Science Bulletin xxx (xxxx) xxx

Contents lists available at ScienceDirect

Science Bulletin

journal homepage: www.elsevier.com/locate/scib

News & Views

Expanding genome editing scopes with artificial intelligence

Li Yang [a,*], Jia Chen [b,c,*]

[a] Center for Molecular Medicine, Children's Hospital, Fudan University and Shanghai Key Laboratory of Medical Epigenetics, International Laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
[b] Gene Editing Center, School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China
[c] Shanghai Clinical Research and Trial Center, Shanghai 201210, China

The successful establishment of the clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) -based genome editing in living eukaryotic cells has encouraged scientists to screen additional CRISPR/Cas systems to expand genome editing scopes [1]. Indeed, a variety of different CRISPR/Cas systems with diverse protospacer adjacent motif (PAM) for different target preferences have been uncovered, mainly by taking advantage of computational approaches to identify homologs with sequences/motifs similar to Cas9 from microbial genomes and meta genomes (Fig. 1, left), providing a rich arsenal of genome engineering toolkits.

Cas protein consists of duel but independent functional domains for both target binding, together with programmable guide RNAs (gRNA), and cleavage on target DNA/RNA, resulting in its predominant application in genome engineering at both single gene and whole genome-wide scales. Interestingly, when combining Cas protein (as the genome locator) with other effectors, such as nucleobase (cytidine or adenosine) deaminases, genome editing can be achieved at single nucleotide resolution, more specifically referred to as base editing [2].

The apolipoprotein B mRNA editing enzyme, catalytic polypeptide (APOBEC)/activation-induced deaminase (AID) family members are classic single-strand (ss) DNA-specific cytidine deaminases, which have been shown to be contributed to the production of mutations in the genomes of retroviruses and cancer cells [3]. By fusing rat APOBEC1 (rA1) with catalytically dead Cas9 (dCas9) or Cas9 nickase (nCas9), first series of base editors (BEs) were firstly reported in 2016 to generate precise and efficient C-to-T changes at nuclear genomic DNAs [4]. Mechanically, dCas9/nCas9 is used as the locator moiety of BEs, together with gRNA, to form an R-loop structure at the target site, while the fused rA1 enzyme functions as the effector moiety to induce C-to-U editing in the ssDNA region of R-loop structure. Subsequent cellular DNA repair or replication processes can eventually install C-to-T base substitutions in nuclear genomic DNA. Later, by fusing different CRISPR/Cas proteins or their mutants with distinct cytidine deaminases, such as human APOBEC3A, efficient C-to-T base editing was also fulfilled in a broader scope, such as in genomic regions with high methylation levels [5,6]. Strikingly, although no enzymes were reported to naturally deaminate adenine in DNA, a transfer RNA (tRNA) specific adenosine deaminase was evolved to catalyze genomic DNA A-to-I editing, thus leading to A-to-G base editing, after fusing with nCas9 [7]. The availability of cytosine and adenosine BEs thus enables all four transition mutations without the requirement of double-stranded DNA cleavage. Interestingly, fusing CRISPR/Cas proteins with cytidine and/or adenosine deaminase or their mutants, variable transition and/or transversion mutations have been reported to be created [8–10].

Despite of these progresses, the application of CRISPR/Cas-based editing in mitochondrial DNA has been impeded by the difficulty of delivering gRNA into mitochondria. Instead, transcriptional activator-like effector (TALE) proteins could be used as the locator protein for the editing of mitochondrial DNA. However, as TALE proteins cannot unwind dsDNA, double-stranded DNA (dsDNA)-specific cytidine deaminases are needed to fuse with TALE proteins to install C-to-T base editing in mitochondrial DNA. In 2020, an interbacterial toxin named as double-stranded DNA deaminase toxin A ($DddA_{tox}$) was identified to catalyze cytidine deamination in dsDNA [11]. Importantly, this newly identified dsDNA-specific cytidine deaminase can be applied in not only mitochondrial DNA base editing and but also nuclear genome base editing. However, the originally identified DddA strongly prefers TC dinucleotides, which limits the application of DddA-derived cytosine base editors (DdCBEs) in broader regions. Despite of a few newly identified DddAs with different substrate preferences [12,13], the number of functional DddAs is still much less than that of APOBEC/AID family deaminases. New approach has been desired to more efficiently identify additional DddA-like deaminase to further expand base editing scopes.

In the past few years, with the advent of AlphaFold2, predicting protein structures became feasible and easy even in individual labs, which in turn benefits biological research in many specific fields. In a recently published study, Gao, Zhao, and colleagues [14] reported to use AlphaFold2 to screen for cytidine deaminases

* Corresponding authors at: Center for Molecular Medicine, Children's Hospital, Fudan University and Shanghai Key Laboratory of Medical Epigenetics, International Laboratory of Medical Epigenetics and Metabolism, Ministry of Science and Technology, Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China (L. Yang) and Gene Editing Center, School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China (J. Chen).

E-mail addresses: liyang_fudan@fudan.edu.cn (L. Yang), chenjia@shanghaitech.edu.cn (J. Chen).

https://doi.org/10.1016/j.scib.2023.10.025

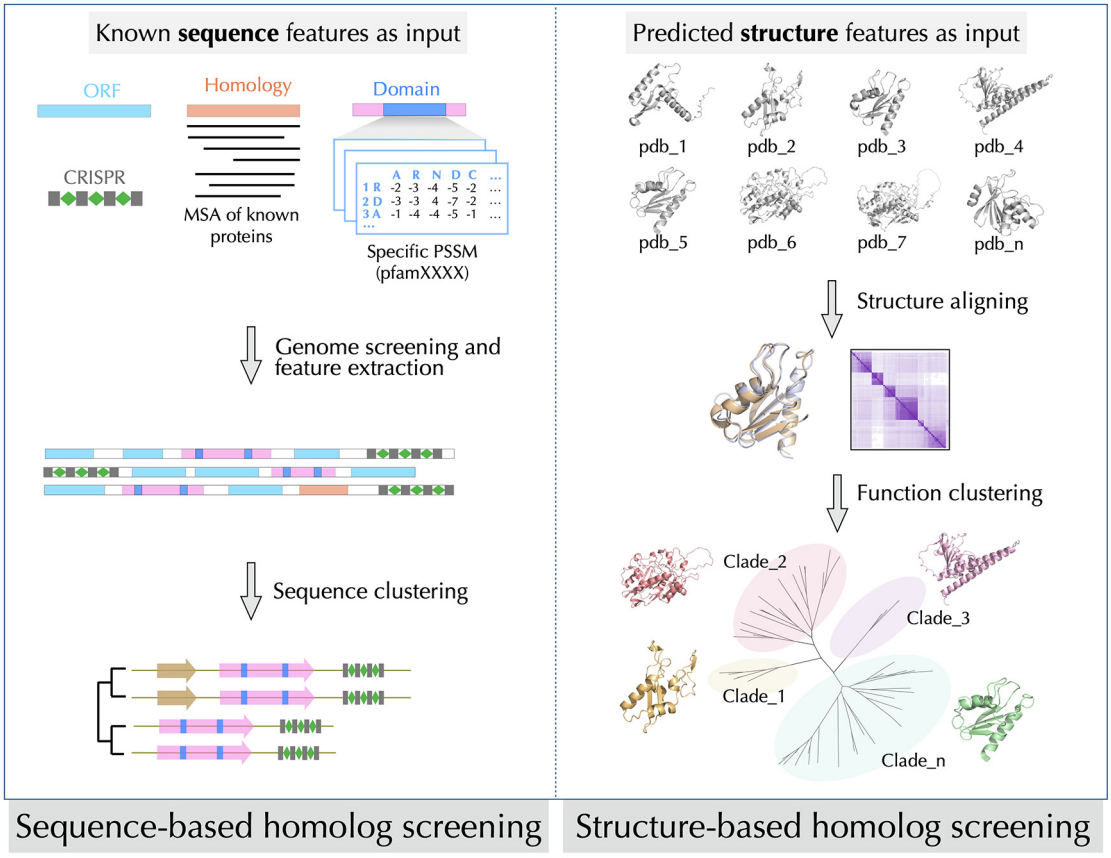


Fig. 1. Sequence-based and structure-based homolog screening strategies to identify new effectors for genome editing. Different to the canonical screening with sequence similarity for new CRISPR/Cas proteins (left), an artificial intelligence-empowered, structure-based clustering method (right) has been successfully developed to identify new cytidine deaminases for the development of new base editors. MSA: multiple sequence alignment; ORF: open reading frame; PSSM: position-specific scoring matrix.

based on structural similarities, instead of applying canonical approaches based on sequence similarity, leading to the identification of a family of DddA-like proteins with more versatile substrate preferences. With a seamless pipeline, including structure prediction by AlphaFold2, correlation by multiple structural alignments and clustering by structural similarity, hundreds of deaminase candidates were successfully sub-grouped into 20 structural clades, each with distinct and conserved structural domains. This artificial intelligence (AI)-empowered, structure-based clustering method (Fig. 1, right) was shown to provide reliable protein classification, indicating distinct catalytic functions and properties within each structural clade.

Next, the authors set a series of experiments to validate candidate proteins with predicted DddA-like structures in the application of constructing new base editors [14]. They found that some predicted DddA-like proteins have cytidine deamination activity on dsDNA. Indeed, the activities of the newly identified Ddd1, Ddd7, Ddd8 and Ddd9 are similar to or even higher than that of the original DddA. Moreover, Ddd1 and Ddd9 demonstrate a preference for GC dinucleotides and Ddd8 demonstrates a preference for WC (W = A or T) dinucleotides, therefore expanding the editing scope of DdCBE. Surprisingly, in addition to a few identified Ddds, most predicted DddA-like proteins exhibited ssDNA deamination activities, referred to as ssDNA deaminases (Sdds). Among the identified Sdds, Sdd7, Sdd9, Sdd5, Sdd6, Sdd4, Sdd76, and Sdd10 have cytidine deamination activities comparable to rA1, which was originally used in base editing [4], and Sdd7 has the highest editing efficiency among them. Of note, Sdd7-derived CBE induced editing efficiencies similar to human APOBEC3A-derived BE (hA3A-BE3) [6], which is one of the most robust CBEs. In terms of the preference for context sequence, Sdd7 and Sdd6 do not show obvious preference but rA1 showed a strong preference on TC or CC, according to the results of a 12-k library of targeted reporter anchored positional sequencing (TRAP-seq) [14].

It has been reported that canonical CBEs can induce gRNA-independent off-target (OT) editing throughout the genome of edited cells [15], raising the safety concern about CBE's application in therapeutics. By using an orthogonal R-loop assay, CBE-induced gRNA-independent OT mutations can be detected conveniently. After evaluating different identified Sdds, Sdd2, Sdd3, Sdd4, Sdd6, Sdd10, and Sdd59 induced lower gRNA-independent OT mutations than rA1 and Sdd6 barely induced OT mutations while maintaining robust on-target editing. When calculating the on-target:off-target ratios of different cytidine deaminases, Sdd6 exhibited higher ratio than rA1 and hA3A.

As an efficient and precise genome editor, BEs have great potential in treating human diseases. Adeno-associated virus (AAV) has been widely used in gene therapy because of its high tissue-specificity, low immunogenicity and inability to integrate into host genome. However, AAV can only package around 4.5 kb of cargo genes, limiting its ability to deliver large genome editors, e.g., most BEs with the canonical deaminase-Cas fusion configuration. Thus, a robust cytidine deaminase with a small size is of high demand for gene therapy. With the aid of AlphaFold2 again, truncated versions of Sdd7, Sdd6, Sdd3, Sdd9, Sdd10, and Sdd4 were generated. Among them, mini-Sdd7, mini-Sdd6, mini-Sdd3, mini-Sdd9, mini-Sdd10, and mini-Sdd4 were found to induce similar or even higher editing efficiency comparing to untruncated ones. After fusing

mini-Sdd6 to SaCas9, a previously identified small Cas9, the resulted small CBE was packaged into a single AAV and then induced up to 43.1% editing efficiency in infected mouse cells.

Although various BEs have been reported to induce efficient base editing in different plants including rice, wheat, maize and potato, it had not been successfully achieved in soybeans with unknown reasons. The research team led by Gao and Zhao [14] then decided to apply newly-identified CBEs to overcome this challenge. As a result, it was shown that Sdd7-derived CBE induced successful base editing at the target site of *GmPPO2* in soybeans, leading to carfentrazone-ethyl-resistance, and 34 heterozygotes were obtained from 154 transgenic seedlings of Sdd7-edited soybean plants. As expected, the edited soybeans by Sdd7-derived CBE grew well in the presence of carfentrazone-ethyl, compared to wild-type plants that were sensitive to wilting and could not generate roots. These results demonstrated that base editing has been realized in soybean, one of the most important staple crops in the world.

Taken together, the study led by Gao and Zhao [14] utilized a structure-based homolog screening strategy to identify new cytidine deaminase empowered by AI, leading to the identification of deaminase candidates clustered by structural features. Surprisingly, the authors found that many of the previously annotated dsDNA-specific cytidine deaminases are ssDNA-specific cytidine deaminases in fact. The newly-identified cytidine deaminases have characteristics different from the previously reported ones and certain Sdds can be further truncated to be packaged into a single AAV or induce base editing in soybeans for the first time. This study thus suggests that the AI-empowered development of new editing systems enriches the genome editing toolbox and broadens its applications in the future.

Conflict of interest

The authors declare that they have no conflict of interest.

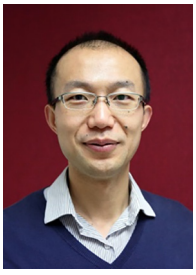
Acknowledgments

This work was supported by the Ministry of Science and Technology of China (2018YFA0801401 and 2019YFA0802804), the National Natural Science Foundation of China (31925011 and 32371514), the Ministry of Agriculture and Rural Affairs of China (NK2022010207) , the Shanghai Municipal Science and Technology Commission (21JC1404600), the Program of Shanghai Academic/Technology Research Leader (23XD1422500),and the Shanghai Municipal Education Commission (SMEC) grant to the Shanghai Frontiers Science Center for Biomacromolecules and Precision Medicine at the ShanghaiTech University. We thank Bao-Qing Gao for helping to prepare the figure.

References

[1] Wang JY, Doudna JA. Crispr technology: A decade of genome editing is only the beginning. Science 2023;379:eadd8643.
[2] Yang L, Chen J. A tale of two moieties: Rapidly evolving crispr/cas-based genome editing. Trends Biochem Sci 2020;45:874–88.
[3] Yang B, Li X, Lei L, et al. APOBEC: from mutator to editor. J Genet Genomics 2017;44:423–37.
[4] Komor AC, Kim YB, Packer MS, et al. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 2016;533:420–4.
[5] Li X, Wang Y, Liu Y, et al. Base editing with a CPF1-cytidine deaminase fusion. Nat Biotechnol 2018;36:324–7.
[6] Wang X, Li J, Wang Y, et al. Efficient base editing in methylated regions with a human apobec3a-cas9 fusion. Nat Biotechnol 2018;36:946–9.
[7] Gaudelli NM, Komor AC, Rees HA, et al. Programmable base editing of A▪T to G▪C in genomic DNA without DNA cleavage. Nature 2017;551:464–71.
[8] Zhao D, Li J, Li S, et al. Glycosylase base editors enable C-to-A and C-to-G base changes. Nat Biotechnol 2021;39:35–40.
[9] Chen L, Hong M, Luan C, et al. Adenine transversion editors enable precise, efficient A▪T-to-C▪G base editing in mammalian cells and embryos. Nat Biotechnol 2023. https://doi.org/10.1038/s41587-023-01821-9.
[10] Tong H, Wang X, Liu Y, et al. Programmable A-to-Y base editing by fusing an adenine base editor with an N-methylpurine DNA glycosylase. Nat Biotechnol 2023;41:1080–4.
[11] Mok BY, de Moraes MH, Zeng J, et al. A bacterial cytidine deaminase toxin enables CRISPR-free mitochondrial base editing. Nature 2020;583:631–7.
[12] Guo J, Yu W, Li M, et al. A DddA ortholog-based and transactivator-assisted nuclear and mitochondrial cytosine base editors with expanded target compatibility. Mol Cell 2023;83:1710–7.
[13] Mi L, Shi M, Li YX, et al. Ddda homolog search and engineering expand sequence compatibility of mitochondrial base editing. Nat Commun 2023;14:874.
[14] Huang J, Lin Q, Fei H, et al. Discovery of deaminase functions by structure-based protein clustering. Cell 2023;186:3182–95.
[15] Jin S, Zong Y, Gao Q, et al. Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. Science 2019;364:292–5.

Li Yang is a Distinguished Principal Investigator at Institutes of Biomedical Sciences, Fudan University and Children's Hospital of Fudan University. He received his B.S. and Ph.D. degrees from Lanzhou University and Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences (CAS), respectively. His research mainly focuses on developing and applying interdisciplinary technologies, including computational pipelines, artificial intelligence modes, and the-state-of-the-art base editing technologies, to decode the complex transcriptome and to recode genetic information for basic research and gene therapeutics.



Jia Chen is a tenured Principal Investigator at School of Life Science and Technology, ShanghaiTech University. He received his B.S. and Ph.D. degrees from Nankai University and Institute of Biochemistry and Cell Biology, CAS, respectively. His research focuses on DNA repair-induced mutagenesis, the development and application of new gene editing systems, and gene editing therapy.