Aftab Ahmad
Sultan Habibullah Khan
Zulqurnain Khan *Editors*

# CRISPR Crops

## The Future of Food Security

Springer

# CRISPR Crops

Aftab Ahmad • Sultan Habibullah Khan • Zulqurnain Khan
Editors

# CRISPR Crops

## The Future of Food Security

Springer

*Editors*
Aftab Ahmad
Center for Advanced Studies in Agriculture and Food Security (CASAFS)
University of Agriculture
Faisalabad, Pakistan

Sultan Habibullah Khan
Centre of Agricultural Biochemistry and Biotechnology (CABB)
University of Agriculture
Faisalabad, Pakistan

Center for Advanced Studies in Agriculture and Food Security (CASAFS)
University of Agriculture
Faisalabad, Pakistan

Zulqurnain Khan
Institute of Plant Breeding and Biotechnology
MNS University of Agriculture
Multan, Pakistan

ISBN 978-981-15-7141-1 ISBN 978-981-15-7142-8 (eBook)
https://doi.org/10.1007/978-981-15-7142-8



This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

This volume is a piece of commitment and dedication with voluntary and extensive collaboration among various parties. We wish to express our sincere gratitude to all the authors and contributors who strived to make this venture fruitful. The genome-editing research work has been emphasized in 2012 by taking a start with TALEs and TALENs. Many reviews and research articles were published focusing on the application of the technology for solving problems in economically important crops. We would like to extend our gratitude to Prof. Caixia Gao and Prof. Weicai Yang, IGDB, Chinese Academy of Sciences, Beijing, China and Professor Yinong Yang, Pennsylvania State University, USA for being our collaborators and part of the workshop on CRISPR-based genome editing in Pakistan. The share of the country rapporteurs and our students cannot be forgotten. Our students made a lot of effort in compiling the final manuscript. We hope that the readers will benefit from this book by getting the latest review and discussion on the subject matter. The way in which the book has been composed is very comprehensive and reader-friendly making concepts palpable. A big proportion of the credit goes to the publisher for their support in publishing the idea. This book is an endeavor toward understanding the concept, application, and regulation of the CRISPR technology in the field of crop sciences. We hope that along with the agricultural scientific community, scientists, researchers, and students of other fields will also found this book beneficial to understand the technology and its safe usage.

At the end, once again, we are thankful to all of the persons who have floated, wrote, composed, and published the idea in the shape of a book for the scientific community.

Faisalabad, Pakistan — Aftab Ahmad
Faisalabad, Pakistan — Sultan Habibullah Khan
Multan, Pakistan — Zulqurnain Khan
January, 2020

# Contents

# About the Editors

**Aftab Ahmad** is currently working as Assistant Professor at Department of Biochemistry, US-Pakistan Center for Advanced Studies in Agriculture and Food Security (CAS-AFS), University of Agriculture Faisalabad (UAF), Pakistan. His research interests include developing virus-resistant and salt-tolerant plants using CRISPR/Cas technology. He has published more than 15 research articles in international peer-reviewed journals including the Journal of Biomedical Science, Scientific Reports-Nature, PLOS One, Plant Molecular Biology Reporter, Molecules, Journal of Cereal Science, and Genetic Resources and Crop Evolution. He pursued his PhD and postdocs in Plant Molecular Biology from University of Shizuoka, Japan under Center of Excellence (COE, Twenty-First Century) program. He also served as visiting scientist in Department of Plant Sciences, University of California, Davis, USA. He has been teaching biochemistry and molecular biology at UAF and his research focuses in the area of CRISPR/Cas-based genome editing in plants.

**Sultan Habibullah Khan** received his PhD from the University of Manchester, England in the field of Plant Sciences in 2004. Since his graduation from UK, he has been working as a faculty at the University of Agriculture Faisalabad (UAF). His postdoctoral experience also includes teaching and research at the University of California, Berkeley and Sultan Qaboos University, Oman. Currently, he is working as the Deputy Chief of Party at the US-Pakistan Center for Advanced Studies in Agriculture and Food Security (CAS-AFS) and a Tenured Associate Professor of Plant Molecular Sciences at the Centre of Agricultural Biochemistry and Biotechnology (CABB), UAF. For the last 20 years, he has been engaged in the fields of plant biotechnology, breeding, genetics, and molecular biology. He is a recipient of national/international scholarships, Research Productivity Award (2014), and over 40 international publications with impact factor.

**Zulqurnain Khan** is currently working as Assistant Professor in the Institute of Plant Breeding and Biotechnology (IPBB), Muhammad Nawaz Sharif University of Agriculture, Multan (MNSUAM), Pakistan. He earned his PhD in Biotechnology from University of Agriculture Faisalabad (UAF), Pakistan and his research focuses on CRISPR/Cas9 system for the development of resistance against begomoviruses

using model plants. He has published one book chapter, three review articles, and three research articles in international peer-reviewed journals including the Journal of Biomedical Sciences, Plant Molecular Biology Reporter, and International Journal of Agriculture and Biology. He also worked as a visiting researcher with a group of Professor Caixia Gao at the Institute of Genetics and Developmental Biology (IGDB), Chinese Academy of Sciences (CAS), Beijing, China.

# Abbreviations

| | |
|---|---|
| ABEs | Adenine base editors |
| BLESS | Breaks labeling and enrichment on streptavidin and sequencing |
| Bt | *Bacillus thuringiensis* |
| Cas | CRISPR-associated protein |
| CBEs | Cytosine base editors |
| CBP | Cartagena biosafety protocol |
| CDE | CRISPR-mediated directed evolution |
| CEO | CRISPR-edited organisms |
| CRE | Cis-regulatory elements |
| CRISPR | Clustered regularly interspaced short palindromic repeats |
| CRISPRi | Cluster interspaced specific short palindromic repeats interference |
| crRNAs | CRISPR ribonucleic acid |
| CTLs | Complex trait loci |
| DSB | Double strand break |
| dsRNA | Double-stranded RNA |
| *EPSPS* | 5-Enolpyruvylshikimate-3-phosphate |
| EU | European Union |
| FAO | Food and Agriculture Organization |
| FSANZ | Food Standards Australia New Zealand |
| GDP | Gross domestic product |
| GE | Genome editing |
| GMOs | Genetically modified organisms |
| gRNAs | Guide RNA |
| HDR | Homology directed repair |
| HRM | High resolution melt |
| HTGTS | High-throughput genome-wide translocation sequencing |
| ICPs | Insecticidal crystalline proteins |
| IDLV | Integrase-deficient lentiviral vectors |
| indel | Insertion or deletion |
| IR | Insect-resistant |
| KI | Knock in |
| KO | Knockout |
| LEA | Late embryogenesis abundant |

| | |
|---|---|
| NHEJ | Non-homologous end joining |
| NPBT | New plant-breeding technique |
| PAGE | Polyacrylamide gel electrophoresis |
| PAM | Protospacer adjacent motif |
| PEG | Polyethylene glycol |
| PNT | Plant novel traits |
| rDNA | Recombinant DNA |
| RGEN | RNA-guided engineered nucleases |
| RISC | RNAi-induced silencing complex |
| RNAi | RNA interference |
| RNPs | Ribonucleoproteins |
| SDNs | Site-directed nucleases |
| siRNA | Short interfering RNA |
| SIT | Sterile insect technique |
| TALENs | Transcription activator-like effector nucleases |
| TDNA | Transfer DNA |
| TET | Ten-eleven translocation |
| TRV | Tobacco rattle virus |
| UAS | Upstream activating system |
| US | United States |
| WHO | World Health Organization |
| ZFNs | Zinc finger nucleases |

# CRISPR/Cas System: An Introduction

1

Nayla Munawar and Aftab Ahmad

**Abstract**

The process of altering genetic code to modify DNA precisely and efficiently within a cell is known as genome editing. All techniques used for changing a gene in a cell are based on cutting of double-stranded DNA through DNA endonucleases at specific site. DNA nucleases used for genome editing are composed of programmable DNA binding domains to bind with target sequence followed by DNA cleavage domain to break double-stranded DNA that excites error prone non-homologous end joining (NHEJ) or homology-directed repair (HDR) at specific genome locations. Programmable nucleases enabled researchers to manipulate practically any genomic sequence, provided opportunities to create cell lines and animal models to study human diseases, and also promoting new possibilities of treating human diseases by gene therapy. Four genome editing technologies, meganucleases, zinc-finger nucleases (ZFNs), transcription activator-like effector nucleases (TALENs), and clustered regularly interspaced short palindromic repeats (CRISPR/Cas9) based on the above-mentioned principle have been developed so far and have been used successfully for the correction of disease-causing mutations, for addition of therapeutic gene and the deletion of specific genes from the specific sites in the genome. Genome editing revolutionized the scientific world and have intense impact on various fields of biotechnology, such as biopharmaceutical production, agriculture, creation of transgenic organisms and cell lines, regulation and function of the genome, etc. All four genome editing techniques, their history, principles,

N. Munawar (✉)
Department of Chemistry, United Arab Emirates University, Al-Ain, UAE

A. Ahmad
Center for Advanced Studies in Agriculture and Food Security (CASAFS), University of Agriculture, Faisalabad, Pakistan
e-mail: aftab.ahmad@uaf.edu.pk



A. Ahmad et al. (eds.), *CRISPR Crops*,
https://doi.org/10.1007/978-981-15-7142-8_1

advantages, disadvantages, and applications, with a special focus on CRISPR/Cas, have been discussed in this chapter.

## 1.1 Genome Editing: An Overview

The alteration of specific DNA sequences in the cellular genome, referred to as genome editing, has gained popularity in recent years. Genome editing can alter the phenotypic characteristics of cells.

The discovery of the DNA double-helix structure in the mid-twentieth century and then an improved understanding of the transfer of genetic material from parents to off-spring have led researchers to conclude that minor changes in the DNA sequence can cause disease. The term "genome" refers to the complete DNA sequence of an organism and can be altered by intrinsic factors such as DNA replication errors during cell division or environmental factors such as UV radiation from the sun. Moreover, DNA mutations can be acquired from parents and can be present in virtually every cell in the body. Most DNA variations do not result in an abnormal protein and are called "polymorphisms." These DNA polymorphisms can lead to diverse phenotypic characteristics, such as differences in hair color and type, eye color, and blood type, and do not affect the health of an individual. However, some DNA variations can increase the risk for certain disorders and may cause serious health issues.

The identification of disease-causing "DNA errors" has led researchers to develop strategies to prevent or reverse genetic diseases. On the basis of this notion, the concept of gene editing or trait alteration was introduced in the late 1980s and later led to the discovery of different genome editing technologies. The first genome editing study to target specific genes by homologous recombination (HR) in mouse embryonic stem cells to generate knockin and knockout cells was reported by Capecchi et al. in 1989. While the process was revolutionary, it was highly inefficient as it required extensive selection and screening to identify the one-in-a-million cell expressing the modified gene.

Although, the recombination frequency was low because of the low rate of HR in mammalian cells, this proof-of-concept study suggested that it was possible to target a specific gene with precision. The method to achieve targeted DNA modifications using a sequence-specific nuclease was first introduced in the 1990s. The Cre-lox system comprising a site-specific DNA endonuclease Cre, which recognizes the 34-nt locus *loxP*, to generate double-stranded DNA breaks (DSBs) and facilitate subsequent recombination at these sites was developed in the early 1990s. This technology led the successful knockout of any desired gene and was used as the most effective gene editing tool during the time to develop transgenic mouse models. While the Cre-lox system provided an opportunity to control gene expression more easily than HR, the efficiency of the system was low because of the large genetic distance between the *loxP* sites.

After developing this successful genome editing technique, the search to develop a reliable method to generate a DSB at any desired location for targeted editing began. Subsequently, zinc-finger nucleases (ZFNs) consisting of a DNA binding domain and a nuclease domain, were used for the first time to incorporate changes in the genome of *Drosophila melanogaster* fruit fly in 2002 and to repair a mutation in the human genome in 2005. Despite being time-consuming and laborious, this technique is being employed in a growing number of applications currently. Almost a decade ago, another genome editing technique called transcription activator-like effector nucleases (TALENs) was developed in 2010 and successfully used for genome modification in more than 40 different organisms and cell types. However, limitations of both ZFNs and TALENs propelled the search for alternative gene targeting technologies. In 2012, the most efficient and widely used genome editing technology of clustered regularly interspaced short palindromic repeats (CRISPR)/Cas associated system 9 (CRISPR/Cas9) was discovered and proved to be the most reliable genome editing technique so far.

## 1.2 Techniques for Genome Editing

The introduction of site-specific DSBs is a prerequisite for genome editing. Hence, by using the intrinsic DNA repair machinery of cellular organisms, tools that generate DNA DSBs can be used to precisely alter the genome. After the success of Cre-lox DNA endonuclease-based gene editing technology in 1992, researchers began to identify nucleases (enzymes that cleave nucleic acids) that can generate DNA DSBs as well as methods to direct these nucleases to specific regions of the genome.

Eventually, four genome editing techniques were established over the following 20 years. These techniques use designer/programmable nucleases to generate DSBs and employ the cellular DNA repair mechanism to precisely modify genomic sequences (Voytas 2013). The foremost characteristics of a successful genome editing technique are (1) target selectivity (nucleotide sequences, epigenetic markers), (2) precision for the target, and (3) regulation of its activity. The following four major classes of customizable DNA binding proteins currently exist for genome editing, and their popularity, applicability, effectiveness, and pitfalls can be determined on the basis of the above-mentioned factors.

1. Meganucleases
2. ZFNs
3. TALENs
4. CRISPR/Cas

### 1.2.1 Meganucleases

Meganucleases (MNs), also known as homing nucleases, are derived from microbial mobile genetic elements (Smith et al. 2006). These are sequence-specific endonucleases with a large restriction site (14–25 nt) and are classified into five families based on their sequences and structure motifs: (1) LAGLIDADG, (2) GIY-YIG, (3) HNH, (4) His-Cys box, and (5) PD-(D/E)XK (Orlowski et al. 2007; Zhao et al. 2007). The LAGLIDADG homing nuclease (LHE family) is the most abundant form and is found in Archaea, Bacteria, Eukarya, and their respective viruses (Stoddard 2005). These meganucleases have been extensively studied and are often termed as "selfish genetic elements" because their role in the host is not yet known. The LHE family endonucleases exist in both homo-dimeric and monomeric forms and can bind to target DNA sequences of up to 24 nt in length. Because meganucleases can recognize specific DNA sequences through protein–DNA interactions, efforts were made to manipulate LAGLIDADG protein–DNA interactions. However, several studies have shown that related proteins can use different subsets of residues to recognize similar DNA sequences (Chevalier et al. 2003; Lucas et al. 2001). Hence, it was difficult to re-engineer the specificity of these meganucleases. Nevertheless, site-specific cleavage of a maize genomic sequence (Gao et al. 2010) and two human genes *XPC* and *RAG1* (Smith et al. 2006; Arnould et al. 2007) by engineered meganucleases has been reported in literature.

Despite being the only nuclease naturally evolved for genome editing, it has been difficult to modify meganucleases to specifically recognize target DNA sequences. Thus, they have not achieved widespread acceptance as a genome editing technique. Moreover, the high probability of off-target binding is another limitation of this technique (Argast et al. 1998). However, a minor advantage of meganucleases is their small size (approximately 40 kDa). They can be used in viral vectors where short coding sequences are required. Nevertheless, because of their inability to predict and modify target DNA sequences, meganucleases are less preferred than other genome editing techniques, including ZFNs and TALENs, which are more flexible with regard to re-targeting the reagent to different sequences using a distinct DNA-cutting domain that can be non-specific and a specific programmable DNA binding domain.

### 1.2.2 Zinc-Finger Nucleases (ZFNs)

ZFNs, a genome editing tool, is based on eukaryotic transcription factors (Urnov et al. 2010; Miller et al. 2011). Zinc fingers are small peptide domains in which a zinc ion is coordinated in a tetrahedral geometry to provide structural stability. These domains bind to specific DNA sequences in a precise manner. The currently used ZFNs consist of restriction endonucleases and a zinc-finger DNA binding domain to target specific DNA sequences within the genome. The natural type IIS restriction enzyme *Fok*I was first discovered by Chandrasegaran (Chandrasegaran and Carroll 2016), and consists of distinct DNA binding and cleavage domains. Researchers

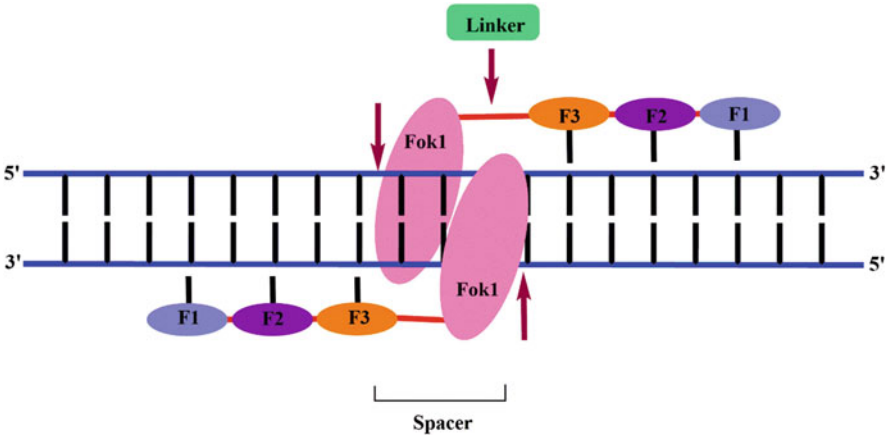


**Fig. 1.1** Mechanism of ZFNs: An illustration of a zinc-finger nuclease (ZFN). A ZFN consists of three to six zinc-finger protein monomers and the *Fok*I restriction enzyme, which cleaves DNA upon dimerization. Each ZFP recognizes a 3-nt target DNA sequence

cleaved the DNA binding domain of *Fok*I and fused its DNA cleavage domain with the zinc-finger motif to construct a distinct DNA recognition and cleavage tool called ZFN (Kim et al. 1996). The association of the zinc-finger DNA binding domain with the DNA cleavage domain enables the enzymatic machinery to target a unique locus in the genome and provoke endogenous DNA repair. The mechanism of action of this technique is shown in Fig. 1.1. The DNA binding domain identifies and binds to a specific DNA sequence. Thereafter, *Fok*I is recruited to the DNA sequence to induce a DSB. To cleave a specific site in the genome, a pair of ZFNs is designed to recognize two sequences flanking the site, one on the forward strand and the other on the reverse strand. Each DNA binding domain of ZFN identifies three nucleotides on either side of the DNA strand. Upon binding to the DNA strand, the *Fok*I domains dimerize and cleave the DNA at the site to generate a DSB, which could be repaired either by non-homologous end joining (NHEJ) or homology-directed repair (HDR) mechanism (Fig. 1.5).

Because each ZFN can recognize a 3-nt DNA sequence, it is possible to assemble 6–7 unique ZFNs to specifically target 18–21 nt in the genome. Hence, sequence-specific chimeric nucleases can be constructed feasibly. Because of its ease of application and high specificity, this technique has gained widespread popularity and has been adopted by many researchers for different applications ranging from genome modification to gene therapy in both animals and plants (Townsend et al. 2009; Shukla et al. 2009; Urnov et al. 2010; Mashimo et al. 2010; Kim et al. 2009). Theoretically, it is possible to target any gene in any organism using this technique. However, the assembly of functional ZFPs with the desired DNA binding specificity is a major limitation of this technique because it requires an extensive screening process. The binding of zinc-finger domains to an extended stretch of nucleotides with high affinity is difficult to achieve (Ramirez et al. 2008). Moreover, the

requirement to assemble zinc fingers in new combinations for each new target DNA sequence made it difficult for non-specialists to routinely engineer ZFNs. To overcome this difficulty, an academic consortium developed an open source library of zinc-finger components and protocols to perform screening to identify ZFNs that bind with high affinity to a desired sequence. Nonetheless, it can take months for non-specialists to construct optimized ZFNs (Maeder et al. 2008). Taken together, the construction of site-specific zinc-finger modules is a rate-limiting step in utilizing this technology for genome editing. It is partially due to this reason, ZFNs were not excessively used for genome editing by the scientific community.

### 1.2.3 Transcription Activator-Like Effector Nucleases (TALENs)

Although, artificially designed ZFNs generated enormous excitement in the field of genome editing, the discovery of TALENs in 2011 was recognized in "Nature Methods" as the "method of precise genome editing." TALENs are based on transcription activator-like effector proteins from plant pathogenic bacteria of *Xanthomonas* spp. (Christian et al. 2010; Miller et al. 2011; Boch et al. 2009; Moscou and Bogdanove 2009). The endogenous function of these proteins is to activate specific host genes by mimicking eukaryotic transcription factors to facilitate the growth of the pathogen (Schornack et al. 2006). These effector proteins contain a DNA binding domain consisting of monomers, with each monomer bound to one nucleotide in the target DNA sequence (Römer et al. 2007). The monomers are tandem repeats of 34 amino acids with two highly variable amino acids at positions 12 and 13 in each repeat. These residues have been termed as repeat variable di-residues (RVDs) and are responsible for DNA specificity of TALE proteins. The DNA specificity of TAL effector proteins can be modulated by manipulating these RVDs. To synthesize artificial DNA binding modules of TALE proteins, monomers containing RVDs, such as Asn and Ile (NI), Asn and Gly (NG), two Asn (NN), and His and Asp (HD) for binding the nucleotides A, T, G, and C, respectively, have been utilized (Nemudryi et al. 2014). The construction of artificial TALENs and their mode of action resemble that of ZFNs (Fig. 1.2). However, unlike ZFNs, the DNA binding domain of TALENs recognize only one nucleotide, thereby enhancing the site specificity with fewer off-target effects (Chandrasegaran and Carroll 2016). Moreover, the chimeric fusion of the *Fok*I nuclease domain with a combination of TALE modules resulted in a robust and effective programmable genome editing tool.

TALENs offer two distinct advantages for genome editing over ZFNs: (1) no requirement of selection or directed evolution to engineer TALE arrays, resulting in a considerable reduction in the amount of time and experience needed to assemble a functional nuclease, (2) provide higher specificity and reduced toxicity than ZFNs because of their increased affinity for target DNA (Meckler et al. 2013) or a greater energetic penalty for association with base mismatches (Mussolino et al. 2014). On the other hand, the construction of a site-specific DNA binding module is challenging because of high similarity between TALE recognition sequences; this restricts

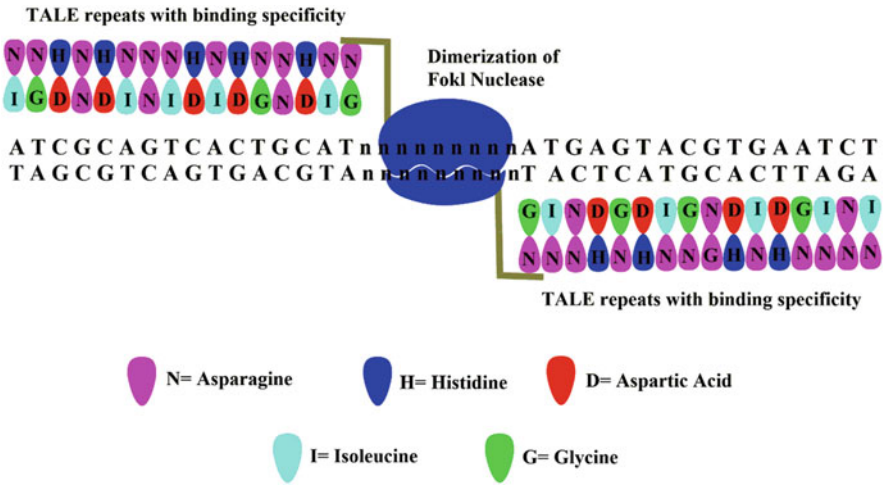


**Fig. 1.2** Mechanism of action of TALENs: An illustration of a transcription activator-like effector nuclease (TALEN) showing left and right monomers of TALE proteins and the *Fok*I restriction enzyme, which cleaves DNA upon dimerization. Each TALE protein recognizes a single DNA base pair

the wide use of TALENs in genome engineering. Moreover, the requirement of the presence of thymine (T) at start of the TALEN binding sequence, specific target-site length, and a spacer between two TALEN arms for the formation of the *Fok*I dimer negatively affect its routine application. Attempts are undertaken to reduce the limitations of TALENs and improve their efficacy by adapting approaches, such as using *Fok*I variants and nickases to reduce off-target binding (Miller et al. 2011), using adenoviral vectors for TALEN delivery to difficult-to-transfect cell types (Holkers et al. 2014; Maggio et al. 2016), and constructing a library of TALENs targeting 18,740 human protein-coding genes to facilitate new studies (Kim et al. 2013). However, the repetitive sequences in TALENs render the construction of novel TALE arrays labor extensive and costly. While the engineering of modular DNA binding proteins is challenging, the discovery of the CRISPR-Cas methodology with a novel mode of DNA recognition simplifies the development of custom nucleases.

### 1.2.4 Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR-Associated Protein 9 (Cas9) System

Most recently, the CRISPR/Cas system has been reported as an efficient and facile alternative of ZFNs and TALENs for inducing targeted genetic modifications. The microbial adaptive immune system CRISPR can target any genome location of choice using a synthetic short guide RNA (sgRNA) (Jinek et al. 2012).

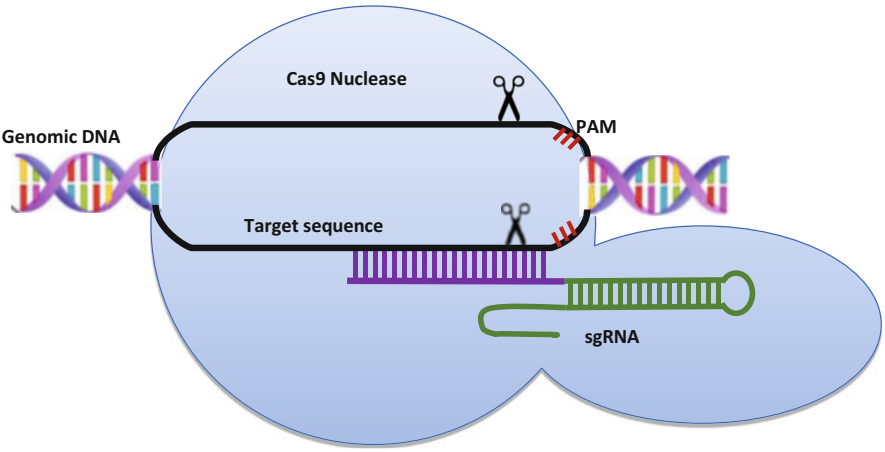


**Fig. 1.3** General structure of CRISPR-Cas locus with all components. Cas9 indicated in blue color. Protospacer adjacent motif (PAM) "NGG" for Cas9 shown in red color. Pairing of 20 nt of sgRNA (purple color) with target site in genome

The general structure of the CRISPR-Cas locus is shown in Fig. 1.3. The locus harbors an array of short repetitive sequences (repeats) interspaced by unique, short non-repetitive sequences (spacers) that originate from mobile genetic elements (MGEs), such as bacteriophages and transposons, followed by an AT-rich leader sequence, and a set of diverse CRISPR-associated *cas* genes that encode Cas proteins (Barrangou et al. 2007; Haft et al. 2005; Makarova et al. 2006; Mojica et al. 2005).

### 1.2.5 Mode of Action

The general mechanism of action of the CRISPR/Cas adaptive immunity system, regardless of its different types, is divided into three steps: (1) adaptation, (2) expression and maturation, and (3) interference.

These three stages are shown in Fig. 1.4. In adaptation, distinct short DNA fragments from the invading pathogen (known as protospacers) are recognized by the Cas protein and integrated into CRISPR repeats as a new spacer, which then serves as a genetic record to develop immunological memory in the host and enables it to recognize same invading pathogens in the future. Several CRISPR/Cas systems require a short 3–5 nt protospacer adjacent motif (PAM) in the target DNA to acquire the protospacer (Deveau et al. 2008; Shah et al. 2013; Bolotin et al. 2005).

Expression and maturation involves the transcription of the CRISPR array into a precursor-CRISPR RNA (pre-RNA), which is further processed to yield short mature CRISPR RNA (crRNA) through endo-nucleolytic cleavage and contain the memorized sequences of the invading pathogen (Carte et al. 2008; Haurwitz et al.

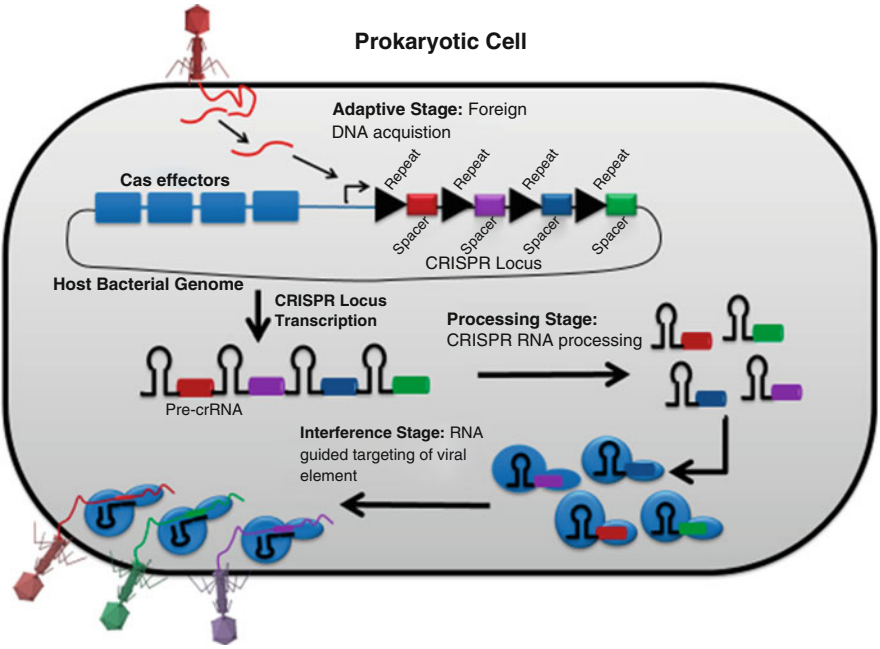


**Fig. 1.4** General mechanism of action of the CRISPR-Cas system. Adaptive stage: repeats and spacer are expressed as pre-crRNAs. Processing stage: pre-crRNAs are processed into mature crRNA.Interference stage: mature crRNA along with Cas nuclease (Cas9) binds with target site in phage genome and creates DSBs

2010). Each crRNA contains a single spacer (short RNA segment, complementary to the DNA sequence of the foreign genetic element) at its 5′ end and a CRISPR repeat sequence at its 3′ end. The mature crRNA associates with one or more Cas effector proteins to form an active Cas-crRNA effector complex.

During the interference stage, this effector complex scans and targets the foreign nucleic acids in the cell. The crRNA part of the complex functions as a guide to recognize target DNA using the specific PAM sequence upstream or downstream of the protospacer (Fonfara et al. 2016; Zetsche et al. 2015; Fineran et al. 2014; Jiang et al. 2013; Semenova et al. 2011; Westra et al. 2013) through Watson–Crick base pairing. The successful recognition of the target DNA leads to its cleavage and destruction by Cas nuclease.

In recent years, diverse classes of the CRISPR immune system based on different effector Cas proteins and PAM recognition sequences have been identified in various microorganisms. These classes have been described in detail below in Sect. 1.3.2.

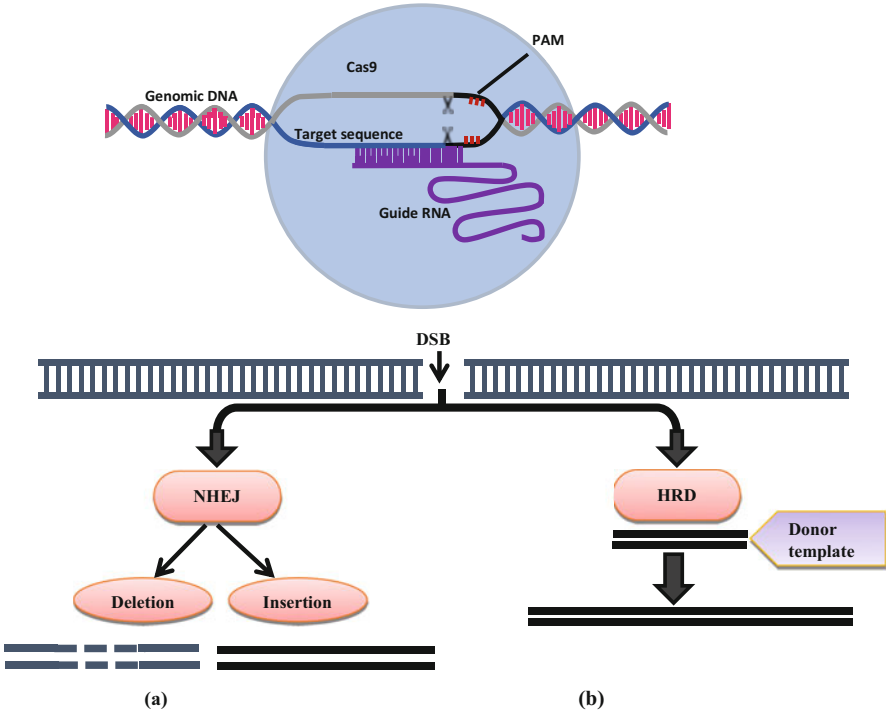


**Fig. 1.5** Mechanisms of gene editing: DNA double strand breaks (DSBs) can be repaired by two mechanisms: (**a**) Non-homologous end joining (NHEJ), which occurs in the absence of a donor template and introduces small base insertions or deletions that can result in gene disruption. (**b**) Homology-directed repair (HDR), which occurs in the presence of a homologous donor template

## 1.3 Molecular Mechanism of Genome Editing

All modern genome editing techniques rely on the principle of inducing a site-specific DNA DSB that can be subsequently repaired using inherent cellular repair mechanisms, such as HDR or the error prone NHEJ pathway (Shalem et al. 2015). These two pathways of DNA repair in the cell permit the insertion, deletion, or mutation of nucleotides in the target DNA, resulting in incorporating desired changes in the genome. Although the NHEJ repair mechanism is error prone and often results in variable lengths of insertion and deletion (indel) variants, it can be used for gene knockout (Fig. 1.5a). On the other hand, the HDR mechanism is error-free because it involves recombination between homologous DNA sequences obtained from the undamaged chromatid of the cell or if engineered, can use an extrinsic homologous donor DNA template to alter the genome (Fig. 1.5b).

In plants, DSBs are mainly repaired by NHEJ. In this process, several enzymes are used to re-anneal the broken ends of DSBs without the requirement of a homologous DNA template (Puchta 2005). NHEJ occurs throughout the cell cycle

in higher eukaryotes and exhibits low fidelity repair. Because of its error prone nature, NHEJ repair often leads to the addition or deletion of nucleotides and causes alterations at targeted DNA DSB sites. In many cases, the indels introduced in exons during NHEJ repair can lead to missense or nonsense mutations, and in some cases, result in a complete loss of gene function. A majority of the published articles on plant genome editing has used the NHEJ pathway to knockout genes (Mladenov and Iliakis 2011).

In the error-free HDR pathway, a homologous sequence serves as a template to repair the DSBs. The HDR pathway can be used to precisely modify nucleotide sequences and perform gene replacement or insertion at target loci in the presence of an exogenous donor DNA as a repair template. Unlike the NHEJ pathway, the HDR pathway occurs mainly during the S and G2 phases of the cell cycle and has low efficiency in higher eukaryotes. Only one in a million treated cells undergoes the desired genome modification (Capecchi 1989). However, the balance between HDR and NHEJ pathways can be shifted towards HDR by inactivating some components of the NHEJ pathway (Beumer et al. 2008, 2013; Bozas et al. 2009). The knockout of DNA ligase IV increased HDR frequency from 20% to 65% in *Drosophila.* This result suggests the possibility of limiting a specific type of repair mechanism in the cell; however, a common method to enhance HDR does not exist because of different cell types and overall complexity of the repair mechanism in the cell (Chandrasegaran and Carroll 2016).

DSBs are resolved by the NHEJ pathway in the absence of an intrinsic donor DNA sequence. This may result in insertion or deletion (indels) and ultimately gene knockout. On the other hand, DSBs are repaired by the HDR pathway in the presence of an intrinsic or extrinsic donor DNA sequence, ultimately resulting in gene knockin (Bibikova et al. 2002). Overall, the combination of these two repair mechanisms is effective for modifying the eukaryotic genome (Hsu et al. 2014). Genome editing systems use sequence-specific nucleases (SSNs) that consist of a DNA binding domain to provide sequence specificity and a nuclease domain to introduce DNA DSBs at the targeted site (Puchta 2005). These DSBs can be repaired by either of the above-mentioned repair mechanisms.

### 1.3.1 A Bacterial and Archaeal Immune System as a Genome Editing Tool in Eukaryotes

All living organisms are continuously exposed to infectious agents such as viruses and transposons. Consequently, the host organisms have developed two mechanisms to defend themselves from these pathogens: (1) innate immunity and (2) adaptive immunity. Innate immunity is naturally present in the genetic system of the host and is responsible for the recognition of the invading pathogen and provides non-specific protection. On the other hand, adaptive immunity is characterized by mounting a specific immune response against the invading pathogen. The immune system memorizes the antigenic properties of the invading pathogen to provide protection to the host upon re-exposure. Adaptive immunity was presumed to be a characteristic

only of eukaryotes until the previous decade. However, the discovery of CRISPR loci and CRISPR-associated proteins (Cas) in both bacteria and archaea demonstrated the presence of the adaptive immune system in prokaryotes as well (Mojica et al. 2000).

Approximately 50% and 90% of the sequenced genomes of bacteria and archaea, respectively, possess adaptive immunity in the form of the CRISPR/Cas system (Van der Oost et al. 2009). This system consists of a "memory chip" and "DNA scissors" in the form of the CRISPR loci. As mentioned in Sect. 1.2.4, the CRISPR loci store the immunological memory of an invading pathogen in the form of a "spacer," a short DNA sequence of the invading pathogen. The second essential component is the variable Cas gene cassette. This cassette is typically located adjacent to a CRISPR locus and produces Cas proteins, which are essential elements of the adaptive immunity system. The defensive role of CRISPR/Cas against pathogens has been well established since its discovery (Barrangou et al. 2007; Brouns et al. 2008; Sapranauskas et al. 2011). Moreover, CRISPR/Cas proteins are involved in other cellular functions such as gene regulation, inhibition of biofilm formation, multicellular development, and virulence (Sampson et al. 2013; Zegans et al. 2009; Vercoe et al. 2013). However, the precise under lying mechanisms of these functions remain unclear (Rath et al. 2015). After identifying the CRISPR/Cas system as an adaptive immune mechanism in various microorganisms, several attempts were undertaken by researchers to understand the mechanism of action of these modules in the cell (Marraffini and Sontheimer 2008; Plagens et al. 2015). These studies identified of distinct classes of the CRISPR/Cas system (Haft et al. 2005) and paved the way for these systems to be used an effective genome editing technology (Jinek et al. 2012; Cong et al. 2013; Mali et al. 2013).

## 1.3.2 Classification of the CRISPR/Cas System

### 1.3.2.1 Classification

The CRISPR-Cas modules (loci) have been identified as an as adaptive immune mechanism in archaea and bacteria and provide sequence-specific protection against foreign DNA or RNA. The CRISPR/Cas system is classified into two broad classes based on variability in the CRISPR loci corresponding to these modules in different organisms: "Class 1" and "Class 2," which are further subdivided into types and subtypes on the basis of comparative genomic analysis, structure and biochemical activities of CRISPR components (Koonin et al. 2017), domain architecture of effector proteins, and PAM sequence identification. The major difference between the two classes lies in the type of the crRNA-effector complexes. In class 1 systems, the crRNA-effector complex exists as a multi-subunit complex, whereas in class 2 systems, a single protein such as Cas9 performs all functions of the effector complex. Moreover, class 2 systems are less complex than class 1 systems.

The remarkable variability in the genomic architecture of CRISPR-Cas loci because of the constant threat to adaptive immunity from viruses and foreign plasmids poses a major challenge for the consistent classification of CRISPR/Cas

systems. However, a consistent classification is essential for a robust characterization of CRISPR-Cas loci in new genomes and to advance research in the field. A strategy to combine signature genes with elements of the *Cas* loci was proposed in 2015 to assign nearly all of the detected CRISPR-Cas loci to specific subtypes. This classification strategy is compatible with the results of sequence-based clustering of crRNA-effector complexes and can be adopted for the classification of newly discovered CRISPR/Cas variants from new genomes.

The class 1 CRISPR/Cas system is categorized into three types: type I, type III, and type IV, which are further divided into subtypes. Similarly, the class 2 CRISPR/Cas system is categorized into three types, which are further divided into subtypes (Fig. 1.6). The two principle modules of the CRISPR/Cas system include the adaptation module, which is responsible for spacer acquisition and the effector module, which involves pre-crRNA processing, target recognition, and cleavage. The adaptation module is largely uniform across the CRISPR/Cas system and consists of endonucleases and the structural subunits Cas1 and Cas2 (Amitai and Sorek 2016). However, the variability of effector modules is high between CRISPR types and subtypes.

Within class 1, type I and type III CRISPR/Cas systems are present most commonly in archaea and less frequently in bacteria. Type IV CRISPR/Cas system is rare and lacks the adaptation module. The type I effector complex (CRISPR-associated complex for antiviral defense) consists of a 61-nt mature crRNA and five Cas proteins with an irregular stoichiometry.

Type I and type III CRISPR/Cas systems possess a complex architecture, with a long backbone of repeat-associated mysterious proteins (RAMPs), including Cas5 and Cas7. These proteins contain an RNA recognition motif (RRM) fold and additional large and small subunits (Zhao et al. 2014; Van Der Oost et al. 2014; Jackson et al. 2014; Jackson and Wiedenheft 2015; Hochstrasser et al. 2014, 2016; Staals et al. 2013, 2014). The effector complexes, which accommodate the guide RNA, consist of one Cas5 and multiple Cas7 subunits. The Cas5 subunit binds to 5′-crRNA and interacts with the large subunit of Cas8 in type I and Cas10 in type III. Cas6 is loosely associated with the effector complex and functions as a repeat-specific RNase in pre-crRNA processing (Charpentier et al. 2015; Niewoehner and Jinek 2016). Therefore, the type I CRISPR/Cas system utilizes Cas5 or Cas6 for pre-processing of crRNA; further cleavage requires Cas3 nucleases, cascade, and crRNA for interference (Khan 2019). Different subunits of the type I system require distinct PAM sequences for target acquisition and recognition (Fig. 1.6). Type I-A system found in *Sulfolobus islandicus* and *Sulfolobus solfataricus* requires a 5′-CNN-3′ PAM motif for interference (Gudbergsdottir et al. 2011; Lillestøl et al. 2009). Type I-B system established in haloarchaea recognizes six distinct PAMs (Fig. 1.6). Type I-C system—characterized by a Cas5-dependent crRNA maturation pathway (Nam et al. 2012)—identifies an "NTTC" consensus PAM sequence in *Bacillus halodurans* (Sorek et al. 2013). Type I-E effector complex from *E. coli* and *Thermobifida fusca* bind to 5′-AAG-3′ PAM sequence (Xiao et al. 2017). Type I-F effector complex from *Pseudomonas aeruginosa* consists of four Cas proteins instead of five protein found in type I-E, and targets foreign DNA using a PAM

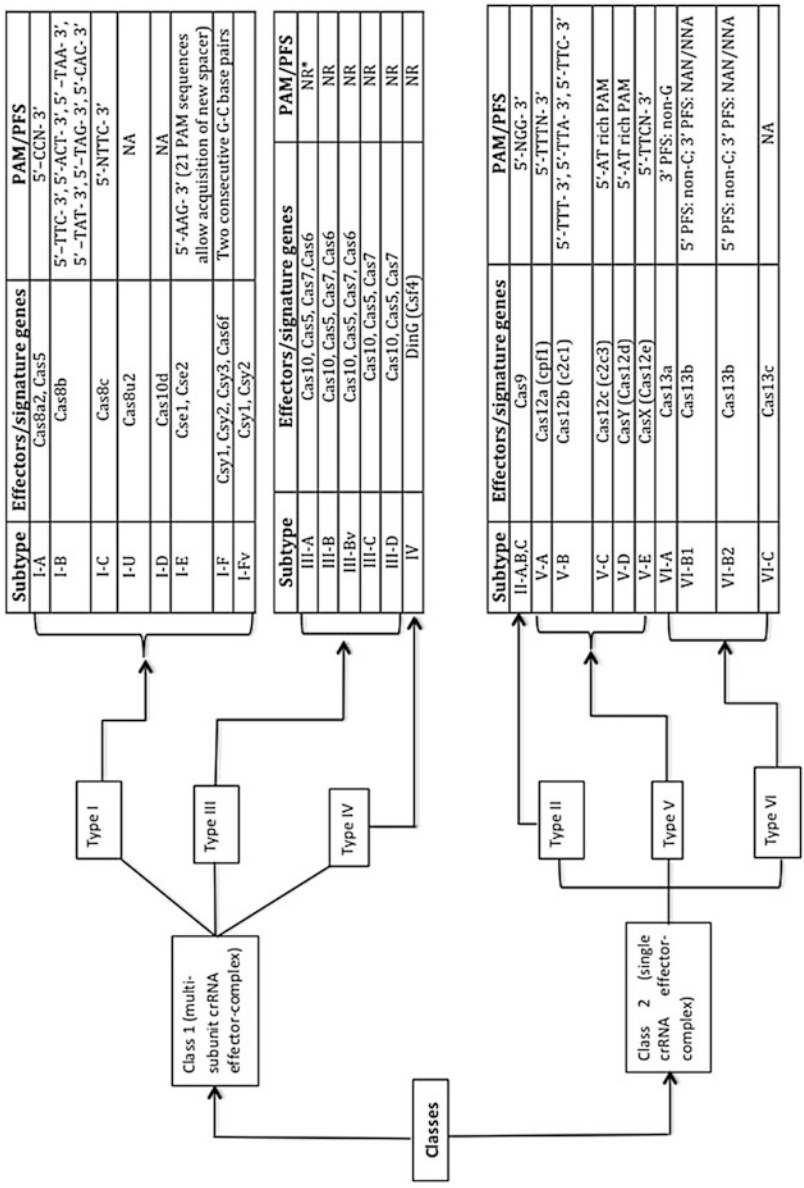


**Fig. 1.6** Experimentally characterized classes and types of CRISPR system consisting of different effectors and PAM sequences. For each CRISPR/Cas subtype, signature genes and PAM/PFS sequences are shown. NR is representing "not required" and NA indicating "not available"

element consisting of two consecutive G-C base pairs (Rollins et al. 2015). Type I-Fv is a variant of type I-F and was discovered in *Shewanella putrefaciens*. The PAM recognition elements are absent in this variant due to the lack of small and large cascade subunits (Rollins et al. 2015).

Type III CRISPR system is largely found in archaea and some bacteria. It has been divided into four subtypes based on the differences in their adaptation, interference, and recognition strategies. Whereas type III-A system possesses an adaptation module, type III-B, -C, and -D systems lack the adaptation genes; hence, these systems depend on the other systems to incorporate new spacer. Similar to the type I system, interference in the type III system is performed by a crRNA-guided multiprotein complex called Csm/Cmr complex (Rath et al. 2015). The Csm complex is present in A and D subtypes, while the Crm complex is present in B and C subtypes. Similar to the type I system, the type III system uses Cas6 for crRNA processing. Because of the scarcity of structural data, PAM recognition sequence for many subtypes of the class 1 CRISPR system is not yet known. A rare type IV CRISPR system includes an elementary CRISPR-Cas locus that lacks the adaptation domain

The effector module in the class 2 CRISPR/Cas system consists of a single, large, multidomain protein. Hence, the CRISPR-Cas loci are organized more uniformly in the class 2 system than in class 1 system. The class 2 system is also categorized into three types: II, V, and VI. Type II is the most studied and well-characterized system among all types of class 2 systems and consists of the effector Cas9 endonuclease that is widely used in genome editing. In this system, the recruitment of Cas9 to target DNA is regulated by crRNA. However, the mechanism by which RNase III, trans-activating RNA (tracrRNA), and protein factors perform the 5′ end processing of crRNA is not yet determined.

The type V CRISPR/Cas system consists of a single effector protein Cas12 (Dong et al. 2016; Yang et al. 2016; Yamano et al. 2016), and is categorized into five known subtypes A-E and a predicted U subtype. Subtype A of the type V (type V-A) CRISPR system consists of Cas12a (Cpf1), an active RNA-guided endonuclease that does not require additional tracrRNA for target cleavage (Zetsche et al. 2015). This subtype uses the "TTN" PAM sequence for target recognition (Gleditzsch et al. 2019). The type V-B system utilizes both tracrRNAs and crRNAs to target dsDNA using a T-rich PAM sequence such as TTT, TTA, or TTC. Other variant systems such as type-VD and -VE consist of CasY (Cas12d) and CasX (Cas12e) signature proteins, respectively. As opposed to CasY, CasX requires tracrRNA for interference. Moreover, while CasY employs a "TA" PAM sequence for target recognition, CasX employs a 5′-TTCN-3′ PAM sequence (Burstein et al. 2017).

All subtypes of the type IV system consist of HEPN domains (higher eukaryotes and prokaryotes nucleotide-binding domains). These subtypes are predicted to possess RNase activity instead of DNA cleavage activity. Moreover, the activity of the type VI system relies on the presence of a protospacer flanking site (PFS), a PAM analogue for RNA targets (Gleditzsch et al. 2019). Type VI-A and -B systems rely on Cas13a and Cas13b effectors, respectively, for their interference activity.

The diversity of the CRISPR-Cas systems in the cell is caused by the presence of novel effector proteins and new molecular mechanisms. Bacteria and archaea possess different CRISPR systems. Type II systems are exclusively found in bacteria, whereas type III systems are more common in archaea. These findings have helped researchers to utilize the unique characteristics of CRISPR systems to design novel and robust technologies for genetic modifications.

Cas12 is frequently used for genome editing due to its relatively small size, no requirement for tracrRNA, and asymmetric cleavage sites. Strategies to expand the range of Cas12 targets by designing different protein variants with different PAM specificities are recently being tested (Gao et al. 2017).

### 1.3.3 History of CRISPR/Cas

The CRISPR/Cas system was developed over 20 years, across 12 cities, in 9 countries (Fig. 1.7). Briefly, this system was first reported by Franscisco Mojica in 1993, who began his doctoral studies at University of Alicante in the Mediterranean port of Santa Pola on Spain's Costa Blanca in 1989. Mojica found multiple copies of a near-perfect, roughly palindromic, repeat sequence of 30 nt, separated by spacers of approximately 36 nt that did not resemble any family of repeats known in

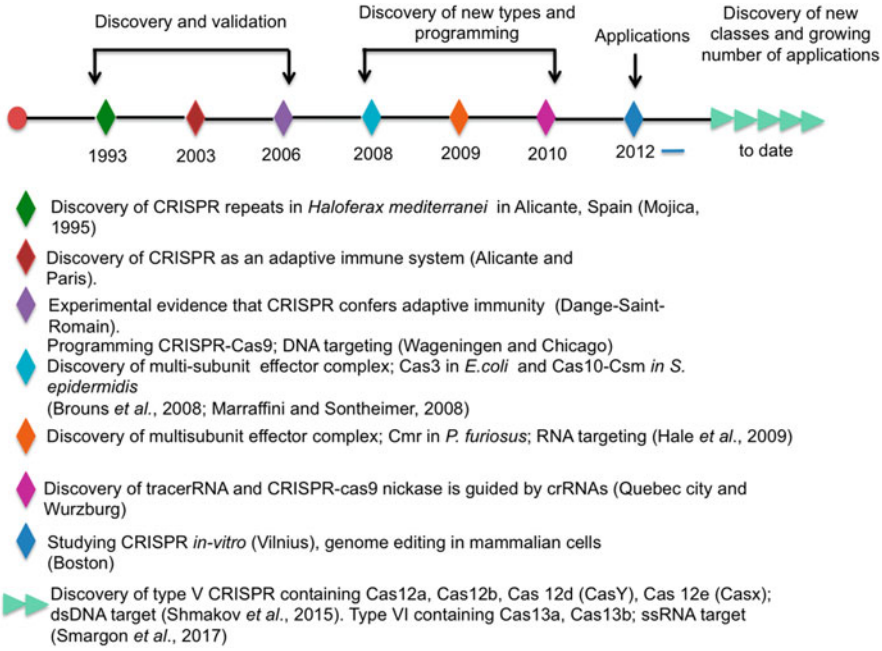


**Fig. 1.7** Timeline of CRISPR transformation from nature to technology

microbes, in a DNA fragment from *Haloferax mediterranei,* a halophilic microorganism. These repeats were named as "clustered regularly interspaced palindromic repeats" or CRISPR in 2002 (Mojica and Garrett 2013). In 2003, after performing a detailed bioinformatics analysis of these repeats, Mojica realized that the CRISPR loci belonged to the adaptive immune system that protects microbes against infections from invading pathogens (Mojica et al. 2005; Bolotin et al. 2005).

Horvath et al. in 2005 confirmed the correlation between CRISPR activity and adaptive immune response in microorganisms. They used a phage-sensitive *Streptococcus thermophilus* strain to study the mechanism of phage resistance in this bacterium. They found that the resistant strains acquired phage-derived sequence at their CRIPSR loci, resulting in enhanced immunity against subsequent infection (Barrangou et al. 2007). They also studied the role of Cas7 and Cas9 proteins. Cas7 was not found to be involved in eliciting an immune response; however, it was required to generate new spacers and repeats. On the other hand Cas9 was found to possess nuclease activity and play an active role in adaptive immunity (Cas9 is referred to as Cas5 in CRISPR-related publications prior to 2012).

Later on, van der Oost et al. in 2008 studied the CRISPR system of *E. coli* by inserting it in another *E. coli* strain lacking its endogenous CRISPR system. They characterized the most complex CRISPR system (class 1, type I) that contains Cas3 instead of Cas9. They also reported that CRISPR arrays are transcribed to small crRNAs that consist of individual spacers to guide Cas nuclease activity.

In the same year, it was demonstrated that Cas possesses DNA nuclease activity, but not RNA cleavage activity, in type III-A CRISPR system from *Staphylococcus epidermidis* (Marraffini and Sontheimer 2008). However, type III-B CRISPR system from *Pyrococcus furious* revealed the presence of crRNA-directed RNA cleavage activity in addition to DNA nuclease activity (Hale et al. 2009, 2012). The importance of the PAM sequence for spacer acquisition and nuclease activity of type II Cas9 was demonstrated by Deveau et al. (2008). More importantly, they demonstrated that type I and II CRISPR systems prevent self-targeting by preventing the binding of PAM sequences to direct repeat sequences. However, a mismatch between the 5′ end of crRNA and target DNA is required for interference in type III CRISPR systems (Marraffini and Sontheimer 2010).

By 2010, studies on the functional mechanism of the native type II CRISPR system revealed that only Cas9 possesses DNA cleavage activity (Garneau et al. 2010). Moreover, it was found that non-coding tracrRNAs hybridize with crRNA to facilitate RNA-guided targeting of Cas9. This finding facilitated the engineering of a simple RNA-programmable DNA endonuclease for genome editing. Moreover, the dual RNA hybrid structure of Cas9 and endogenous RNAase III is required for the formation of mature crRNA from the CRISPR array primary transcript (Deltcheva et al. 2011). These studies demonstrated that at least three components (Cas9, mature crRNA, and tracrRNA) are essential for the reconstitution of the type II CRISPR system. In 2011, the successful reconstitution of CRISPR interference of type II CRISPR locus obtained from *Streptococcus thermophiles* and transformed into *E. coli* validated the transferable properties of the CRISPR system (Sapranauskas et al. 2011). Until 2012, - in-vitro DNA cleavage using purified Cas9 from

*Streptococcus thermophiles* or *Streptococcus pyogenes* guided by crRNA provided a strong basis for utilizing type II CRISPR system for genome editing. In 2013, a few research groups, Cong et al. (2013) and Mali et al. (2013) successfully engineered the type II CRISPR system from *Streptococcus thermophiles* and *Streptococcus pyogenes* for genome editing in eukaryotic cells.

Since then, researchers have reported the use of CRISPR-based genome editing in organisms such as yeast, fruit fly, nematode, mouse, zebrafish, and monkey. Moreover, new types of CRISPR systems have been discovered, their biology has been elucidated in detail, the existing technology for genome editing has been improved (Barrangou and Marraffini 2014; Hsu et al. 2014; Van Der Oost et al. 2014; Sander and Joung 2014; Jiang and Marraffini 2015; Sternberg and Doudna 2015; Wright et al. 2016; Smargon et al. 2017).

## 1.4 CRISPR: Experimental Design and Considerations

Scientific advancement requires continuous development of novel tools and technologies. Protein-based nuclease systems such as meganucleases, ZFNs, and TALENs have been used for DNA sequence editing; however, their restricted application have led researchers to design more convenient gene editing strategies to modify DNA sequences. The discovery of the CRISPR/Cas system and the understanding of its physiological role in prokaryotic cells paved the way to modify the innate genome editing mechanisms into targeted genome editing technology. Different classes and types of the CRISPR system require different components to target specific DNA sequences using different mechanisms. Among all types of natural CRISPR systems, the type II CRISPR system is the best re-engineered genome editing system in eukaryotes. The type II CRISPR/Cas9 system consists of two components: Cas9 protein that harbor the DNA binding and cleavage domains, and crRNA to guide Cas9 protein to target DNA with the help of tracrRNA. The crRNA and tracrRNA duplex work synergistically to dock Cas9 protein onto the target DNA. However, Cas9 can recognize the target DNA only in the presence of a short (2–5 nt) PAM sequence adjacent to the target DNA. Cas9 obtained from different sources can identify different PAM sequences, as described in the previous section. The ease of operation and the following beneficial characteristics of the type II CRISPR system facilitated its use as a robust, specific, and simple genome editing technology.

1. crRNA and tracrRNA can be combined into a single guide RNA (sgRNA) (Jinek et al. 2012).
2. Unlike Cas3, which degrades the target DNA, Cas9 produces a single DNA DSB, a prerequisite for gene editing.
3. Cas9 can be re-targeted by designing a sgRNA sequence.
4. Cas9 nuclease domain can be inactivated while retaining its target specificity, thereby allowing researchers to regulate the transcription of a desired gene.

The CRISPR/Cas system can be transformed into cells using conventional methods as well as microinjection. The generation of DNA DSBs by Cas nuclease triggers the activation of DNA repair pathways in the cell. The NHEJ pathway repairs DSBs in a non-template directed manner, resulting in the addition or deletion (indels) of nucleotides in the target DNA. On the other hand, precise nucleotide sequence modifications, gene replacement, and gene insertion at target loci is achieved by HDR in the presence of a homologous template.

Knowledge regarding different CRISPR systems and cellular DNA repair mechanisms has been extensively utilized to design the most popular gene editing technology—CRISPR-Cas—today a billion-dollar industry.

## 1.5 CRISPR/Cas: Beyond Genome Editing

CRISPR/Cas9 is the most widely used genome editing technique in both medical and agricultural fields. In contrast to ZFNs and TALENs where target DNA specificity and nuclease activity rely on the fusion of the sequence-specific DNA binding domain and the nuclease domain of the protein, the CRISPR system uses Cas9 protein to target any desired DNA sequence using a sequence-specific guide RNA (sgRNA), which is a hybrid of the naturally occurring tracrRNA: crRNA duplex. The use of sgRNA has simplified the application of the CRISPR system for genome editing. The application of ZFNs and TALENs for genome editing is restricted because of their complex re-engineering requirement for each target. However, the simple design of CRISPR/Cas9 has facilitated its use for genome editing in virtually all commonly studied eukaryotes, ranging from yeast to plants and from zebrafish to humans (Terns and Terns 2014; Sampson and Weiss 2014).

The application of the CRISPR technology is not restricted to genome editing. Owing to the programmable targeting capability of catalytically inactive Cas9, dCas9-based techniques have been used for several other genome manipulating applications. The Cas9 molecular scissors have been engineered into a versatile delivery tool by deactivating its catalytic activity through D10A and H804A substitutions. The deactivated Cas9 (dCas9)/sgRNA system can deliver different gene regulatory components to a specific target DNA, thereby acting as a shuttle without disrupting the double-stranded DNA. Therefore, dCas9/sgRNA can be used as a powerful tool for the transcriptional control of target genes. The CRISPRi (CRISPR interference) technique has been developed to knockdown gene expression with the aid of dCas9 (Qi et al. 2013). The fusion of dCas9 with a strong repressor complex such as Kruppel-associated box (KRAB) improves gene repression to a higher degree than dCas9 alone (Gilbert et al. 2013). Conversely, the fusion of dCas9 with the transcriptional activation domains VP16 and VP64 generated a gene activation platform. The fusion of dCas9 with a tripartite transcriptional complex, which is composed of VP64, P65, and Rta (VPR) proteins, induced gene expression (Chavez et al. 2015).

The programmable capacity of dCas9 has also been exploited for epigenome editing by recruiting various epigenetic writers and erasers to specific loci.