The CRISPR Journal
Volume 3, Number 6, 2020
© Mary Ann Liebert, Inc.
DOI: 10.1089/crispr.2020.29111.cpo

The CRISPR Journal

FIRST CUT

New Insights into CRISPR Arrays

Christine Pourcel

A new study in this issue from Eugene Koonin and his colleagues (Shmakov *et al.* page 535) investigates isolated CRISPR arrays and defines true "orphan" arrays that may be associated with yet unknown CRISPR-Cas systems

Downloaded by 116.237.143.130 from www.liebertpub.com at 12/28/20. For personal use only.

CRISPR-Cas systems have evolved into a very complex family, as reflected by the composition and functionality of the *cas* gene clusters and the diversity of CRISPR repeats. Despite a remarkably conserved structure, CRISPR arrays display heterogeneity in size, nature and length of their leader, and their capacity to expand, all characteristics that are poorly understood.

Whereas the arrays in archaea and many extremophiles can store >100 spacers, in most species, CRISPR arrays are no longer than a few dozen spacers, which may be due to deletion by recombination between repeats. The leader, a 100–200 bp sequence flanking the array on its proximal side,[1] is involved in transcription and in adaptation through binding of the Cas1–Cas2 complex onto the leader-repeat junction.[2] It coevolves with repeats, the Cas1 protein, and the protospacer adjacent motif.[3] Interestingly, in many genomes, a single cluster of *cas* genes may be associated with several arrays (up to 24 in *Nocardiopsis alba* ATCC BAA-2165 for a single Type IA *cas* gene cluster; https://crisprcas.i2bc.paris-saclay.fr/[4]), which possess the same or very similar repeat sequences and do not share any spacer. They are generally flanked by a leader with 70–90% similarity over >100 bp, and can be targets for spacer acquisition.

Multiple arrays are found more frequently in genomes bearing a Type I and/or Type III system, as in most archaea and bacteria, but are rare in Type II systems. They can be scattered or grouped, often near the *cas* gene cluster. The multiplication of CRISPR arrays with their leader is probably not an accident. An analysis of 5,796 genomes bearing a single Type I or Type III *cas* gene cluster shows that the number of spacers increases with the number of CRISPR arrays (Fig. 1A). This may suggest that in cells where the size of the CRISPR arrays seems to be limited, the existence of multiple copies that can be transcribed efficiently and are available for insertion of spacers provides a selective advantage. This would prevent the loss of isolated CRISPR arrays, as observed for different gene families.[5]

Writing in this issue of *The CRISPR Journal*, Shmakov *et al.* present the results of their exploration of 13,116 complete genome sequences in search for isolated CRISPR arrays, that is, those with no nearby *cas* genes.[6] Over the past 15 years, Eugene Koonin's group has been a leader in cataloguing Cas proteins, allowing the precise definition and classification of the CRISPR-Cas types and subtypes. In the latest work, they apply the minCED tool (https://github.com/ctSkennerton/minced) and a filtering pipeline to identify CRISPR arrays that are not adjacent to *cas* genes. In agreement with previous observations made with a more limited number of sequences, they find that 90% of arrays that have no *cas* gene in their vicinity possess a repeat shared with a *cas*-accompanied array in the same or different genomes and a leader. Interestingly, they identify 116 unique *bona fide* arrays distributed across 89 clusters, for which repeats show no similarity to known CRISPR, and are therefore ''orphans'' until the *cas* they complement is uncovered.

To explain the existence of isolated arrays with similarity to *cas*-adjacent arrays, the authors analyzed their region of insertion, and found that transposable elements were three times more frequent than in other part of the genome. Therefore, they propose a plausible scenario: the loss of the *cas* gene cluster following transfer of a complete CRISPR-Cas system. Another hypothesis is *de novo* generation by off-target insertion of spacers into a repeat-like sequence as observed experimentally[7] and dissemination by mobile genetic elements.

Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France.

Address correspondence to: Christine Pourcel, PhD, Institute for Integrative Biology of the Cell (I2BC), SSFA, Université Paris-Saclay, Bât 400, 91400 Orsay, France, Email: christine.pourcel@i2bc.paris-saclay.fr

Downloaded by 116.237.143.130 from www.liebertpub.com at 12/28/20. For personal use only.

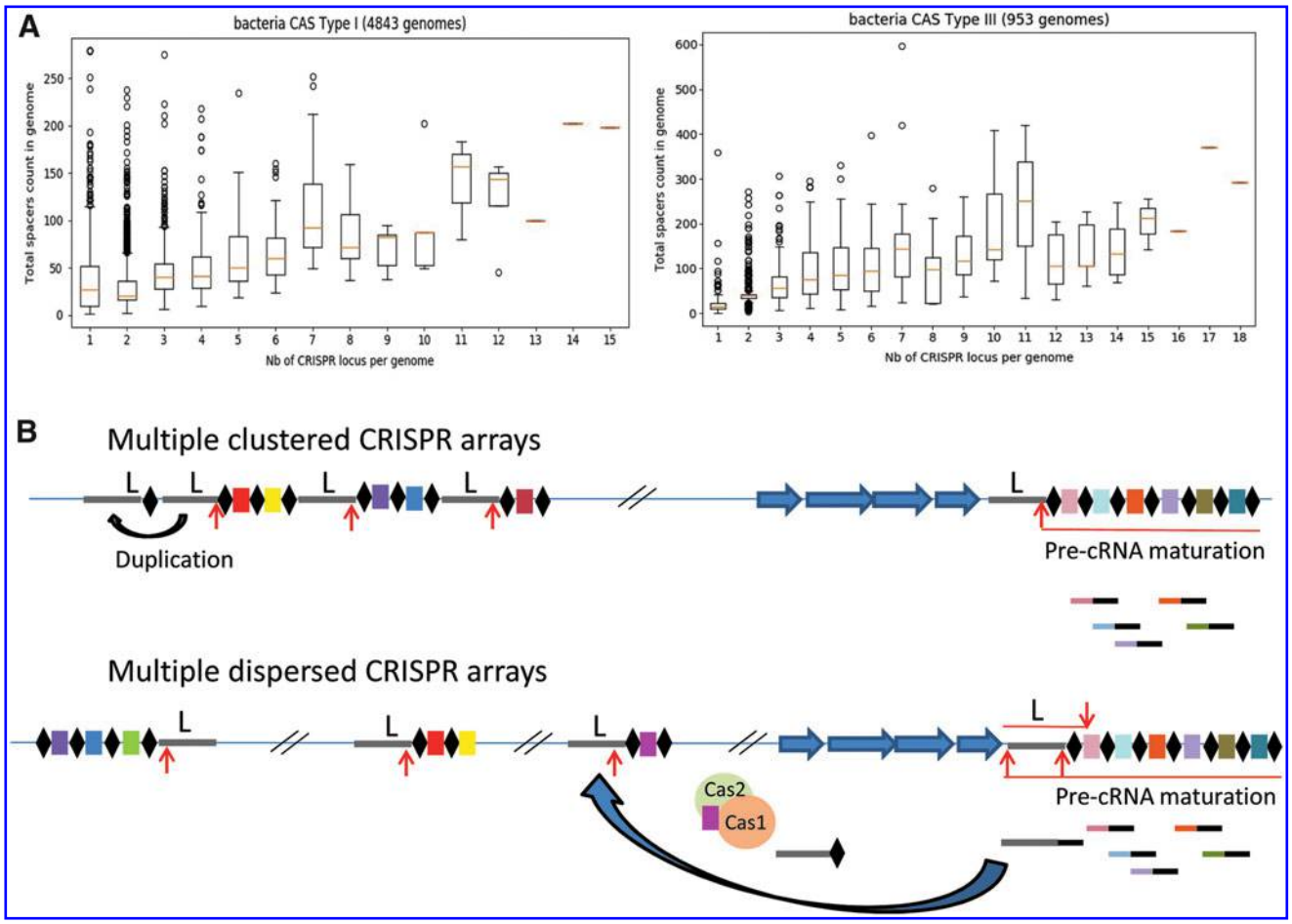


FIG. 1. Multiplication of CRISPR arrays. **(A)** Box plot representation of the correlation between the number of CRISPR arrays and the total number of spacers in bacteria possessing a single Type I or Type III CRISPR-Cas system. Half (50%) of the observations are inside the box, and the median is represented by a continuous orange line. Outliers are represented by dots outside the whiskers. **(B)** Proposed mechanisms to account for the creation of new CRISPR arrays by duplication or reverse transcription and insertion of a leader-repeat RNA. The black diamonds represent repeats. Colored boxes are spacers. The leader is shown with a gray bar marked L. The red arrows are transcription starts. The Cas1–Cas2 complex binds to the leader repeat junction and allows insertion of a new spacer and duplication of the repeat.

These mechanisms may apply in some situations but can hardly explain the majority of isolated-CRISPR arrays. The systematic deletion of *cas* clusters seems improbable, and the creation of arrays following off-target insertion of spacers is incompatible with the presence of the conserved leader sequence in a majority of isolated arrays. I propose two additional mechanisms that could lead to production of copies of the leader and the first repeat of a *cas*-associated array and allow their insertion at different positions. In cases where arrays are clustered and separated only by their leader, such as in *Fusobacterium pseudoperiodonticum* KCOM 2555 (with 11 arrays within 9 kb), it might result from tandem duplication of the original array and exclusion of preexisting spacers by recombination between repeats (Fig. 1B). Excision of some arrays and nonhomologous recombination could account for scattered distribution of new copies.

Another mechanism could involve retro-transcription of a mature RNA bearing the leader and the proximal repeat (Fig. 1C). RNA-seq analysis has shown that in addition to transcription initiated at the leader, antisense transcripts can be common due to the presence of promoters within repeats or in some spacers.[8,9] Furthermore, in some arrays such as in *Myxococcus xhantus*, the leader is transcribed from the *cas* gene cluster readthrough.[10] Ancillary proteins with reverse transcriptase activity, group II intron or RT-Cas1 fusion proteins in Type III systems, might be actors in the generation of new CRISPR arrays.[11,12]

In the past few years, the term ''orphan'' has been used somewhat indiscriminately to define any array that is not physically associated with *cas* genes, whether the repeat was shared with a *cas*-associated array or not. They were also called ''split arrays,''[13] although the presence of the leader shows that the majority do not correspond to an array that has been fragmented. Shmakov *et al.* propose a list of arrays that do not seem to be associated with known *cas* genes and will be the subject of future investigations. It is important that each genome presumably holding orphan arrays be reexamined with caution such as the cited *Ktedonobacterales* bacterium SCAWS-G2 genome which does possess Type IIID and Type ID *cas* genes and 8 isolated arrays.

Much work remains in order to understand the dynamics of the CRISPR-Cas systems and particularly the birth and multiplication of CRISPR arrays. The organization we see today is the result of millions of years of slow evolution, and we need to analyze more genomes to reconstruct these ancient events.

References

1. Alkhnbashi OS, Shah SA, Garrett RA, et al. Characterizing leader sequences of CRISPR loci. *Bioinformatics* 2016;32:i576–i585. DOI: 10.1093/bioinformatics/btw454.
2. Sasnauskas G, Siksnys V. CRISPR adaptation from a structural perspective. *Curr Opin Struct Biol* 2020;65:17–25. DOI: 10.1016/j.sbi.2020.05.015.
3. Shah SA, Garrett RA. CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res Microbiol* 2011;162:27–38. DOI: 10.1016/j.resmic.2010.09.001.
4. Pourcel C, Touchon M, Villeriot N, et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res* 2020;48:D535–D544. DOI: 10.1093/nar/gkz915.
5. Reams AB, Neidle EL. Selection for gene clustering by tandem duplication. *Annu Rev Microbiol* 2004;58:119–142. DOI: 10.1146/annurev.micro.58.030603.123806.
6. Shmakov SA, Utkina I, Wolf YI, et al. CRISPR arrays away from *cas* genes. *CRISPR J* 2020.
7. Nivala J, Shipman SL, Church GM. Spontaneous CRISPR loci generation *in vivo* by non-canonical spacer integration. *Nat Microbiol* 2018;3:310–318. DOI: 10.1038/s41564-017-0097-z.
8. Zoephel J, Randau L. RNA-Seq analyses reveal CRISPR RNA processing and regulation patterns. *Biochem Soc Trans* 2013;41:1459–1463. DOI: 10.1042/BST20130129.
9. Heidrich N, Dugar G, Vogel J, et al. Investigating CRISPR RNA biogenesis and function using RNA-seq. *Methods Mol Biol* 2015;1311:1–21. DOI: 10.1007/978-1-4939-2687-9_1.
10. Bernal-Bernal D, Abellon-Ruiz J, Iniesta AA, et al. Multifactorial control of the expression of a CRISPR-Cas system by an extracytoplasmic function sigma/anti-sigma pair and a global regulatory complex. *Nucleic Acids Res* 2018;46:6726–6745. DOI: 10.1093/nar/gky475.
11. Toro N, Martinez-Abarca F, Gonzalez-Delgado A. The reverse transcriptases associated with CRISPR-Cas systems. *Sci Rep* 2017;7:7089. DOI: 10.1038/s41598-017-07828-y.
12. Gonzalez-Delgado A, Mestre MR, Martinez-Abarca F, et al. Spacer acquisition from RNA mediated by a natural reverse transcriptase-Cas1 fusion protein associated with a type III-D CRISPR-Cas system in *Vibrio vulnificus*. *Nucleic Acids Res* 2019;47:10202–10211. DOI: 10.1093/nar/gkz746.
13. Crawley A, Henriksen JR, Barrangou R. CRISPRdisco: an automated pipeline for the discovery and analysis of CRISPR-Cas systems. *CRISPR J* 2018;1:171–181. DOI: 10.1089/crispr.2017.0022.

Downloaded by 116.237.143.130 from www.liebertpub.com at 12/28/20. For personal use only.