Genome Editing

Ma Hongbao [1], Margaret Young [2], Yang Yan [1]

[1] Brookdale Hospital, Brooklyn, New York 11212, USA; [2] Cambridge, MA 02138, USA
ma8080@gmail.com

Abstract: The genome is the genetic material of an organism, which consists of DNA (or RNA in RNA viruses). The genome includes both the genes (the coding regions) and the noncoding DNA. Genome editing is a technique in which DNA is deleted, inserted or replaced in the genome of an organism using engineered nucleases that create site-specific double-strand breaks at desired locations in the genome. The induced double-strand breaks are repaired through nonhomologous end-joining (NHEJ) or homologous recombination (HR) which resulting in targeted mutations. NHEJ uses a variety of enzymes to directly join the DNA ends in a double-strand break and in HDR a homologous sequence is utilized as a template for regeneration of missing DNA sequence at the break point. Clustered regularly interspaced short palindromic repeats (CRISPR) are certain prokaryotic DNA segments that contain short repetitions of DNA sequences. Each repetition is followed by some short segments related to a bacteriophage virus or plasmid. The CRISPR/Cas system is a prokaryotic immune system. CRISPR spacers recognize and cut the exogenous DNA for an immunological function. CRISPRs exist in the bacterial genomes. The Cas9 nuclease and guide RNAs can cut genome in certain location and remove the existing genes. CRISPRs can be used for genome editing and gene regulation. Using CRISPR interference technique can alter the germline of animals.

[Ma H, Young M, Yang Y. **Genome Editing.** *Researcher* 2016;8(8):25-29]. ISSN 1553-9865 (print); ISSN 2163-8950 (online). http://www.sciencepub.net/researcher. 6. doi:10.7537/marsrsj080816.06.

Keywords: genome; editing; clustered regularly interspaced short palindromic repeats (CRISPR); DNA; bacteriophage; immune system; life; gene

The genome is the genetic material of an organism, which consists of DNA (or RNA in RNA viruses). The genome includes both the genes (the coding regions) and the noncoding DNA.

A gene is a locus of DNA which is made up of nucleotides and is the molecular unit of heredity. The transmission of genes to an organism's offspring is the basis of the inheritance of phenotypic traits. Most biological traits are under the influence of polygenes as well as the gene–environment interactions. Genes can make mutations in their sequence, leading to different variants (alleles) in the population. The alleles encode slightly different versions of a protein, which cause different phenotype traits. Genes evolve due to natural selection or survival of the fittest of the alleles. DNA (deoxyribonucleic acid) is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses on the Earth.

Genome editing is a technique in which DNA is deleted, inserted or replaced in the genome of an organism using engineered nucleases that create site-specific double-strand breaks at desired locations in the genome. The induced double-strand breaks are repaired through nonhomologous end-joining (NHEJ) or homologous recombination (HR) which resulting in targeted mutations. NHEJ uses a variety of enzymes to directly join the DNA ends in a double-strand break and in HDR a homologous sequence is utilized as a template for regeneration of missing DNA sequence at the break point.

There are currently four types of engineered nucleases being used: zinc finger nucleases, meganucleases, transcription activator-like effector-based nucleases (TALEN), and the CRISPR-Cas system.

An important method in modern biological research is to modify the DNA sequence (genotype) of an organism (or a single cell) and observe the impact of this change on the organism (phenotype), which is called reverse genetics. The forward genetics is that a new phenotype is first observed and then its genetic basis is studied.

To modify the DNA sequence of the target organism it can be achieved by: (1) Site-directed mutagenesis employing either phage- or polymerase chain reaction (PCR)-mediated methods and oligonucleotides containing the desired mutation. (2) Recombination based methods that utilize the natural ability of cells to exchange DNA between its own genetic information and an exogenous DNA.

There are 3 classes of nucleases: (1) The Zinc finger nucleases (ZFNs); (2) Transcription-activator like effector nucleases (TALEN); (3) Meganucleases.

In the genome editing work, the important thing is to find an endonuclease whose DNA recognition site and cleaving site are separate from each other. A

restriction enzyme with such properties is FokI. FokI has the advantage of requiring dimerization to have nuclease activity and this means the specificity increases dramatically as each nuclease partner would recognize a unique DNA sequence. FokI nucleases have been engineered that can only function as heterodimers and have increased catalytic activity. The heterodimer functioning nucleases would avoid the possibility of unwanted homodimer activity and thus increase specificity of the DSB. Although the nuclease portions of both ZFNs and TALEN constructs have similar properties, the difference between these engineered nucleases is in their DNA recognition peptide. ZFNs rely on Cys2-His2 zinc fingers and TALEN constructs on TALEs. Both of these DNA recognizing peptide domains have the characteristic that they are naturally found in combinations in their proteins. Cys2-His2 Zinc fingers typically happen in repeats that are 3 bp apart and are found in diverse combinations in a variety of nucleic acid interacting proteins such as transcription factors. TALEs on the other hand are found in repeats with a one-to-one recognition ratio between the amino acids and the recognized nucleotide pairs. Because both zinc fingers and TALEs happen in repeated patterns, different combinations can be tried to create a wide variety of sequence specificities. Zinc fingers have been more established in these terms and approaches such as modular assembly.

Gene therapy can replace the defective gene with a normal allele at its natural location. The expression of the partially replaced genes is more consistent with normal cell biology than full genes that are carried by viral vectors. Gene targeting through ZFNs or TALEN-based approaches can also be used to modify defective genes at their endogenous chromosomal locations. Examples include the treatment of X-linked severe combined immunodeficiency (X-SCID) by ex vivo gene correction with DNA carrying the interleukin-2 receptor common gamma chain (IL-2Rγ). Insertional mutagenesis by the retroviral vector genome induced leukemia in some patients, a problem that is predicted to be avoided by these technologies.

Clustered regularly interspaced short palindromic repeats (CRISPR) are certain prokaryotic DNA segments that contain short repetitions of DNA sequences and each repetition is followed by short segments of spacer DNA. And, each repetition is followed by some short segments related to a bacteriophage virus or plasmid. CRISPRs are found in approximately 40% of sequenced bacterial genomes and 90% of sequenced archaea.

Cas9 is an RNA-guided DNA endonuclease enzyme in bacteria. Cas9 can be used to interrogate and cleave foreign DNA, such as invading bacteriophage DNA or plasmid DNA in bacteria. If the DNA substrate is complementary to the guide RNA, Cas9 cleaves the invading DNA. CRISPR-Cas9 is a genome editing tool. The Cas9 endonuclease is a four-component system that includes two small RNA molecules. By manipulating the nucleotide sequence of the guide RNA, the artificial Cas9 system could be programmed to target any sequence in DNA for cleavage. This technological advance has fueled efforts to edit genomes with the re-engineered CRISPR-Cas9 system. CRISPR has been modified to make programmable transcription factors that allow scientists to target and activate or silence specific genes. Libraries of tens of thousands of guide RNAs are available.

The CRISPR/Cas system is a prokaryotic immune system. CRISPR spacers recognize and cut the exogenous DNA for an immunological function. CRISPRs exist in the bacterial genomes. The Cas9 nuclease and guide RNAs can cut genome in certain location and remove the existing genes. Using CRISPR interference technique can alter the germline of animals.

CRISPR's function is related to the cas genes. The cas genes express a helicase that unwinds DNA and a nuclease that digests DNA. CRISPR sequences directed cas enzymes to degrade viral DNA. Bacteria use spacers in their immune defenses. In 2015, the nuclease Cpf1 was discovered in the CRISPR/Cpf1 system of the bacterium Francisella novicida.

CRISPR repeats about 24 to 48 base pairs that show some dyad symmetry normally in a secondary structure as a hairpin. Repeats are separated by spacers. Some CRISPR spacer sequences exactly match sequences from plasmids and phages and some other spacers match the prokaryote's genome as self-targeting spacers. New spacers can be added rapidly as part of the immune response to phage infection. Small clusters of cas genes are often located next to CRISPR repeat-spacer arrays.

The 4 stages of CRISPR immunity: (1) Invading DNA is recognized and cleaved to protospacer by Cas1 and Cas2; (2) Protospacer is ligated to the direct repeat adjacent to the leader sequence; (3) Single strand extension repairs the CRISPR and duplicates the direct repeat; (4) The primary CRISPR transcript is cleaved by cas genes to produce crRNAs.

When a microbe is invaded by a virus, the first stage of the immune response is to capture viral DNA and insert it into a CRISPR locus in the form of a spacer.

Cas1 proteins have diverse amino acid sequences, which are metal-dependent nucleases/integrases that bind to DNA in a sequence-independent manner. Cas2 proteins have been characterized and possess either ssRNA- or dsDNA- specific endoribonuclease activity.

The PAM sequence appears to be important

during spacer insertion in type I-E systems. That sequence contains a strongly conserved final nucleotide (nt) adjacent to the first nt of the protospacer. This nt becomes the final base in the first direct repeat. This suggests that the spacer acquisition machinery generates single-stranded overhangs in the second-to-last position of the direct repeat and in the PAM during spacer insertion. However, not all CRISPR-Cas systems appear to share this mechanism as PAMs in other organisms do not show the same level of conservation in the final position. It is likely that in those systems, a blunt end is generated at the very end of the direct repeat and the protospacer during acquisition. Multiple CRISPRs contain many spacers to the same phage.

CRISPR-RNA (crRNA) guides the Cas nuclease to the target during the interference step. The crRNA is initially transcribed as part of a single long transcript encompassing much of the CRISPR array. This transcript is then cleaved by Cas proteins to form crRNAs. CrRNAs associate with Cas proteins to form ribonucleotide complexes that recognize foreign nucleic acids. CrRNAs show no preference between the coding and non-coding strands, which is indicative of an RNA-guided DNA-targeting system. All crRNAs contain a spacer sequence and some portion of the repeat at one or both ends. It is the partial repeat sequence that prevents the CRISPR-Cas system from targeting the chromosome as base pairing beyond the spacer sequence signals self and prevents DNA cleavage. RNA-guided CRISPR enzymes are classified as type V restriction enzymes.

The basic model of CRISPR evolution is newly incorporated spacers driving phages to mutate their genomes to avoid the bacterial immune response, creating diversity in both the phage and host populations.

CRISPRs are widely distributed among bacteria and archaea and show some sequence similarities. CRISPRs are evolutionarily conserved and cluster into related types. Many show signs of a conserved secondary structure.

Through the CRISPR/Cas mechanism, bacteria can acquire immunity to certain phages and thus halt further transmission of targeted phages. CRISPR/Cas-based RNA-guided nucleases can be used to target virulence factors, genes encoding antibiotic resistance, and other medically relevant sequences of interest.

Although researchers came to an understanding that CRISPR-Cas was involved in microbial immunity, the mechanism of how the spacers conferred immunity remained a mystery. Koonin and colleagues proposed that spacers produce small RNA guides to target RNA transcribed from viral DNA, analogous to the RNA interference system used by eukaryotic cells. Others hypothesized that CRISPR sequences directed Cas enzymes to degrade viral DNA.

The CRISPR-Cas9 system consists of two key molecules that introduce a change into the DNA: (1) The enzyme Cas9, which acts as a pair of molecular scissors that can cut the two strands of DNA at a specific location in the genome so that bits of DNA can then be added or removed. (2) The guide RNA (gRNA), which consists of a small piece of pre-designed RNA sequence (about 20 bases long) located within a longer RNA scaffold. The scaffold part binds to DNA and the pre-designed sequence guides Cas9 to the right part of the genome. This makes sure that the Cas9 enzyme cuts at the right point in the genome.

The guide RNA is designed to find and bind to a specific sequence in the DNA. The guide RNA has RNA bases that are complementary to those of the target DNA sequence in the genome. The guide RNA will only bind to the target sequence and no other regions of the genome. The Cas9 follows the guide RNA to the same location in the DNA sequence and makes a cut across both strands of the DNA.

Spacer DNA is a region of non-coding DNA between genes. In bacteria, spacer DNA sequences are only a few nucleotides long, and in eukaryotes they can be extensive and include repetitive DNA, comprising the majority of the DNA of the genome. In ribosomal DNA the spacers are within or between gene clusters that are called internal transcribed spacer and external transcribed spacers, respectively. In animal the mitochondrial DNA genes generally have very short spacers. In fungi mitochondrial DNA spacers are common and variable in length, and they may also be mobile.

Repeated sequences are DNA or RNA multiple copies throughout the genome. In many organisms, a significant fraction of the genomic DNA is highly repetitive, with over two-thirds of the sequence consisting of repetitive elements in human. The repetitive DNA segments could be remainders from past evolution or autonomous self-replicating sequences hacking the cell machinery to proliferate. The disposition of repetitive elements consists either in arrays of tandemly repeated sequences, or in repeats dispersed throughout the genome. In human, several classes of repetitive elements present a high tendency for co-localization within the nuclear space, suggesting that DNA repeats positions can be used by the cell as a genome folding map.

The immune system is a host defense system comprising many biological structures and processes within an organism that protects against disease. To function properly, an immune system must detect a wide variety of agents, known as pathogens, from viruses to parasitic worms, and distinguish them from the organism's own healthy tissue. In many species, the immune system can be classified into subsystems,

such as the innate immune system versus the adaptive immune system, or humoral immunity versus cell-mediated immunity. In humans, the blood–brain barrier, blood–cerebrospinal fluid barrier, and similar fluid–brain barriers separate the peripheral immune system from the neuroimmune system which protects the brain.

Cpf1 is an RNA-guided endonuclease of a class II CRISPR/Cas system. This acquired immune mechanism is found in Prevotella and Francisella bacteria. It prevents genetic damage from viruses. Cpf1 genes are associated with the CRISPR locus, coding for anendonuclease that use a guide RNA to find and cleave viral DNA. Cpf1 is a smaller and simpler endonuclease than Cas9, overcoming some of the CRISPR/Cas9 system limitations. CRISPR/Cpf1 could have multiple applications, including treatment of genetic illnesses and degenerative conditions. In 2015, the nuclease Cpf1 was discovered in the CRISPR/Cpf1 system of the bacterium *Francisella novicida*. Cpf1 showed several key differences to Cas9 including, such as: (1) causing a staggered cut in double stranded DNA as opposed to the blunt cut produced by Cas9, (2) relying on a T rich Protospacer adjacent motif, (3) requiring only a CRISPR RNA (crRNA) for successful targeting.

Clustered Regularly Interspaced Short Palindromic Repeats from Prevotella and Francisella 1 (CRISPR/Cpf1) is a DNA-editing technology analogous to the CRISPR/Cas9 system. Cpf1 is an RNA-guided endonuclease of a class II CRISPR/Cas system found in Prevotella and Francisella bacteria, which prevents genetic damage from viruses. Cpf1 genes are associated with the CRISPR locus coding for an endonuclease that use a guide RNA to find and cleave viral DNA. The Cpf1-crRNA complex cleaves target DNA or RNA by identification of a protospacer adjacent motif 5'-YTN-3' (where Y is a pyrimidine and N is any nucleobase) or 5'-TTN-3', in contrast to the G-rich PAM targeted by Cas9. After identification of PAM, Cpf1 introduces a sticky-end-like DNA double-stranded break of 4 or 5 nucleotides overhang. Cpf1 requires 1 RNA molecule to cut DNA while Cas9 needs 1 RNA molecule.

Cpf1 leaves one strand longer than the other, creating sticky ends and Cas9 cuts both strands in a DNA molecule at the same position, leaving behind blunt ends. Using Cpf1 it is more able to insert new sequences at the cut site.

The CRISPR/Cpf1 system consists of a Cpf1 enzyme and a guide RNA that finds and positions the complex at the correct spot on the double helix to cleave target DNA. CRISPR/Cpf1 systems activity has three stages: (1) Adaptation: Cas1 and Cas2 proteins facilitate the adaptation of small fragments of DNA into the CRISPR array. (20 Formation of crRNAs: processing of pre-cr-RNAs producing of mature crRNAs to guide the Cas protein. (3) Interference: the Cpf1 is bound to a crRNA to form a binary complex to identify and cleave a target DNA sequence.

As a new tool in CRISPR genome editing, Cpf1 has many advantages comparing to Cas9: It requires only a single RNA that CRISPR RNA assembly is simpler; its staggered cleavage patterns may facilitate substituting existing DNA with desired sequences; and it recognizes thymidine-rich DNA sequences, which has been less explored than the guanosine-rich sequences recognized by Cas9. Cpf1 as a highly specific programmable tool that is suitable for precision genome editing.

CRISPR repeats range in size from 24 to 48 base pairs. Small clusters of *cas* genes are often located next to CRISPR repeat-spacer arrays. Comparative genomics identified multiple *cas* genes have as many as 45 *cas* gene families. The classification groups CRISPR-Cas systems are classified as 2 classes. Class 1 systems use a complex of multiple Cas proteins to degrade foreign nucleic acids. Class 2 systems use a single large Cas protein for the same purpose. Class 1 is divided into types I, III, and IV; class 2 is divided into types II and V. The five system types are divided into 16 subtypes. Each type and most subtypes are characterized by a signature gene found exclusively in the category. Classification is also based on the complement of *cas* genes that are present. Most CRISPR-Cas systems have a Cas1 protein. The phylogeny of Cas1 proteins generally agrees with the classification system. Many organisms contain multiple CRISPR-Cas systems suggesting that they are compatible and may share components. The sporadic distribution of the CRISPR/Cas subtypes suggests that the CRISPR/Cas system is subject to horizontal gene transfer during microbial evolution. CRISPR/Cas9 is a very useful tool in genome editing with it's simplicity, high efficiency and versatility of the system.

For simple gene editing, a single cut site can generate indel mutations using the NHEJ repair pathway. When within coding exons, such indels can cause frameshifts and disrupt protein expression. Targeting coding exons towards the beginning of the gene may be preferable, as mutations may create more complete gene disruption and be less likely to accidentally generate truncated protein artefacts with residual biological activity. The technology for genetic engineering has progressed rapidly in the past few years. The ability to easily and efficiently edit animal genomes using custom-engineered nucleases has already greatly expanded studies of gene function and holds great potential for constructing modified

animal cell what could be as safer gene therapies, drug production and agriculture.

References

1. Baidu. http://www.baidu.com. 2016.
2. Google. http://www.google.com. 2016.
3. http://dharmacon.gelifesciences.com/gene-editing/crispr-cas9/?wt.mc_id=PaidSearch_Google_ALL&wm_camp=Google_ALL&gclid=CNiMyeSi3c4CFQNZhgod-TsE_w#all. 2006.
4. http://phys.org/news/2016-06-crispr-genome-cpf1-specificity-mutant.html#jCp.
5. http://www.genscript.com/CRISPR-gRNA-constructs.html?src=google&gclid=CMWA-Y2m3c4CFZdZhgod9E8Hog. 2016.
6. http://www.yourgenome.org/facts/what-is-crispr-cas9. 2016.
7. https://en.wikipedia.org/wiki/Cas9. 2016.
8. https://en.wikipedia.org/wiki/CRISPR. 2016
9. Ma H, Chen G. Stem cell. The Journal of American Science 2005;1(2):90-92. doi:10.7537/marsjas010205.14. http://www.jofamericanscience.org/journals/am-sci/0102/14-mahongbao.pdf.
10. Ma H, Cherng S. Eternal Life and Stem Cell. Nature and Science. 2007;5(1):81-96. doi:10.7537/marsnsj050107.10. http://www.sciencepub.net/nature/0501/10-0247-mahongbao-eternal-ns.pdf.
11. Ma H, Cherng S. Nature of Life. Life Science Journal 2005;2(1):7-15. doi:10.7537/marslsj020105.03. http://www.lifesciencesite.com/lsj/life0201/life-0201-03.pdf.
12. Ma H, Yang Y. Turritopsis nutricula. Nature and Science 2010;8(2):15-20. doi:10.7537/marsnsj080210.03. http://www.sciencepub.net/nature/ns0802/03_1279_hongbao_turritopsis_ns0802_15_20.pdf.
13. Ma H. The Nature of Time and Space. Nature and science 2003;1(1):1-11. doi:10.7537/marsnsj010103.01. http://www.sciencepub.net/nature/0101/01-ma.pdf.
14. National Center for Biotechnology Information, U.S. National Library of Medicine**.** http://www.ncbi.nlm.nih.gov/pubmed. 2016.
15. Susan M. Byrne, Prashant Mali, George M. Church. Genome Editing In Human Stem Cells. Methods Enzymol. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4408990/. 2016.

8/22/2016