Recent advances in genome engineering technologies based on the CRISPR-associated RNA-guided endonuclease Cas9 are enabling the systematic interrogation of mammalian genome function. Analogous to the search function in modern word processors, Cas9 can be guided to specific locations within complex genomes by a short RNA search string. Using this system, DNA sequences within the endogenous genome and their functional outputs are now easily edited or modulated in virtually any organism of choice. Cas9-mediated genetic perturbation is simple and scalable, empowering researchers to elucidate the functional organization of the genome at the systems level and establish causal linkages between genetic variations and biological phenotypes. In this Review, we describe the development and applications of Cas9 for a variety of research or translational applications while highlighting challenges as well as future directions. Derived from a remarkable microbial defense system, Cas9 is driving innovative applications from basic biology to biotechnology and medicine.

Introduction

The development of recombinant DNA technology in the 1970s marked the beginning of a new era for biology. For the first time, molecular biologists gained the ability to manipulate DNA molecules, making it possible to study genes and harness them to develop novel medicine and biotechnology. Recent advances in genome engineering technologies are sparking a new revolution in biological research. Rather than studying DNA taken out of the context of the genome, researchers can now directly edit or modulate the function of DNA sequences in their endogenous context in virtually any organism of choice, enabling them to elucidate the functional organization of the genome at the systems level, as well as identify causal genetic variations.

Broadly speaking, genome engineering refers to the process of making targeted modifications to the genome, its contexts (e.g., epigenetic marks), or its outputs (e.g., transcripts). The ability to do so easily and efficiently in eukaryotic and especially mammalian cells holds immense promise to transform basic science, biotechnology, and medicine (Figure 1).

For life sciences research, technologies that can delete, insert, and modify the DNA sequences of cells or organisms enable dissecting the function of specific genes and regulatory elements. Multiplexed editing could further allow the interrogation of gene or protein networks at a larger scale. Similarly, manipulating transcriptional regulation or chromatin states at particular loci can reveal how genetic material is organized and utilized within a cell, illuminating relationships between the architecture of the genome and its functions. In biotechnology, precise manipulation of genetic building blocks and regulatory machinery also facilitates the reverse engineering or reconstruction of useful biological systems, for example, by enhancing biofuel production pathways in industrially relevant organisms or by creating infection-resistant crops. Additionally, genome engineering is stimulating a new generation of drug development processes and medical therapeutics. Perturbation of multiple genes simultaneously could model the additive effects that underlie complex polygenic disorders, leading to new drug targets, while genome editing could directly correct harmful mutations in the context of human gene therapy (Tebas et al., 2014).

Eukaryotic genomes contain billions of DNA bases and are difficult to manipulate. One of the breakthroughs in genome manipulation has been the development of gene targeting by homologous recombination (HR), which integrates exogenous repair templates that contain sequence homology to the donor site (Figure 2A) (Capecchi, 1989). HR-mediated targeting has facilitated the generation of knockin and knockout animal models via manipulation of germline competent stem cells, dramatically advancing many areas of biological research. However, although HR-mediated gene targeting produces highly precise alterations, the desired recombination events occur extremely infrequently (1 in 10⁶–10⁹ cells) (Capecchi, 1989), presenting enormous challenges for large-scale applications of gene-targeting experiments.

To overcome these challenges, a series of programmable nuclease-based genome editing technologies have been
A series of studies by Haber and Jasin (Rudin et al., 1989; Plessis Precise Genome Editing Programmable Nucleases as Tools for Efficient and

Figure 1. Applications of Genome Engineering
Genetic and epigenetic control of cells with genome engineering technologies is enabling a broad range of applications from basic biology to biotechnology and medicine. (Clockwise from top) Causal genetic mutations or epigenetic variants associated with altered biological function or disease phenotypes can now be rapidly and efficiently recapitulated in animal or cellular models (Animal models, Genetic variation). Manipulating biological circuits could also facilitate the generation of useful synthetic materials, such as algae-derived, silica-based diatoms for oral drug delivery (Materials). Additionally, precise genetic engineering of important agricultural crops could confer resistance to environmental deprivation or pathogenic infection, improving food security while avoiding the introduction of foreign DNA (Food). Sustainable and cost-effective biofuels are attractive sources for renewable energy, which could be achieved by creating efficient metabolic pathways for ethanol production in algae or corn (Fuel). Direct in vivo correction of genetic or epigenetic defects in somatic tissue would be permanent genetic solutions that address the root cause of genetically encoded disorders (Gene surgery). Finally, engineering cells to optimize high yield generation of drug precursors in bacterial factories could significantly reduce the cost and accessibility of useful therapeutics (Drug development).

Programmable Nucleases as Tools for Efficient and Precise Genome Editing
A series of studies by Haber and Jasin (Rudin et al., 1989; Plessis et al., 1992; Rouet et al., 1994; Choulika et al., 1995; Bibikova et al., 2001; Bibikova et al., 2003) led to the realization that targeted DNA double-strand breaks (DSBs) could greatly stimulate genome editing through HR-mediated recombination events. Subsequently, Carroll and Chandrasegaran demonstrated the potential of designer nucleases based on zinc finger proteins for efficient, locus-specific HR (Bibikova et al., 2001, 2003). Moreover, it was shown in the absence of an exogenous homology repair template that localized DSBs can induce insertions or deletion mutations (indels) via the error-prone nonhomologous end-joining (NHEJ) repair pathway (Figure 2A) (Bibikova et al., 2002). These early genome editing studies established DSB-induced HR and NHEJ as powerful pathways for the versatile and precise modification of eukaryotic genomes.

To achieve effective genome editing via introduction of site-specific DNA DSBs, four major classes of customizable DNA-binding proteins have been engineered so far: meganucleases derived from microbial mobile genetic elements (Smith et al., 2006), zinc finger (ZF) nucleases based on eukaryotic transcription factors (Urnov et al., 2005; Miller et al., 2007), transcription activator-like effectors (TALEs) from Xanthomonas bacteria (Christian et al., 2010; Miller et al., 2011; Boch et al., 2009; Moscou and Bogdanove, 2009), and most recently the RNA-guided DNA endonuclease Cas9 from the type II bacterial adaptive immune system CRISPR (Cong et al., 2013; Mali et al., 2013a).

Meganuclease, ZF, and TALE proteins all recognize specific DNA sequences through protein-DNA interactions. Although meganucleases integrate its nuclease and DNA-binding domains, ZF and TALE proteins consist of individual modules targeting 3 or 1 nucleotides (nt) of DNA, respectively (Figure 2B). ZFs and TALEs can be assembled in desired combinations and attached to the nuclease domain of FokI to direct nucleolytic activity toward specific genomic loci. Each of these platforms, however, has unique limitations.

Meganuclease have not been widely adopted as a genome engineering platform due to lack of clear correspondence between meganuclease protein residues and their target DNA sequence specificity. ZF domains, on the other hand, exhibit context-dependent binding preference due to crosstalk between adjacent modules when assembled into a larger array (Maeder et al., 2008). Although multiple strategies have been developed to account for these limitations (Gonzalez et al., 2010; Sander et al., 2011), assembly of functional ZFPs with the desired DNA binding specificity remains a major challenge that requires an extensive screening process. Similarly, although TALE DNA-binding monomers are for the most part modular, they can still suffer from context-dependent specificity (Juillerat et al., 2014), and their repetitive sequences render construction of novel TALE arrays labor intensive and costly.

Given the challenges associated with engineering of modular DNA-binding proteins, new modes of recognition would significantly simplify the development of custom nucleases. The CRISPR nuclease Cas9 is targeted by a short guide RNA that recognizes the target DNA via Watson-Crick base pairing (Figure 2C). The guide sequence within these CRISPR RNAs typically corresponds to phage sequences, constituting the natural mechanism for CRISPR antiviral defense, but can be easily replaced by a sequence of interest to retarget the Cas9 nuclease. Multiplexed targeting by Cas9 can now be achieved at unprecedented scale by introducing a battery of short guide
elements now known as CRISPR (Figure 3), which are found underpinning the biological function of the mysterious repetitive editing draws upon more than a decade of basic research into CRISPR-Cas9: From Yogurt to Genome Editing

The recent development of the Cas9 endonuclease for genome editing has opened up a broad range of biological applications across basic research to biotechnology and medicine.

The utility of customizable DNA-binding domains extends far beyond genome editing with site-specific endonucleases. Fusing them to modular, sequence-agnostic functional effector domains allows flexible recruitment of desired perturbations, such as transcriptional activation, to a locus of interest (Xu and Bestor, 1997; Beerli et al., 2000a; Konermann et al., 2013; Maeder et al., 2013a; Mendenhall et al., 2013). In fact, any modular enzymatic component can, in principle, be substituted, allowing facile additions to the genome engineering toolbox. Integration of genome- and epigenome-modifying enzymes with inducible protein regulation further allows precise temporal control of dynamic processes (Beerli et al., 2000b; Konermann et al., 2013).

**CRISPR-Cas9: From Yogurt to Genome Editing**

The recent development of the Cas9 endonuclease for genome editing draws upon more than a decade of basic research into understanding the biological function of the mysterious repetitive elements now known as CRISPR (Figure 3), which are found throughout the bacterial and archaeal diversity. CRISPR loci typically consist of a clustered set of CRISPR-associated (Cas) genes and the signature CRISPR array—a series of repeat sequences (direct repeats) interspaced by variable sequences (spacers) corresponding to sequences within foreign genetic elements (protospacers) (Figure 4). Whereas Cas genes are transcribed into proteins, most CRISPR arrays are first transcribed as a single RNA before subsequent processing into shorter CRISPR RNAs (crRNAs), which direct the nuclease activity of certain Cas enzymes to degrade target nucleic acids.

The CRISPR story began in 1987. While studying the iap enzyme involved in isozyme conversion of alkaline phosphatase in *E. coli*, Nakata and colleagues reported a curious set of 29 nt repeats downstream of the *iap* gene (Ishino et al., 1987). Unlike most repetitive elements, which typically take the form of a form of tandem repeats like TALE repeat monomers, these 29 nt repeats were interspaced by five intervening 32 nt nonrepetitive sequences.

Over the next 10 years, as more microbial genomes were sequenced, additional repeat elements were reported from genomes of different bacterial and archaeal strains. Mojica and colleagues eventually classified interspaced repeat sequences as a unique family of clustered repeat elements present in >40% of sequenced bacteria and 90% of archaea (Mojica et al., 2000).

These early findings began to stimulate interest in such microbially repeat elements. By 2002, Jansen and Mojica coined the acronym CRISPR to unify the description of microbial genomic loci consisting of an interspaced repeat array (Jansen et al., 2002; Barrangou and van der Oost, 2013). At the same time, several clusters of signature CRISPR-associated (cas) genes were identified to be well conserved and typically adjacent to the repeat elements (Jansen et al., 2002), serving as a basis for the eventual classification of three different types of CRISPR systems (types I–III) (Haft et al., 2005; Makarova et al., 2011b). Types I and III CRISPR loci contain multiple Cas proteins, now known to form complexes with crRNA (CASCADE complex for type I; Crm or Csm RAMP complexes for type III) to facilitate the recognition and destruction of target nucleic acids (Grouns...
et al., 2008; Hale et al., 2009) (Figure 4). In contrast, the type II system has a significantly reduced number of Cas proteins. However, despite increasingly detailed mapping and annotation of CRISPR loci across many microbial species, their biological significance remained elusive.

A key turning point came in 2005, when systematic analysis of the spacer sequences separating the individual direct repeats suggested their extrachromosomal and phage-associated origins (Mojica et al., 2005; Pourcel et al., 2005; Bolotin et al., 2005). This insight was tremendously exciting, especially given previous studies showing that CRISPR loci are transcribed (Tang et al., 2002) and that viruses are unable to infect archaeal cells carrying spacers corresponding to their own genomes (Mojica et al., 2005). Together, these findings led to the speculation that CRISPR arrays serve as an immune memory and defense mechanism, and individual spacers facilitate defense against bacteriophage infection by exploiting Watson-Crick base-pairing between nucleic acids (Mojica et al., 2005; Pourcel et al., 2005). Despite these compelling realizations that CRISPR loci might be involved in microbial immunity, the specific mechanism of how the spacers act to mediate viral defense remained a challenging puzzle. Several hypotheses were raised, including thoughts that CRISPR spacers act as small RNA guides to degrade viral transcripts in a RNAi-like mechanism (Makarova et al., 2006) or that CRISPR spacers direct Cas enzymes to cleave viral DNA at spacer-matching regions (Bolotin et al., 2005).

Working with the dairy production bacterial strain *Streptococcus thermophilus* at the food ingredient company Danisco, Horvath and colleagues uncovered the first experimental evidence for the natural role of a type II CRISPR system as an adaptive immunity system, demonstrating a nucleic-acid-based immune system in which CRISPR spacers dictate target specificity while Cas enzymes control spacer acquisition and phage defense (Barrangou et al., 2007). A rapid series of studies illuminating the mechanisms of CRISPR defense followed shortly and helped to establish the mechanism as well as function of all three types of CRISPR loci in adaptive immunity. By studying the type I CRISPR locus of *Escherichia coli*, van der Oost and colleagues showed that CRISPR arrays are transcribed and converted into small crRNAs containing individual spacers to guide Cas nuclease activity (Brouns et al., 2008). In the same year, CRISPR-mediated defense by a type III-A CRISPR system from *Staphylococcus epidermidis* was demonstrated to block plasmid conjugation, establishing the target of Cas enzyme activity as DNA rather than RNA (Marraffini and Sontheimer, 2005).

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>First report of CRISPR clustered repeats Ishino et al.</td>
</tr>
<tr>
<td>2000</td>
<td>Recognition that CRISPR families are present throughout prokaryotes Mojica et al.</td>
</tr>
<tr>
<td>2002</td>
<td>Coined “CRISPR” name, defined signature Cas genes Jansen et al.</td>
</tr>
<tr>
<td>2007</td>
<td>First experimental evidence for CRISPR adaptive immunity Barrangou et al.</td>
</tr>
<tr>
<td>2008</td>
<td>CRISPR acts upon DNA targets Marrasini et al. Spacers are converted into mature crRNAs that act as small guide RNAs Brouns et al.</td>
</tr>
<tr>
<td>2009</td>
<td>Type III-B Cmr CRISPR complexes cleave RNA Hale et al.</td>
</tr>
<tr>
<td>2010</td>
<td>Cas9 is guided by spacer sequencess and deavies target DNA via DSBs Gameau et al.</td>
</tr>
<tr>
<td>2011</td>
<td>tracrRNA forms a duplex structure with crRNA in association with Cas9 Delcheva et al. Type II CRISPR systems are modular and can be heterologously expressed in other organisms Sapranaskas et al.</td>
</tr>
<tr>
<td>2012</td>
<td>In vitro characterization of DNA targeting by Cas9 Jinek et al. Gasminas et al.</td>
</tr>
<tr>
<td>2013</td>
<td>First demonstration of Cas9 genome engineering in eukaryotic cells Cong et al. Mali et al.</td>
</tr>
</tbody>
</table>

Figure 3. Key Studies Characterizing and Engineering CRISPR Systems Cas9 has also been referred to as Cas5, Csx12, and Cas1 in literature prior to 2012. For clarity, we exclusively adopt the Cas9 nomenclature throughout this Review. CRISPR, clustered regularly interspaced short palindromic repeats; Cas, CRISPR-associated; crRNA, CRISPR RNA; DSB, double-strand break; tracrRNA, trans-activating CRISPR RNA.
target degradation. In type III CRISPR, crRNAs associate either with Csm or Cmr complexes that bind and cleave DNA and RNA substrates, respectively. In contrast, the type II system requires only the Cas9 nuclease to degrade DNA matching its dual guide RNA consisting of a crRNA-tracrRNA hybrid.

As the pace of CRISPR research accelerated, researchers quickly unraveled many details of each type of CRISPR system (Figure 4). Building on an earlier speculation that protospacer-adjacent motifs (PAMs) may direct the type II Cas9 nuclease to cleave DNA (Bolotin et al., 2005), Moineau and colleagues highlighted the importance of PAM sequences by demonstrating that PAM mutations in phage genomes circumvented CRISPR interference (Deveau et al., 2008). Additionally, for types I and II, the lack of PAM within the direct repeat sequence within the CRISPR array prevents self-targeting by the CRISPR system. In type III systems, however, mismatches between the 5’ end of the crRNA and the DNA target are required for plasmid interference (Marraffini and Sontheimer, 2010).

By 2010, just 3 years after the first experimental evidence for CRISPR in bacterial immunity, the basic function and mechanisms of CRISPR systems were becoming clear. A variety of groups had begun to harness the natural CRISPR system for various biotechnological applications, including the generation of phage-resistant dairy cultures (Quiberoni et al., 2010) and phylogenetic classification of bacterial strains (Horvath et al., 2008, 2009). However, genome editing applications had not yet been explored.

Around this time, two studies characterizing the functional mechanisms of the native type II CRISPR system elucidated the basic components that proved vital for engineering a simple RNA-programmable DNA endonuclease for genome editing. First, Moineau and colleagues used genetic studies in Streptococcus thermophilus to reveal that Cas9 (formerly called Cas5, Csn1, or Csx12) is the only enzyme within the cas gene cluster that mediates target DNA cleavage (Garneau et al., 2010). Next, Charpentier and colleagues revealed a key component in the biogenesis and processing of crRNA in type II CRISPR systems—a noncoding trans-activating crRNA (tracrRNA) that hybridizes with crRNA to facilitate RNA-guided targeting of Cas9 (Deltcheva et al., 2011). This dual RNA hybrid, together with Cas9 and endogenous RNase III, is required for processing the CRISPR array transcript into mature crRNAs (Deltcheva et al., 2011). These two studies suggested that there are at least three components (Cas9, the mature crRNA, and tracrRNA) that are essential for reconstituting the type II CRISPR nuclease system. Given the increasing importance of programmable site-specific nucleases based on ZFs and TALEs for enhancing eukaryotic genome editing, it was tantalizing to think that perhaps Cas9 could be developed into an RNA-guided genome editing system. From this point, the race to harness Cas9 for genome editing was on.
In 2011, Siksnys and colleagues first demonstrated that the type II CRISPR system is transferrable, in that transplantation of the type II CRISPR locus from *Streptococcus thermophilus* into *Escherichia coli* is able to reconstitute CRISPR interference in a different bacterial strain (Saparansauskas et al., 2011). By 2012, biochemical characterizations by the groups of Charpentier, Doudna, and Siksnys showed that purified Cas9 from *Streptococcus thermophilus* or *Streptococcus pyogenes* can be guided by crRNAs to cleave target DNA in vitro (Jinek et al., 2012; Gasiunas et al., 2012), in agreement with previous bacterial studies (Garneau et al., 2010; Deltcheva et al., 2011; Saparansauskas et al., 2011). Furthermore, a single guide RNA (sgRNA) can be constructed by fusing a crRNA containing the targeting guide sequence to a tracrRNA that facilitates DNA cleavage by Cas9 in vitro (Jinek et al., 2012).

In 2013, a pair of studies simultaneously showed how to successfully engineer type II CRISPR systems from *Streptococcus thermophilus* (Cong et al., 2013) and *Streptococcus pyogenes* (Cong et al., 2013; Mali et al., 2013a) to accomplish genome editing in mammalian cells. Heterologous expression of mature crRNA-tracrRNA hybrids (Cong et al., 2013) as well as sgRNAs (Cong et al., 2013; Mali et al., 2013a) directs Cas9 cleavage within the mammalian cellular genome to stimulate NHEJ or HDR-mediated genome editing. Multiple guide RNAs can also be used to target several genes at once. Since these initial studies, Cas9 has been used by thousands of laboratories for genome editing applications in a variety of experimental model systems (Sander and Joung, 2014). The rapid adoption of the Cas9 technology was also greatly accelerated through a combination of open-source distributors such as Addgene, as well as a number of online user forums such as http://www.genome-engineering.org and http://www.egenome.org.

**Structural Organization and Domain Architecture of Cas9**

The family of Cas9 proteins is characterized by two signature nuclease domains, RuvC and HNH, each named based on homology to known nuclease domain structures (Figure 2C). Though HNH is a single nuclease domain, the full RuvC domain is divided into three subdomains across the linear protein sequence, with RuvC I near the N-terminal region of Cas9 and RuvC II/II flanking the HNH domain near the middle of the protein. Recently, a pair of structural studies shed light on the structural mechanism of RNA-guided DNA cleavage by Cas9.

First, single-particle EM reconstructions of the *Streptococcus pyogenes* Cas9 (SpCas9) revealed a large structural rearrangement between apo-Cas9 unbound to nucleic acid and Cas9 in complex with crRNA and tracrRNA, forming a central channel to accommodate the RNA-DNA heteroduplex (Jinek et al., 2014). Second, a high-resolution structure of SpCas9 in complex with sgRNA and the complementary strand of target DNA further revealed the domain organization to comprise of an α-helical recognition (REC) lobe and a nuclease (NUC) lobe consisting of the HNH domain, assembled RuvC subdomains, and a PAM-interacting (PI) C-terminal region (Nishimasu et al., 2014) (Figure 2A and Movie S1).

Together, these two studies support the model that SpCas9 unbound to target DNA or guide RNA exhibits an autoinhibited conformation in which the HNH domain active site is blocked by the RuvC domain and is positioned away from the REC lobe (Jinek et al., 2014). Binding of the RNA-DNA heteroduplex would additionally be sterically inhibited by the orientation of the C-terminal domain. As a result, apo-Cas9 likely cannot bind nor cleave target DNA. Like many ribonucleoprotein complexes, the guide RNA serves as a scaffold around which Cas9 can fold and organize its various domains (Nishimasu et al., 2014).

The crystal structure of SpCas9 in complex with an sgRNA and target DNA also revealed how the REC lobe facilitates target binding. An arginine-rich bridge helix (BH) within the REC lobe is responsible for contacting the 9–12 nt of the RNA-DNA heteroduplex (Nishimasu et al., 2014), which correspond with the seed sequence identified through guide sequence mutation experiments (Jinek et al., 2012; Cong et al., 2013; Fu et al., 2013; Hsu et al., 2013; Pattanayak et al., 2013; Mali et al., 2013b).

The SpCas9 structure also provides a useful scaffold for engineering or refactoring of Cas9 and sgRNA. Because the REC2 domain of SpCas9 is poorly conserved in shorter orthologs, domain recombination or truncation is a promising approach for minimizing Cas9 size. SpCas9 mutants lacking REC2 retain roughly 50% of wild-type cleavage activity, which could be partly attributed to their weaker expression levels (Nishimasu et al., 2014). Introducing combinations of orthologous domain recombination, truncation, and peptide linkers could facilitate the generation of a suite of Cas9 mutant variants optimized for different parameters such as DNA binding, DNA cleavage, or overall protein size.

**Metagenomic, Structural, and Functional Diversity of Cas9**

Cas9 is exclusively associated with the type II CRISPR locus and serves as the signature type II gene. Based on the diversity of associated Cas genes, type II CRISPR loci are further subdivided into three subtypes (IIA–IIC) (Figure 5B) (Makarova et al., 2011a; Chylinski et al., 2013). Type II CRISPR loci mostly consist of the cas9, cas1, and cas2 genes, as well as a CRISPR array and tracrRNA. Type IIC CRISPR systems contain only this minimal set of cas genes, whereas types IIA and IIB have an additional signature cas2 or cas4 gene, respectively (Chylinski et al., 2013).

Subtype classification of type II CRISPR loci is based on the architecture and organization of each CRISPR locus. For example, type IIA and IIB loci usually consist of four cas genes, whereas type IIC loci only contain three cas genes. However, this classification does not reflect the structural diversity of Cas9 proteins, which exhibit sequence homology and length variability irrespective of the subtype classification of their parental CRISPR locus. Of >1,000 Cas9 nucleases identified from sequence databases (UniProt) based on homology, protein length is rather heterogeneous, roughly ranging from 900 to 1600 amino acids (Figure 5C). The length distribution of most Cas9 proteins can be divided into two populations centered around 1,100 and 1,350 amino acids in length. It is worth noting that a third population of large Cas9 proteins belonging to subtype IIA, formerly called Csx12, typically contain around 1,500 amino acids.

Despite the apparent diversity of protein length, all Cas9 proteins share similar domain architecture (Makarova et al., 2011a;