the green arrow. The branches are colored according to the assignment of *cas1* genes to CRISPR-Cas subtypes based on the analysis of 10 upstream and 10 downstream genes. X denotes systems of unknown type or those that are predicted to be derivatives of the respective system (when colored). The trees are shown only schematically, the complete trees are available in Supplementary Figure S2. (**B**) Logoplots of CRISPR repeats for the genomes that belong to several branches that are neighbors of the type II-B branch on the Cas1 phylogenetic tree. Clusters 1 and 2 are indicated by dashed lines. The type II-B (cluster 2) logoplot is shown separately. See details in Supplementary Figure S3.

To this end, we collected all *cas2* and *cas4* genes co-localized with the representative *cas1* genes and constructed phylogenetic trees for the corresponding proteins. Although the tree of Cas2 is not as reliable as the Cas1 tree due to the low sequence conservation and small size of the Cas2 proteins, the overall topology of this tree is largely consistent with that of Cas1 and also shows clustering of type II-B with type I-A (along with several type I-B and type III sequences; Figure 3A). A similar observation was made as a result of the phylogenetic analysis of Cas4 (Figure 3A). Thus, all three genes (*cas1*, *cas2* and *cas4*) that type II-B CRISPR-Cas systems share with other CRISPR-Cas types clearly originate from type I, most likely subtype I-A.

We further examined the characteristics of CRISPR repeats in the organisms that possess type I and/or type III CRISPR-Cas systems but cluster with type II-B in the Cas1 tree (Supplementary Figure S3). The repeats of type II-B more closely resemble the repeats from these organisms than those of the other type II systems including the characteristic length of the repeats, 37 nt, in contrast to the repeats of other type II systems most of which consist of 36 nucleotides and encompass distinct signature motifs (Figure 3B, Figure 4 and Supplementary Figure S3; see also a more detailed discussion below). These observations are compatible with the results of independent clustering of the repeats that was based on repeat length and sequence similarity analysis (73). Thus, *cas1*, *cas2* and *cas4* as well as the CRISPR repeats of type II-B CRISPR-Cas systems appear to originate from type I systems, probably via recombination. Given the extensive diversity of the type I CRISPR-

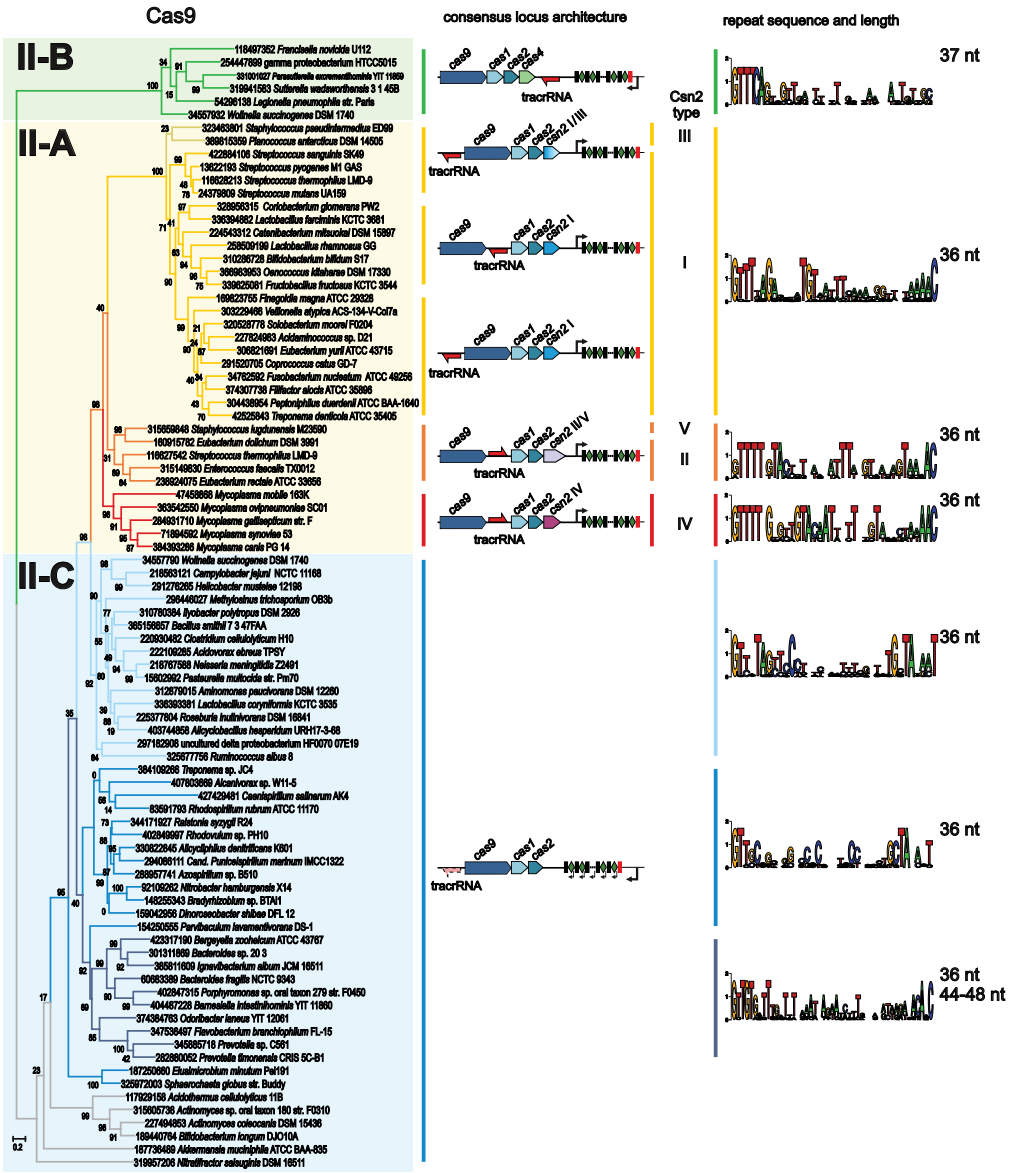

Figure 4. Cas9 phylogeny as a basis for type II system classification. The multiple alignment for the representative set of Cas9 sequences was constructed using the MUSCLE program followed by manual adjustment based on the results of pairwise alignments by PSI-BLAST, HHPRED and secondary structure predictions (see Supplementary Materials and Methods for details).

Cas systems, which is in a sharp contrast to the uniformity and narrow spread of the type II-B systems, the opposite direction of evolution can be effectively ruled out.

To investigate the evolution of type II CRISPR-Cas systems in greater detail, we performed phylogenetic analysis of Cas9 as described earlier (31). A representative set of Cas9 sequences from the genomes that encompass type II systems was constructed (see Supplementary Table S1 and Supplementary Materials and Methods for details). This sequence set was used to construct a phylogenetic tree from the conserved sequence blocks that are present in all Cas9 sequences (Figure 4 and Supplementary Figure S4). For comparison, we also reconstructed a phylogenetic tree for the Cas1 proteins from the same operons as the selected Cas9 (Supplementary Figures S5 and S6) and analyzed their genomic neighborhoods (Figure 4). Both Cas9 and Cas1 trees reveal monophyly of the II-A and II-B subtypes (Figure 4). All the genomes that encode subtype II-A possess either a *csn2* gene or previously uncharacterized genes that, as shown below, encode derived members of the Csn2 family (22). In both trees, the II-A branch is embedded within type II-C in which the *cas* operon contains only the *cas1*, *cas2* and *cas9* genes. Thus, type II-A appears to be a derivative of type II-C, with the *csn2* gene acquired by the type II-A ancestor rather than lost during the evolution of type II-C. The conservation of *csn2* in all members of the monophyletic type II-A group is most likely linked to its essential function in adaptation. The absence of the *csn2* gene in type II-C implies that in these systems adaptation occurs via a

distinct molecular mechanism that might involve additional bacterial factors.

Comparative genomics as well as experimental studies indicate that CRISPR-Cas loci are prone to HGT (5,7). Apart from the hybrid origin of the type II-B systems (see above), the present analysis reveals many likely cases of HGT. Generally, neither the Cas9 nor the Cas1 phylogenetic trees agree with the phylogeny of the respective organisms (Figure 4). Even at relatively short evolutionary distances, ample evidence of HGT is apparent (31). Furthermore, when the Cas9 protein sequences were clustered by sequence similarity using BLASTCLUST, many of the resulting clusters included bacteria of different classes (e.g. Clostridia and Bacilli in cluster 9; Supplementary Table S1). However, no clustering of sequences from different phyla was observed (e.g. no clusters that include both Firmicutes and Proteobacteria). Thus, barriers appear to exist that have prevented frequent HGT between distantly related bacteria.

Csn2, a common but fast evolving component of type II-A CRISPR-Cas systems

The next step in our exploration of the type II-A subgroup (Figure 4) involved an in-depth analysis of several genes in the respective operons that could not be confidently identified as Csn2 homologs using either PSI-BLAST or search of the Conserved Domain Database (74) which includes sequence profiles for all major domains of Cas proteins (5). These uncharacterized genes co-localize with *cas9* genes that belong to three deep branches within the type II-A clade, namely *Mycoplasmas*, *Planococcus antarcticus*, *Staphylococcus pseudintermedius* and *Staphylococcus lugdunensis* (Figure 4) (22). Using PSI-BLAST, homologs of these uncharacterized proteins in the Non-redundant protein sequence database were identified and a representative set for each subfamily was constructed (see Supplementary Table S1 and Supplementary Figure S7 for details). To search for remote sequence similarity, we used the HHPRED method with query sequences representing each of the three groups. In all three cases, HHPRED identified significant similarity of these proteins to known Csn2 sequences, suggesting that all three groups of uncharacterized proteins are diverged members of the Csn2 family (Supplementary Figure S7). This finding is in agreement with the positions of the respective Cas9 and Cas1 proteins in the phylogenetic trees and shows that Csn2 remains a perfect signature of type II-A. Superposition of the Csn2 protein architectures onto the phylogenetic trees of Cas9 and Cas1 implies that the long form of Csn2 is ancestral whereas the short form is derived (Figure 4).

We used the sequence similarity identified with HHPRED to adjust the multiple alignment of all five Csn2 subfamilies (Figure 5 and Supplementary Figure S7). In agreement with the structure comparison results, HHPRED identified the similarity between the Csn2 sequences and various ATP-Binding Cassette (ABC) ATPases (Figure 5A and Supplementary Figure S7). The similarity with ATPases is limited to the fold core including the region spanning the Walker A and B motifs (75). However, the characteristic amino acids involved in ATP and Mg^{2+} binding are replaced and thus Csn2 proteins are not predicted to bind ATP or Mg^{2+}, consistent with the experimental results (15,16,19). We also found that previously identified lysine residues involved in DNA binding are not highly conserved even among the short forms of Csn2 (Supplementary Figure S7). The C-terminal regions of the three Csn2 families are extremely diverse. The short variants lack the C-terminal extension whereas in the other subfamilies these regions are not alignable. However, based on the secondary structure prediction, it can be suggested that at least three families are structurally similar whereas the fourth is clearly distinct and might have an additional subdomain at the C-terminus (Figure 5B).

Origin of the enzymatic domains of Cas9 from common transposon-related proteins

The detailed analysis of Cas9 domains has been reported previously (8,12). Three regions of homology with the RuvC-like nucleases (RNase H fold) were identified; the predicted RuvC-like domain of Cas9 is interrupted by inserts of variable lengths (Figure 1 and Supplementary Figure S4). An uninterrupted HNH (McrA-like) nuclease domain is located in the middle of the protein (12). The arginine-rich region recently proposed to be involved in RNA binding (39) is present in all Cas9 proteins immediately downstream of the first region of RuvC homology. The large N-terminal insert is predominantly alpha-helical and shows the greatest variability among subtypes and even within subtypes of type II CRISPR-Cas systems (Supplementary Figure S4).

After the original version of this Survey and Summary was submitted for publication, two independent studies have been published reporting crystal structures of Cas9 proteins (76,77). The first of these studies describes the structures of the type II-A Cas9 from *S. pyogenes* (PDB: 4CMP) and the type II-C Cas9 from *Actinomyces naeslundii* (PDB: 4OGE) (77). The second study describes the structure of the *S. pyogenes* Cas9 co-crystalized with the guide RNA and the target DNA, and provides detailed information on the amino acid residues involved in the interaction with both RNA and DNA and the catalysis of the DNA cleavage (76). Both reports (76,77) describe two-lobed structures, with the target DNA and guide RNA positioned in the interface between the two lobes. In agreement with the secondary structure prediction, the N-terminal lobe is mostly alpha helical and could be divided into two subdomains whereas the second lobe encompasses two beta-stranded subdomains. Two specific loops in both lobes contribute to the recognition of the PAM. Several portions of the Cas9 molecule are intrinsically disordered, at least under certain conditions; in particular, this is the case of the HNH domain prior to its interaction with the dual-RNA and DNA. Outside the RuvC and HNH domains, the Cas9 structure shows no structural similarity to other proteins.

The sequence downstream of the last RuvC homology region belongs to the nuclease lobe of Cas9 and consists of two subdomains that adopt distinct alpha/beta secondary structures, and a loop that is involved in PAM recognition. This region of Cas9 showed no significant sequence or structural similarity to any known proteins (76,77) but contains a small structural motif, the Greek key, that is found

Figure 5. Multiple alignment of Csn2 subfamilies and comparison of their specific structural elements. (**A**) The multiple sequence alignment was constructed using the MUSCLE program for each Csn2 subfamily, separately. The alignments were then superimposed on the basis of conserved regions identified by HHPRED with some manual adjustment based on secondary structure predictions (see Supplementary Materials and Methods for details). The alignment with several ATPase sequences is based on Vector Alignment Search Tool (VAST) structural alignments with the structure of Csn2 of *S. thermophilus* (3ZTH) (17) used as a query (see Supplementary Materials and Methods for details). . The sequences are denoted by their GI numbers and species names. Secondary structure predictions and the secondary structure elements mapped to the respective crystal structures of the Csn2 long and short subfamilies are shown above the alignment for each Csn2 family. The positions of the first and last residues of the aligned region in the corresponding protein are indicated for each sequence. The numbers within the alignment represent poorly conserved inserts that are not shown. Secondary structure prediction is shown as follows: H indicates α-helix and E indicates extended conformation (β-strand). The positions strongly conserved in three families with a larger number of representatives are shown by reverse shading. The coloring is based on the 70% consensus built for a larger alignment (Supplementary Figure S7). Specific 90% consensus is also shown underneath the alignment for each family: 'h' indicates hydrophobic residues (WFYMLIVA), 'c' indicates charged residues (EDKRH) and 's' indicates small residues (AGS). (**B**) Schematic representation of structures (actual and predicted) of five distinct Csn2 subfamilies. Cylindrical shape represents α-helix and arrow β-strand.

in numerous OB-fold domains in diverse RNA and DNA-binding proteins (77). Here we show several additional motifs that are conserved in different Cas9 subfamilies and might be of interest for future mutagenesis studies, especially in the context of recent structural data (Figure 2 and Supplementary Figure S4).

Previously, several families of distant Cas9 homologs that are encoded outside of CRISPR-Cas loci have been identified (12). These proteins contain the RuvC-like and arginine-rich domains, and in some cases, also contain the HNH domain (Figure 2). We updated the set of these uncharacterized Cas9 homologs that are encoded in the current collection of sequenced genomes. *Ktedonobacter racemifer* remains the species with the largest number of stand-alone Cas9 homologs (49). Furthermore, 15 more genomes encode 10 or more paralogs suggesting that these genes belong to uncharacterized mobile elements. To address this possibility, we performed additional searches aimed at the identification of distant homologs of Cas9. Indeed, HHPRED searches detected significant similarity between the Cyan7822_6324-like subfamily of Cas9 homologs (GI: 297585104 from *Bacillus selenitireducens* was used as the query) and OrfB protein (also known as TnpB) from a variety of transposons of the IS605 family (Supplementary Figure S8). The similarity covers the N-terminal RuvC-like motif and the arginine-rich motif. The reverse search started from a TnpB protein (GI: 386713960 IS1341-type transposase from *Halobacillus halophilus*) identified a highly significant similarity with the RuvC nuclease and many RuvC homologs. The TnpB superfamily has recently been analyzed in detail, and many homologs of these proteins, dubbed Fanzors, have also been detected in diverse eukaryotic mobile elements (78). However, the similarity of the conserved motifs identified in the TnpB–Fanzor family to those of RuvC-like nucleases was missed in these analyses. The TnpB–Fanzor proteins do not contain transposase domains, and it has been shown that OrfB is not required for the transposition of IS605 transposons (79). The *TnpB* genes are often associated with various transposases but there are also many transposons that encompass the *TnpB* gene only (78). Accordingly, it has been hypothesized that such TnpB-only transposons employ a transposase *in trans* (78). It seems likely that both families of distant homologs of Cas9 retain this ability and are transposable with the help of transposases of other mobile elements. Such mobility would explain the extensive proliferation of these genes in many genomes. In accordance with this hypothesis, we identified several cases when these proteins are associated with group II introns [e.g. GI: 428315656 *Oscillatoria* PCC 7112, which transposes via a reverse transcription mechanism (80)]. The numerous Cas9 and TnpB homologs that contain a RuvC-like nuclease domain and an arginine-rich region can be predicted to possess DNase activity, the role of which in the life cycle of the respective transposons remains to be determined.

Evolution of type II associated CRISPR repeats and tracrRNA

Although CRISPR repeats and tracrRNA are much less amenable to phylogenetic analysis than *cas* genes because of their small size and low conservation, comparison of these sequences might provide insights into the evolution of these components within the subtypes and reveal potential exchange of repeats between CRISPR-Cas loci. Recently we reported a detailed analysis of tracrRNA anti-repeats and their base-pairing with cognate repeat sequences within the three type II subtypes and delineated several subgroups based on the similarity of repeat lengths and localization of the *tracrRNA* genes (22). Here, we identified full-length tracrRNA sequences for 63 of the 88 representative type II loci and validated the CRISPR repeat sequences according to their transcription direction (Supplementary Table S4). The updated information about the consensus locus architecture is summarized in Figure 4. We also provide a thermal map representation of the matrix of similarity within the CRISPR repeat and tracrRNA sequences (Supplementary Figure S9). The pairwise comparison of repeat sequences results in a grouping of repeats that is consistent with the Cas9 clustering. The repeat sequences within the groups share similar characteristics, especially at the 5′ and 3′ ends (Figure 4). Generally, CRISPR repeats of type II system are weakly palindromic (i.e. presumably unfolded), and typically 36 nt in length. However, as already pointed out above, type II-B repeats are 37 nt long with a distinct motif and with only the 5′-terminal part of the repeats conserved. Some longer repeats of up to 48 nt in length were found exclusively in a subset of type II-C loci, mainly in *Bacteroidetes*. Except one analyzed locus, all the repeat sequences start with G.

Despite considerable divergence of the respective Cas9 proteins and the existence of several clearly defined subgroups, which correlate with distinct subfamilies of Csn2 proteins and with distinct position and orientation of tracrRNA, all CRISPR repeats of type II-A share a clear pattern of similarity. This is less obvious in the type II-C where a certain degree of similarity is only observed within subclusters (Figure 4 and Supplementary Figure S9).

As discussed above, type II-C is extremely diverse in terms of sequence similarity of protein components, RNA components and repeats. Most of the repeats, except the *Bacteroidetes* cluster with longer repeat sequences, show conservation of the last nucleotide (T, rather than C like in type II-A; Figure 4). The repeat–spacer array is in most cases transcribed from the strand opposite to the *cas* operon and the first, not the last (as in many type I and most type II-A systems), repeat is degenerated (Supplementary Figure S10). *N. meningitidis* and *C. jejuni* contain functional promoters within the repeats (38), specifically, inside the AT-rich, conserved 3′ part of the repeats of at least two groups of type II-C. However, the lack of perfect conservation suggests that these promoters either are absent in other genomes encoding the type II-C system or are very weak. The latter possibility is consistent with recent experimental observations (38).

The tracrRNA is required for both RNase III-dependent pre-crRNA processing and interference with DNA and has to be complementary to the cognate repeat (23). TracrRNAs were identified in the majority of representative type II CRISPR-Cas systems in various locations within CRISPR loci and in different orientations with respect to the *cas* operon and the repeat–spacer array, although usu-

ally closely related genomes share similar characteristics (22). The length of the predicted tracrRNAs is highly variable (#72–171 nt) and cannot be confidently aligned even for some closely related loci (Supplementary Table S4). The pairwise comparison further corroborates the diversity in tracrRNA sequences showing clusters of similar tracrRNA sequences only for very closely related loci (Supplementary Figure S9). Base-pairing of tracrRNA and CRISPR repeats typically involves long sequence stretches, although only within tracrRNA anti-repeats. The diversity of tracrRNA length and nucleotide sequences upstream and downstream of the anti-repeats (22,23), together with the functionality of substantially truncated tracrRNA-derived parts of single guide RNAs that have been used for genome editing, suggest that the tracrRNA sequences outside the functional anti-repeat can be essentially random (33). However, the recent structural study of *S. pyogenes* Cas9 in complex with single guide RNA and target DNA (76) demonstrates the importance of certain structural features within tracrRNA. A comparative analysis of DNA cleavage efficiency in human cells using a series of mutations of the single guide RNA based on the original *S. pyogenes* dual-tracrRNA-CRISPR repeat suggests that Cas9 utilizes most efficiently long RNAs with structural features of naturally occurring tracrRNA, namely a three short stem-loop structure following the anti-repeat (76). Notably, the first stem-loop located immediately downstream of the repeat:anti-repeat duplex seems to be essential for the recognition of the guide RNA by Cas9 whereas deletion of the second and third stem-loop structures substantially decreases but does not abolish the DNA cleavage activity. The solved structure of the complex further shows that all these structural elements interact with Cas9 (76). Mutations that affect the stem sequence but not the structure have only a slight negative effect on the DNA cleavage efficiency. Even the heavily mutated guide RNA with almost one-third of the sequence exchanged within the duplex and all three stem loops, but with preserved secondary structure, remained functional in the assay, albeit with a 2-fold lower efficiency (76). Here, we modeled the putative secondary structures of tracrRNAs base-paired with cognate CRISPR repeats for all the predicted sequences (Supplementary Tables S4 and S5). Although Cas9 binding might affect the guide RNA structure and thus the predictions are unlikely to be fully accurate, we identified the three similar stem-loop structures with short G/C-rich stems downstream of the repeat:anti-repeat duplex in RNAs from loci closely related to *S. pyogenes*, e.g. in *S. mutans*, *S. thermophilus*, *Listeria innocua*, *Coriobacterium glomerans* or *Lactobacillus farciminis*. Similar short stem-loops were also found in other type II-A and many type II-C RNAs. Along the same lines, closely related loci of *N. meningitidis* and *P. multocida* seem to encode RNAs with similar secondary structure characteristics, namely two stem-loops downstream of the repeat:anti-repeat duplex.

Despite the diversity and low conservation of tracrRNA and crRNA repeat sequences, there are some indications of their co-evolution (22,23). The functionality of tracrRNA requires complementarity of the tracrRNA anti-repeat with the CRISPR repeat. A comparison of predicted secondary structures of tracrRNA antirepeat:crRNA repeat duplexes among closely related species shows compensatory mutations within RNAs resulting in the maintenance of both base-pairing and the secondary structure of the ds region [Supplementary Figure S11; (23)]. Our recent results further support the co-evolution of the type II crRNAs with Cas9 (31). Orthologous Cas9 proteins can utilize non-cognate tracrRNA and crRNA as guide sequences only when these RNAs originate from loci with highly similar Cas9 sequences, as exemplified for the groups of *S. pyogenes*, *S. mutans* and *S. thermophilus* (CRISPR3) and the pair of *N. meningitidis* and *P. multocida* (31). The RNAs and Cas9 proteins from more distantly related loci can still be exchangeable although with lowered cleavage efficiency (*N. meningitidis* or *P. multocida* with *C. jejuni*; see Figure 4). The same study suggests the involvement of the secondary structure of tracrRNA anti-repeat:crRNA repeat duplex in the specific recognition by Cas9 based on similar structure characteristics shared among exchangeable RNAs (31). This hypothesis is corroborated by the recent structural study that indicates the importance of the asymmetric bulge of tracrRNA within the *S. pyogenes* anti-repeat:repeat duplex (76). A mutation in the tracrRNA that removes the bulge and a mutation in the bulge sequence both abrogate the DNA cleavage activity of the Cas9 complex. Moreover, the bulge has been shown to directly interact with Cas9 (76). According to *in silico* predictions, this bulge is a conserved structural feature of RNAs from type II-A loci closely related to that of *S. pyogenes* (22) including the exchangeable RNAs of *S. pyogenes*, *S. mutans* and *S. thermophilus* (CRISPR3) (31).

Role of type II CRISPR-Cas in bacterial virulence and origin of scaRNA

The ability of CRISPR-Cas to limit horizontal transfer of mobile genetic elements can impact bacterial fitness and pathogenicity. An inverse relationship between the presence of a CRISPR-Cas locus and acquired antibiotic resistance has been described in *Enterococcus faecalis* (81). Similarly, an association between CRISPR spacer content and antibiotic susceptibility in *S. pyogenes* has been reported (82). By targeting multiple temperate phages, CRISPR-Cas can limit the acquisition of virulence genes in *S. pyogenes* (23). A comparative analysis of *S. pyogenes* genomes demonstrates an inverse correlation between the presence of CRISPR-Cas systems and the amount of CRISPR spacers with the number of integrated prophages (23,83).

Remarkably, type II CRISPR-Cas systems have been shown to affect virulence in *C. jejuni*, *F. novicida* and *L. pneumophila* (41). In the type II-B system of *F. novicida* U112, scaRNA has been recently identified as part of a complex that includes also Cas9 and tracrRNA. The three components are involved in the repression of the mRNA of a lipoprotein that contributes to the virulence of this bacterium (39). The regulation is likely to involve interaction between sequences located within tracrRNA and the mRNA of the lipoprotein FTN_1103, a process that would trigger mRNA degradation. We analyzed the multiple alignment of the CRISPR loci from the genomes of two *F. novicida* strains, U112 and 3523. The sequences from the two strains contain a variable region upstream of the CRISPR array, around the scaRNA sequence. The vari-

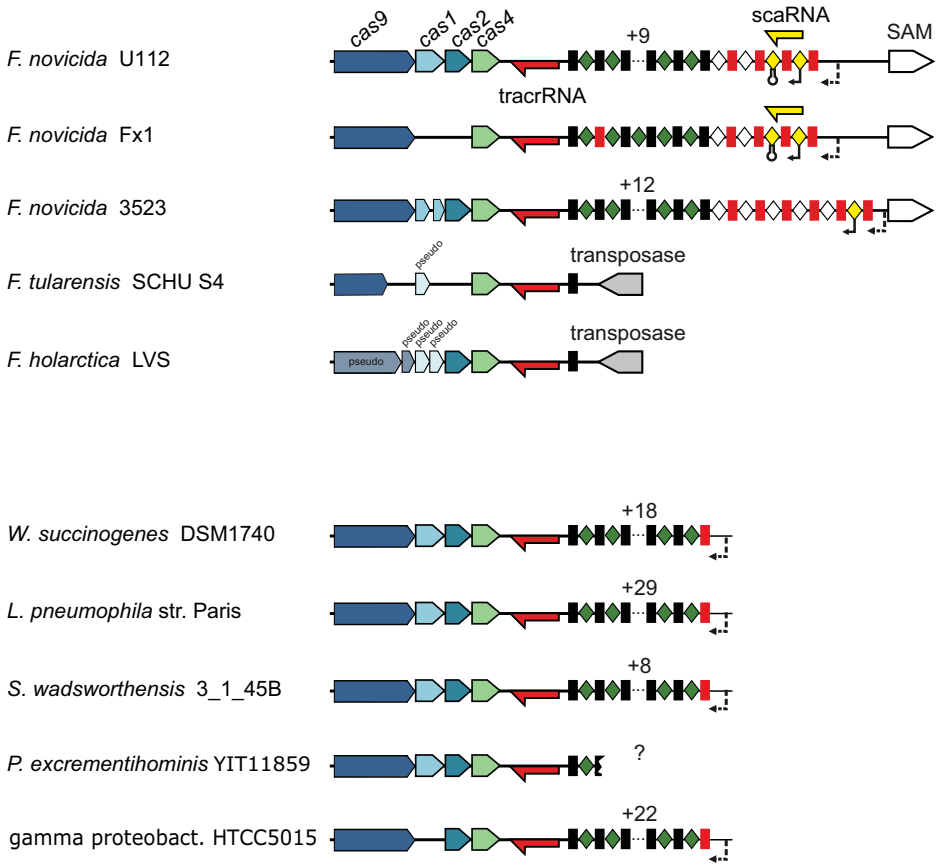

Figure 6. A schematic representation of the scaRNA-tracrRNA locus in *Francisella* strains. The type II-B CRISPR-Cas locus architecture of representative species (see Figure 4) and diverse *Francisella* species is shown. Red and yellow arrows: tracrRNA and scaRNA with indicated confirmed (22) or predicted transcription direction, accordingly; black rectangles and green diamonds: repeat–spacer arrays; red rectangles: degenerated repeats; white diamonds: putative spacers of degenerated arrays. Degenerated array spacers with the scaRNA promoter and transcriptional terminator are shown in yellow. Putative promoters of repeat–spacer arrays are shown with dotted arrows. The scaRNA-encoding spacer–repeat–spacer unit was found only in two of the analyzed strains and is incomplete in *F. novicida* 3523, lacking transcriptional terminator-encoding spacer. Note also the degenerate repeats that are commonly found at the 5′-end of the repeat–spacer array. See Supplementary Figure S12.

able region contains highly degenerated but regularly interspaced repeat sequences, most likely a remnant of an old, not currently active CRISPR array, which is longer in *F. novicida* 3523 than in the U112 strain (Figure 6 and Supplementary Figure S12). The scaRNA sequence covers one degenerated repeat completely and another one partially. The scaRNA appears to be transcribed from a promoter located in the putative spacer sequence to a transcriptional terminator located within the next putative spacer sequence. In *F. novicida* 3523, only the 5′ part of the scaRNA sequence containing the degenerated repeat that can base-pair with tracrRNA is conserved, whereas the 3′ part of the sequence is absent or not conserved and the transcriptional terminator cannot be confidently predicted. Thus, scaRNA might not be expressed as a single RNA species in *F. novicida* 3532. We hypothesize that scaRNA is not a novel small RNA transcript but rather a fragment of a highly degenerated, old CRISPR array that might be present only in genomes of closely related isolates of *Francisella* (e.g. *F. novicida* Fx1). This fragment would still be transcribed (albeit with substantially reduced efficiency) but not processed into mature crRNAs. The degenerated sequence cannot serve as a template for new spacer integration either. Such an inactivated transcript would be prone to mutations and recombination and accordingly would have the potential for the acquisition of a new function. In the case of *F. novicida* U112, this region might have acquired its own promoter and terminator and started to be transcribed as a single RNA, still retaining the ability to interact with both tracrRNA and Cas9. The promoter and terminator sequences could have been acquired as CRISPR spacer units while the array was still functional.

CONCLUSIONS

Since the demonstration of the activity of type II CRISPR-Cas system in bacterial adaptive defense against phages and plasmids (13,14), and the discovery of the essential role of tracrRNA in crRNA maturation and DNA targeting by Cas9 (23,33), type II CRISPR-Cas systems have been actively studied. Thanks to the minimalist architecture of the interfering complex that consists of the single, multidomain endonuclease Cas9 guided by dual-tracrRNA:crRNA, type II is the obvious top choice of a CRISPR-Cas system for application in genome engineering. Despite the major differences in the mechanisms of crRNA maturation and ap-

paratus required for target DNA cleavage, type I, type II and type III systems share important mechanistic details. Such shared features between different types of CRISPR-Cas systems include the role of PAM as both a prerequisite for target DNA binding and a motif that is required for spacer acquisition and self–non-self discrimination (types I and II) as well as R-loop formation during DNA targeting (types I and II). The duplex structure of the natural dual-tracrRNA:crRNA and the minimalist single engineered guide RNA (sufficient for Cas9 activation) can both be considered mimics of the 3′ hairpin of mature crRNAs that are transcribed from palindromic repeats in CRISPR-Cas types I and III (Figure 1).

We present evidence that reinforces the previously published hypothesis on the origin of type II system (12). First, it is shown that type II-B systems evolved by recombination between Cas9 and a type I CRISPR-Cas locus. Second, we describe the homology between Cas9 and transposon-encoded-predicted nucleases. Conceivably, transposons were the original source of the nuclease moieties of Cas9, and the mobility of these elements would have greatly facilitated recombination. Although there are some mechanistic similarities of the interference processes between type I and type II CRISPR-Cas systems, the N- and C-terminal inserts in the recently determined structures of Cas9 do not show any structural similarity with subunits of Cascade complexes or any other proteins. Therefore, the currently available Cas9 structures do not directly contribute to deciphering the origin of Cas9 beyond the two nuclease domains. In that regard, the structure of type II-B Cas9, which is the largest protein in the family and, as shown here, belongs to the system in which other components most likely originated from type I CRISPR-Cas, could be informative. Of further interest will be structures of distant homologs of Cas9 which are encoded outside CRISPR-Cas loci.

The evolutionary and functional plasticity of type II CRISPR-Cas systems is further demonstrated by the apparent acquisition of the dual-RNA-guided RNA interfering enzyme function by the Cas9 protein of *Francisella* (39). This transition would be accompanied by the recruitment of the transcript of a degraded array of CRISPR repeat–spacer to tracrRNA as a novel guide dual-RNA. The contribution of this derived type II system to the pathogenicity of *Francisella* and the involvement of type II CRISPR-Cas in the virulence of *C. jejuni* (40), *N. meningitidis* (39) and *L. pneumophila* (41), taken together with the significant enrichment of type II systems among pathogens and commensals, suggest that rewiring of this system for functions distinct from antivirus defence is a general evolutionary trend. The potential involvement in the regulation of bacterial pathogenicity is another important incentive for detailed study of type II CRISPR-Cas systems.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGMENTS

K.C. was a fellow of the Austrian Doctoral Program in RNA Biology.

FUNDING

US Department of Health and Human Services (tothe National Library of Medicine) [K.S.M., E.V.K.]; Swedish Research Council [K2010-57X-21436-01-3, K2013-57X-21436-04-3, 621-2011-5752-LiMS to E.C.]; Kempe Foundation [#SMK-1136.1]; Umeå University [Dnr: 223-2728-10, Dnr: 223-2836-10, Dnr: 223-2989-10 to E.C.]; Laboratory for Molecular Infection Medicine Sweden, the Umeå Centre for Microbial Research, the Helmholtz Association and the Alexander von Humboldt Foundation, Federal Ministry of Education and Research [E.C.]. Funding for open access charge: US Department of Health and Human Services (to the National Library of Medicine); Swedish Research Council [K2010-57X-21436-01-3, K2013-57X-21436-04-3, 621-2011-5752-LiMS]; Kempe Foundation; Umeå University [Dnr: 223-2728-10, Dnr: 223-2836-10, Dnr: 223-2989-10].

Conflict of interest statement. None declared.

REFERENCES

1. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
2. Barrangou,R. and Horvath,P. (2012) CRISPR: new horizons in phage resistance and strain identification. *Annu. Rev. Food Sci. Technol.*, **3**, 143–162.
3. Wiedenheft,B., Sternberg,S.H. and Doudna,J.A. (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, **482**, 331–338.
4. van der Oost,J., Jore,M.M., Westra,E.R., Lundgren,M. and Brouns,S.J. (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends Biochem. Sci.*, **34**, 401–407.
5. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
6. Jansen,R., Embden,J.D., Gaastra,W. and Schouls,L.M. (2002) Identification of genes that are associated with DNA repeats in prokaryotes. *Mol. Microbiol.*, **43**, 1565–1575.
7. Haft,D.H., Selengut,J., Mongodin,E.F. and Nelson,K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.
8. Makarova,K.S., Grishin,N.V., Shabalina,S.A., Wolf,Y.I. and Koonin,E.V. (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.
9. Makarova,K.S., Aravind,L., Grishin,N.V., Rogozin,I.B. and Koonin,E.V. (2002) A DNA repair system specific for thermophilic archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Res.*, **30**, 482–496.
10. Reeks,J., Naismith,J.H. and White,M.F. (2013) CRISPR interference: a structural perspective. *Biochem. J.*, **453**, 155–166.
11. Koonin,E.V. and Makarova,K.S. (2013) CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes. *RNA Biol.*, **10**, 679–686.
12. Makarova,K.S., Aravind,L., Wolf,Y.I. and Koonin,E.V. (2011) Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. *Biol. Direct.*, **6**, 38.
13. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
14. Garneau,J.E., Dupuis,M.E., Villion,M., Romero,D.A., Barrangou,R., Boyaval,P., Fremaux,C., Horvath,P., Magadan,A.H. and Moineau,S. (2010) The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, **468**, 67–71.

15. Nam,K.H., Kurinov,I. and Ke,A. (2011) Crystal structure of clustered regularly interspaced short palindromic repeats (CRISPR)-associated Csn2 protein revealed Ca2+-dependent double-stranded DNA-binding activity. *J. Biol. Chem.*, **286**, 30759–30768.
16. Koo,Y., Jung,D.K. and Bae,E. (2012) Crystal structure of Streptococcus pyogenes Csn2 reveals calcium-dependent conformational changes in its tertiary and quaternary structure. *PLoS ONE*, **7**, e33401.
17. Lee,K.H., Lee,S.G., Eun Lee,K., Jeon,H., Robinson,H. and Oh,B.H. (2012) Identification, structural, and biochemical characterization of a group of large Csn2 proteins involved in CRISPR-mediated bacterial immunity. *Proteins*, **80**, 2573–2582.
18. Arslan,Z., Wurm,R., Brener,O., Ellinger,P., Nagel-Steger,L., Oesterhelt,F., Schmitt,L., Willbold,D., Wagner,R., Gohlke,H. *et al.* (2013) Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. *Nucleic Acids Res.*, **41**, 6347–6359.
19. Ellinger,P., Arslan,Z., Wurm,R., Tschapek,B., MacKenzie,C., Pfeffer,K., Panjikar,S., Wagner,R., Schmitt,L., Gohlke,H. *et al.* (2012) The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. *J. Struct. Biol.*, **178**, 350–362.
20. Zhang,J., Kasciukovic,T. and White,M.F. (2012) The CRISPR associated protein Cas4 Is a 5′ to 3′ DNA exonuclease with an iron-sulfur cluster. *PLoS ONE*, **7**, e47232.
21. Makarova,K.S., Anantharaman,V., Aravind,L. and Koonin,E.V. (2012) Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. *Biol. Direct.*, **7**, 40.
22. Chylinski,K., Le Rhun,A. and Charpentier,E. (2013) The tracrRNA and Cas9 families of type II CRISPR-Cas immunity systems. *RNA Biol.*, **10**, 726–737.
23. Deltcheva,E., Chylinski,K., Sharma,C.M., Gonzales,K., Chao,Y., Pirzada,Z.A., Eckert,M.R., Vogel,J. and Charpentier,E. (2011) CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature*, **471**, 602–607.
24. Yosef,I., Goren,M.G. and Qimron,U. (2012) Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. *Nucleic Acids Res.*, **40**, 5569–5576.
25. Datsenko,K.A., Pougach,K., Tikhonov,A., Wanner,B.L., Severinov,K. and Semenova,E. (2012) Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. *Nat. Commun.*, **3**, 945.
26. Swarts,D.C., Mosterd,C., van Passel,M.W. and Brouns,S.J. (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS ONE*, **7**, e35888.
27. Horvath,P., Romero,D.A., Coute-Monvoisin,A.C., Richards,M., Deveau,H., Moineau,S., Boyaval,P., Fremaux,C. and Barrangou,R. (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.*, **190**, 1401–1412.
28. Bhaya,D., Davison,M. and Barrangou,R. (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. *Annu. Rev. Genet.*, **45**, 273–297.
29. Mojica,F.J., Diez-Villasenor,C., Garcia-Martinez,J. and Almendros,C. (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.
30. Deveau,H., Barrangou,R., Garneau,J.E., Labonte,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2008) Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. *J. Bacteriol.*, **190**, 1390–1400.
31. Fonfara,I., Le Rhun,A., Chylinski,K., Makarova,K.S., Lecrivain,A.L., Bzdrenga,J., Koonin,E.V. and Charpentier,E. (2013) Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic Acids Res.*, **42**, 2577–2590.
32. Sapranauskas,R., Gasiunas,G., Fremaux,C., Barrangou,R., Horvath,P. and Siksnys,V. (2011) The Streptococcus thermophilus CRISPR/Cas system provides immunity in Escherichia coli. *Nucleic Acids Res.*, **39**, 9275–9282.
33. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
34. Haurwitz,R.E., Jinek,M., Wiedenheft,B., Zhou,K. and Doudna,J.A. (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, **329**, 1355–1358.
35. Deng,L., Kenchappa,C.S., Peng,X., She,Q. and Garrett,R.A. (2012) Modulation of CRISPR locus transcription by the repeat-binding protein Cbp1 in Sulfolobus. *Nucleic Acids Res.*, **40**, 2470–2480.
36. Carte,J., Wang,R., Li,H., Terns,R.M. and Terns,M.P. (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes Dev.*, **22**, 3489–3496.
37. Nam,K.H., Haitjema,C., Liu,X., Ding,F., Wang,H., DeLisa,M.P. and Ke,A. (2012) Cas5d protein processes pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. *Structure*, **20**, 1574–1584.
38. Zhang,Y., Heidrich,N., Ampattu,B.J., Gunderson,C.W., Seifert,H.S., Schoen,C., Vogel,J. and Sontheimer,E.J. (2013) Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. *Mol. Cell*, **50**, 488–503.
39. Sampson,T.R., Saroj,S.D., Llewellyn,A.C., Tzeng,Y.L. and Weiss,D.S. (2013) A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature*, **497**, 254–257.
40. Louwen,R., Horst-Kreft,D., de Boer,A.G., van der Graaf,L., de Knegt,G., Hamersma,M., Heikema,A.P., Timms,A.R., Jacobs,B.C., Wagenaar,J.A. *et al.* (2013) A novel link between Campylobacter jejuni bacteriophage defence, virulence and Guillain-Barre syndrome. *Eur. J. Clin. Microbiol. Infect. Dis.*, **32**, 207–226.
41. Gunderson,F.F. and Cianciotto,N.P. (2013) The CRISPR-associated gene cas2 of Legionella pneumophila is required for intracellular infection of amoebae. *mBio*, **4**, e00074-13.
42. Mali,P., Yang,L., Esvelt,K.M., Aach,J., Guell,M., DiCarlo,J.E., Norville,J.E. and Church,G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823–826.
43. Mali,P., Aach,J., Stranges,P.B., Esvelt,K.M., Moosburner,M., Kosuri,S., Yang,L. and Church,G.M. (2013) CAS9 transcriptional activators for target specificity screening and paired nickases for cooperative genome engineering. *Nat. Biotechnol.*, **31**, 833–838.
44. Ran,F.A., Hsu,P.D., Lin,C.Y., Gootenberg,J.S., Konermann,S., Trevino,A.E., Scott,D.A., Inoue,A., Matoba,S., Zhang,Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
45. Cong,L., Ran,F.A., Cox,D., Lin,S., Barretto,R., Habib,N., Hsu,P.D., Wu,X., Jiang,W., Marraffini,L.A. *et al.* (2013) Multiplex genome engineering using CRISPR/Cas systems. *Science*, **339**, 819–823.
46. Perez-Pinera,P., Kocak,D.D., Vockley,C.M., Adler,A.F., Kabadi,A.M., Polstein,L.R., Thakore,P.I., Glass,K.A., Ousterout,D.G., Leong,K.W. *et al.* (2013) RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nat. Methods*, **10**, 973–976.
47. Gilbert,L.A., Larson,M.H., Morsut,L., Liu,Z., Brar,G.A., Torres,S.E., Stern-Ginossar,N., Brandman,O., Whitehead,E.H., Doudna,J.A. *et al.* (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell*, **154**, 442–451.
48. Bikard,D., Jiang,W., Samai,P., Hochschild,A., Zhang,F. and Marraffini,L.A. (2013) Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.*, **41**, 7429–7437.
49. Qi,L.S., Larson,M.H., Gilbert,L.A., Doudna,J.A., Weissman,J.S., Arkin,A.P. and Lim,W.A. (2013) Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell*, **152**, 1173–1183.
50. Mali,P., Esvelt,K.M. and Church,G.M. (2013) Cas9 as a versatile tool for engineering biology. *Nat. Methods*, **10**, 957–963.
51. Gasiunas,G., Barrangou,R., Horvath,P. and Siksnys,V. (2012) Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, E2579–E2586.
52. Cho,S.W., Kim,S., Kim,J.M. and Kim,J.S. (2013) Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nat. Biotechnol.*, **31**, 230–232.
53. Jinek,M., East,A., Cheng,A., Lin,S., Ma,E. and Doudna,J. (2013) RNA-programmed genome editing in human cells. *Elife*, **2**, e00471.
54. Niu,Y., Shen,B., Cui,Y., Chen,Y., Wang,J., Wang,L., Kang,Y., Zhao,X., Si,W., Li,W. *et al.* (2014) Generation of gene-modified cynomolgus monkey via Cas9/RNA-mediated gene targeting in one-cell embryos. *Cell*, **156**, 836–843.
55. Hai,T., Teng,F., Guo,R., Li,W. and Zhou,Q. (2014) One-step generation of knockout pigs by zygote injection of CRISPR/Cas system. *Cell Res.*, **24**, 372–375.

56. Shen,B., Zhang,J., Wu,H., Wang,J., Ma,K., Li,Z., Zhang,X., Zhang,P. and Huang,X. (2013) Generation of gene-modified mice via Cas9/RNA-mediated gene targeting. *Cell Res.*, **23**, 720–723.
57. Wang,H., Yang,H., Shivalila,C.S., Dawlaty,M.M., Cheng,A.W., Zhang,F. and Jaenisch,R. (2013) One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. *Cell*, **153**, 910–918.
58. Hwang,W.Y., Fu,Y., Reyon,D., Maeder,M.L., Kaini,P., Sander,J.D., Joung,J.K., Peterson,R.T. and Yeh,J.R. (2013) Heritable and precise zebrafish genome editing using a CRISPR-Cas system. *PLoS ONE*, **8**, e68708.
59. Gratz,S.J., Cummings,A.M., Nguyen,J.N., Hamm,D.C., Donohue,L.K., Harrison,M.M., Wildonger,J. and O'Connor-Giles,K.M. (2013) Genome engineering of Drosophila with the CRISPR RNA-guided Cas9 nuclease. *Genetics*, **194**, 1029–1035.
60. DiCarlo,J.E., Norville,J.E., Mali,P., Rios,X., Aach,J. and Church,G.M. (2013) Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Res.*, **41**, 4336–4343.
61. Shan,Q., Wang,Y., Li,J., Zhang,Y., Chen,K., Liang,Z., Zhang,K., Liu,J., Xi,J.J., Qiu,J.L. *et al.* (2013) Targeted genome modification of crop plants using a CRISPR-Cas system. *Nat. Biotechnol.*, **31**, 686–688.
62. Xie,K. and Yang,Y. (2013) RNA-guided genome editing in plants using a CRISPR-Cas system. *Mol. Plant*, **6**, 1975–1983.
63. Waaijers,S., Portegijs,V., Kerver,J., Lemmens,B.B., Tijsterman,M., van den Heuvel,S. and Boxem,M. (2013) CRISPR/Cas9-targeted mutagenesis in Caenorhabditis elegans. *Genetics*, **195**, 1187–1191.
64. Jiang,W., Bikard,D., Cox,D., Zhang,F. and Marraffini,L.A. (2013) RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.*, **31**, 233–239.
65. Wei,C., Liu,J., Yu,Z., Zhang,B., Gao,G. and Jiao,R. (2013) TALEN or Cas9 - rapid, efficient and specific choices for genome modifications. *J. Genet. Genomics*, **40**, 281–289.
66. Pennisi,E. (2013) The CRISPR craze. *Science*, **341**, 833–836.
67. Wolf,Y.I., Makarova,K.S., Yutin,N. and Koonin,E.V. (2012) Updated clusters of orthologous genes for archaea: a complex ancestor of the archaea and the byways of horizontal gene transfer. *Biol. Direct.*, **7**, 46.
68. Rho,M., Wu,Y.W., Tang,H., Doak,T.G. and Ye,Y. (2012) Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.*, **8**, e1002441.
69. Pride,D.T., Salzman,J. and Relman,D.A. (2012) Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses. *Environ. Microbiol.*, **14**, 2564–2576.
70. Berg Miller,M.E., Yeoman,C.J., Chia,N., Tringe,S.G., Angly,F.E., Edwards,R.A., Flint,H.J., Lamed,R., Bayer,E.A. and White,B.A. (2012) Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome. *Environ. Microbiol.*, **14**, 207–227.
71. Horvath,P. and Barrangou,R. (2010) CRISPR/Cas, the immune system of bacteria and archaea. *Science*, **327**, 167–170.
72. Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2013) The basic building blocks and evolution of CRISPR-Cas systems. *Biochem. Soc. Trans.*, **41**, 1392–1400.
73. Kunin,V., Sorek,R. and Hugenholtz,P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.*, **8**, R61.
74. Marchler-Bauer,A., Anderson,J.B., Chitsaz,F., Derbyshire,M.K., DeWeese-Scott,C., Fong,J.H., Geer,L.Y., Geer,R.C., Gonzales,N.R., Gwadz,M. *et al.* (2009) CDD: specific functional annotation with the Conserved Domain Database. *Nucleic Acids Res.*, **37**, D205–D210.
75. Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, **1**, 945–951.
76. Nishimasu,H., Ran,F.A., Hsu,P.D., Konermann,S., Shehata,S.I., Dohmae,N., Ishitani,R., Zhang,F. and Nureki,O. (2014) Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, **156**, 935–949.
77. Jinek,M., Jiang,F., Taylor,D.W., Sternberg,S.H., Kaya,E., Ma,E., Anders,C., Hauer,M., Zhou,K., Lin,S. *et al.* (2014) Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, **343**, doi:10.1126/science.1247997.
78. Bao,W. and Jurka,J. (2013) Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. *Mob DNA*, **4**, 12.
79. Kersulyte,D., Velapatino,B., Dailide,G., Mukhopadhyay,A.K., Ito,Y., Cahuayme,L., Parkinson,A.J., Gilman,R.H. and Berg,D.E. (2002) Transposable element ISHp608 of Helicobacter pylori: nonrandom geographic distribution, functional organization, and insertion specificity. *J. Bacteriol.*, **184**, 992–1002.
80. Lambowitz,A.M. and Zimmerly,S. (2011) Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harb. Perspect. Biol.*, **3**, a003616.
81. Palmer,K.L. and Gilmore,M.S. (2010) Multidrug-resistant enterococci lack CRISPR-cas. *mBio*, **1**, e00227–10.
82. Zheng,P.X., Chiang-Ni,C., Wang,S.Y., Tsai,P.J., Kuo,C.F., Chuang,W.J., Lin,Y.S., Liu,C.C. and Wu,J.J. (2013) Arrangement and number of clustered regularly interspaced short palindromic repeat spacers are associated with erythromycin susceptibility in emm12, emm75 and emm92 of group A streptococcus. *Clin. Microbiol. Infect.*, doi:10.1111/1469-0691.12379.
83. Nozawa,T., Furukawa,N., Aikawa,C., Watanabe,T., Haobam,B., Kurokawa,K., Maruyama,F. and Nakagawa,I. (2011) CRISPR inhibition of prophage acquisition in Streptococcus pyogenes. *PLoS ONE*, **6**, e19543.