CHARACTERIZING NGAGO AND EXPLORING ITS ACTIVITIES FOR BIOTECHNOLOGICAL APPLICATIONS

by

Kok Zhi Lee

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Agricultural and Biological Engineering
West Lafayette, Indiana
May 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Kevin V. Solomon, Chair
Department of Agricultural and Biological Engineering

Dr. Michael R. Ladisch
Department of Agricultural and Biological Engineering

Dr. Jenna L. Rickus
Department of Agricultural and Biological Engineering

Dr. Frederick S. Gimble
Department of Biochemistry

Dr. Kari L. Clase
Department of Agricultural and Biological Engineering

Approved by:
Dr. Nathan Mosier

Dedicated to my family who loves me unconditionally and all of my friends who have supported me

ACKNOWLEDGMENTS

This work was possible through the financial support of the College of Engineering, College of Agriculture, Ralph W. & Grace M. Showalter Research Trust, and Air Force Institute of Technology.

I would like to thank Dr. Solomon for providing a flexible training environment to explore my ideas to start new projects.

Next, I would like to thank my committees, including Dr. Ladisch, Dr. Gimble, Dr. Rickus, and Dr. Clase, for their effort and time in advising my project and career.

There are million moments in my graduate school experience that I wouldn’t trade. Thank you for being part of my grad school experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

BLAST:	Basic Local Alignment Search Tool
CRISPR:	clustered regularly interspaced short palindromic repeats
DSB:	DNA double-stranded break
E.coli:	*Escherichia coli*
FW strand:	forward strand
MID:	middle domain
pAgos:	prokaryotic argonautes
PAM:	protospacer adjacent motif
IPTG:	Isopropyl β- d-1-thiogalactopyranoside
MjAgo:	*Methanocaldococcus jannaschii* argonaute
MpAgo:	*Marinitoga piezophila* argonaute
N:	N-terminal domain
NgAgo:	pAgo from *Natronobacterium gregoryi* argonaute
NHEJ:	non-homologous end-joining
OB:	oligonucleotide/oligosaccharide-binding
PAZ:	PIWI-Argonaute-Zwille domain
PfAgo:	*Pyrococcus furiosus* argonaute
PIWI:	P element-induced wimpy testis domain
P-ssDNA:	phosphorylated guide ssDNA
rNgAgo:	refolded NgAgo
RV strand:	reverse strand
sNgAgo:	soluble NgAgo
tGreen:	truncated mNeonGreen
tKanR:	truncated Kanamycin resistant gene
TtAgo:	*Thermus thermophilus* argonaute
TXTL:	transcription-translation

ABSTRACT

Prokaryotic Argonautes (pAgos) have been proposed as more flexible tools for gene-editing as they do not require sequence motifs adjacent to their targets for function. One promising pAgo candidate from the halophilic archaeon *Natronobacterium gregoryi* (NgAgo) has been the subject of intense debate regarding its potential in eukaryotic systems. NgAgo was initially claimed to edit genes in mammalian cells, but the report was retracted due to replication failure. Due to low solubility, subsequent studies refolded NgAgo and suggested that it cuts RNA but not DNA; however, mutation of the conserved active site does not abolish cleavage activity, raising the possibility of nuclease contamination. Another independent study demonstrated gene-editing via NgAgo in bacteria. These inconsistent results underscore the knowledge gap and roadblock for NgAgo-based gene-editing tool development.

In this work, I revisit this enzyme and characterize its function *in vitro* and in a bacterial system. The halophilic features of NgAgo have been neglected in the literature, leading to inconclusive results. Like other halophilic proteins, NgAgo has modified amino acid composition, leading to failure of domain identification/function prediction via sequence alignment. Indeed, using more sensitive structural alignments, I identified a new single-stranded DNA binding domain, repA, in NgAgo and other halophilic pAgos. Due to its halophilic nature, NgAgo expresses poorly in low-salt environments, with the majority of protein being insoluble and inactive even after refolding. However, soluble NgAgo indeed cuts DNA. NgAgo DNA-cleaving activity can only be abolished via mutation in the canonical PIWI domain and repA deletion, revealing a new catalytic behavior in pAgos. Moreover, NgAgo requires both repA and PIWI domains to create double-stranded DNA breaks, leading to cell death or enhancing homologous recombination, or gene-editing, at a modest level in bacteria. Rational protein engineering of NgAgo was also pursued to increase solubility. Although three out of seven mutants showed significant increases in solubility, they lost the ability to cleave DNA in *E.coli*. Structural modeling revealed some subtle but important differences in the protein structures, explaining why the mutants lose their function. Besides, a selection system for improving endonuclease activity was optimized for future pAgo optimization. Collectively, this work revealed that NgAgo possesses unique catalytic behavior in the pAgo family and has some gene-editing application potential. More importantly, this work

expands knowledge of the pAgo family, providing a foundation for future pAgo-based gene-editing tool development.

CHAPTER 1. INTRODUCTION

1.1 Background and motivation

Gene-editing tools revolutionize biotechnology by enabling *in vivo* modification of many gene loci across many species that may cure genetic diseases[1], increase crop yields[2], and optimize bioprocess[3], among other applications. For example, gene editing was used to treat muscular dystrophy in a dog model by rescuing expression of aberrantly-low mutant dystrophin by 3-90% in various tissue types[4]. Another study also showed that gene-editing can also be achieved in germ cells, which generate offspring without a gene mutation[1]. This is exciting as edits that treat mutant alleles can be inherited by offspring, reducing transmission of genetic diseases. Gene-editing may also be used in plants. For instance, scientists have mutated two genes, including Gn1a and GS3, in rice to enhance grain number and size, respectively, increasing overall yields[5]. Other examples of modifying the traits include resistance to viral infection[6] and environmental stress[2]. These examples either improve the crop yields directly or provide protection against natural disasters that decrease crop yield. Gene-editing also creates technologies that allow scientists to rapidly modify multiple loci simultaneously to optimize the production of valuable chemicals with high efficiency, accelerating the bioprocess optimization process[7]. These applications can be attributed to the advance of gene-editing tools.

Clustered regularly interspaced short palindromic repeats (CRISPR)-derived tools are currently the most popular gene-editing tool because they are easy to use in many species across kingdoms[8–12]. CRISPR relies on a Cas enzyme and single guide RNA (sgRNA) to edit DNA[9,10]. Users can express a guide sequence complementary to the target sequence to direct guide-bound Cas enzyme for DNA cleavage. Upon DNA cleavage, introduced DNA with desired modifications serves as a repair template to edit the genome via homologous recombination[13]. The final product incorporates the desired modifications within the genome to edit the gene. CRISPR/Cas systems require a sequence-specific motif adjacent to their target for function; the recognition of the sequence-specific motif (protospacer adjacent motif, PAM) by Cas-guide complex initiates hybridization of guide RNA and target DNA, and subsequent DNA cleavage[14]. Despite the versatility of the CRISPR/Cas system, editing GC-biased genomic regions or species is challenging[9,10]. For example, the most popular CRISPR/Cas9 needs three nucleotides, “5’NGG” (N: A, T, C, or G), to induce DNA cleavage[14]. For genomic regions or species that have few “5’NGG” motif, gene-editing becomes challenging. For example, soybeans on average have a Cas9 PAM site every 20 nucleotides due to their biased genomic GC content[15]. As the distance between the motif and the target modification site increases to 35-base-pair, gene-editing efficiency decreases to ~20%[16]. To resolve this motif-restriction issue, scientists have developed different CRISPR/Cas systems such as Cas12a (Cpf1) that require different sequence-specific motifs. Cpf1 requires “5’TTN” (N: A, T, C, or G) to increase the coverage of the genome of interest[17]. Alternatively, scientists have engineered the CRISPR/Cas9 system so it can bind to expanded motifs. For example, xCas9 can bind to NG, GAA, and GAT motifs[1]. Nonetheless, CRISPR/Cas systems still require motifs to edit DNA. As different CRISPR/Cas systems have different guide design rules, editing efficiency[19], specificity[20], and off-target activity[21], finding the right tool with high performance is complicated and challenging, underscoring a need for other programmable endonucleases without motif restrictions.

Prokaryotic Argonautes (pAgos) have been proposed as more flexible tools because they have been shown to cut DNA without motifs requirement *in vitro* [22,23]. Guide DNA or RNA directs pAgos to the DNA or RNA region complementary to the guide for single-stranded cleavage of the nucleic acids (nicking) by the pAgo[23]. To induce a double-stranded DNA break, two pAgo molecules with two different guides are needed[23]. Several pAgos that cut DNA have been characterized. However, the majority of characterized pAgos only function at high temperatures (>55°C), restricting their use in species living at mesophilic temperatures relevant to biotechnology[24].

The first identified mesophilic pAgo, NgAgo, from halophilic (salt-loving) *Natronobacterium gregoryi* was claimed to edit DNA in mammalian cells at temperatures relevant to biotechnology. However, this claim was later retracted[25]. Recently, *in vitro* studies suggested that refolded NgAgo cleaves RNA, not DNA, but the inactivation of the active site of NgAgo is still capable of cleavage[26]. On the other hand, another study demonstrated the gene-editing capability of NgAgo in prokaryotes[27]. The lack of loss-of-function mutant and conflicting conclusions from eukaryotes and prokaryotes indicate knowledge gaps in our understanding of NgAgo. These knowledge gaps limit the development of NgAgo as it dictates how we can repurpose NgAgo to manipulate DNA in mesophilic temperatures.

1.2 Scope and Objectives

In order to overcome the knowledge gaps that restrict the use of mesophilic NgAgo for gene-editing in mesophilic temperatures relevant to biotechnology, this dissertation aims to dissect the function of NgAgo and leverage the knowledge to develop NgAgo as a gene-editing tool. Specifically, the work in this dissertation will:

- Analyze NgAgo via more sensitive structure alignment tools
- Optimize NgAgo production for *in vitro* function characterization
- Demonstrate DNA-cleaving and gene-editing in bacteria
- Optimize the solubility of NgAgo for improving its activity
- Optimize a selection method for improving endonuclease activity

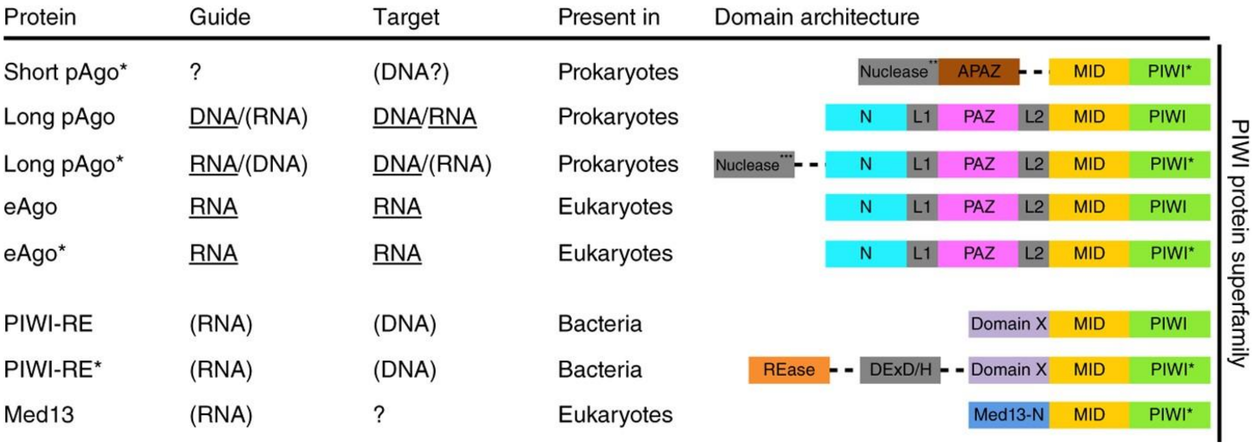
1.3 Thesis organization

This dissertation is structured around developing strategies to realize the potential gene-editing activity of NgAgo in mesophilic temperatures relevant to biotechnology. Chapter 2 lays out and analyzes the strategies for pAgo characterization. Chapter 3 focuses on the functional characterization of NgAgo. Here, more sensitive structure alignment tools were used to analyze domains of NgAgo before characterization. I then developed a cell-free platform to produce NgAgo, NgAgo variants with domain deletion and point mutations were tested for DNA cleavage activity. The NgAgo variants were also tested for their DNA cleavage and gene-editing activities in bacteria. Chapter 4 focuses on overcoming the low solubility of halophilic NgAgo via rational protein engineering design. Chapter 5 discusses optimization of a current selection system for evolving active endonucleases. Finally, Chapter 6 concludes the works from chapter 3-4 and lays out the impact and future directions.

CHAPTER 2. LITERATURE REVIEW

2.1 Prokaryotic Argonautes as programmable endonucleases for flexible gene-editing tool development

P element–induced wimpy testis (PIWI) superfamily proteins contain diverse protein families with additional domains in different kingdoms (Figures 2-1 and 2-2). PIWI superfamily proteins have been either predicted or shown to use DNA or RNA as guides to interfere with target DNA or RNA via binding or cleavage, depending on the presence of an intact catalytic tetrad, DEDX, in the PIWI domain or additional nuclease domain (Figure 2-1). Besides eukaryotic Argonautes (eAgos), prokaryotic Argonautes (pAgos), including long pAgos, short pAgos, and PIWI-RE, are the most well-characterized family[28] (Figures 2-1 and 2-2). Among pAgos, long pAgos are programmable endonucleases recently proposed as flexible tools for genome editing[24] without motifs requirement[22,23]. Like CRISPR/Cas9-based gene editing strategies, single-stranded nucleic acids bind to pAgos and enhance pAgo cleavage of complementary target nucleic acids sequences, enabling DNA repair and editing. However, pAgos have the distinct advantage of not requiring a protospacer adjacent motif (PAM) for function[22,23,29], which means that pAgos are not limited to targets flanked by PAM sites and can potentially cut any DNA target regardless of composition. Despite this potential, no pAgo has been developed that rivals the simplicity and function of Cas9-based strategies



Protein	Guide	Target	Present in	Domain architecture
Short pAgo*	?	(DNA?)	Prokaryotes	Nuclease** APAZ - - MID PIWI*
Long pAgo	DNA/(RNA)	DNA/RNA	Prokaryotes	N L1 PAZ L2 MID PIWI
Long pAgo*	RNA/(DNA)	DNA/(RNA)	Prokaryotes	Nuclease*** - - N L1 PAZ L2 MID PIWI*
eAgo	RNA	RNA	Eukaryotes	N L1 PAZ L2 MID PIWI
eAgo*	RNA	RNA	Eukaryotes	N L1 PAZ L2 MID PIWI*
PIWI-RE	(RNA)	(DNA)	Bacteria	Domain X MID PIWI
PIWI-RE*	(RNA)	(DNA)	Bacteria	REase - - DExD/H - - Domain X MID PIWI*
Med13	(RNA)	?	Eukaryotes	Med13-N MID PIWI*

Figure 2-1. Classification of PIWI superfamily proteins. PIWI superfamily contains short pAgo, long pAgo, eukaryotic argonaute (eAgo), PIWI-RE, and Med13. *: Ago proteins without an intact DEDX catalytic tetrad in the PIWI domain. Guide and target nucleic acid annotations are based on available biochemical characterization (underlined) or prediction (in parentheses). **: predicted nucleases from Sir2, Mrr, or TIR protein families. ***: predicted nucleases from Sir2, Mrr, Cas4, or PLD protein families. REase: restriction endonuclease; DExD/H: superfamily II helicase (denoted after a signature amino acid motif). Dotted lines indicate separate genes located in the same (predicted) operon. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Structural & Molecular Biology (2014)[28].

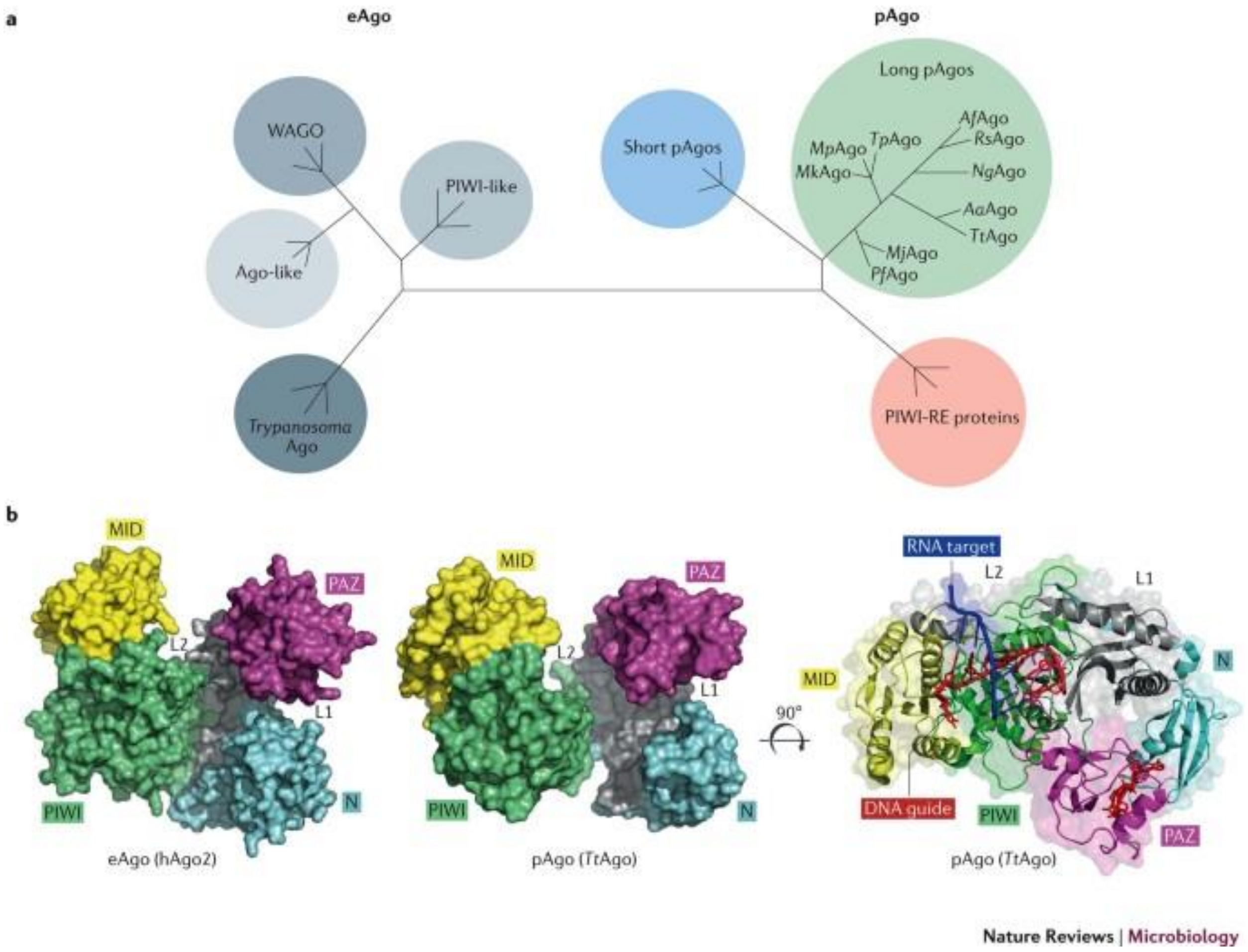


Figure 2-2. a, Phylogenetic analysis based on the amino acid alignments of conserved MID and PIWI domains from eukaryotic Argonaute proteins (eAgos) and prokaryotic Argonaute proteins (pAgos). aAgo, *Aquifex aeolicus* Argonaute; AfAgo, *Archaeoglobus fulgidus* Argonaute; MjAgo, *Methanocaldococcus jannaschii* Argonaute; MkAgo, *Methanopyrus kandleri* Argonaute; MpAgo, *Marinitoga piezophila* Argonaute; NgAgo, *Natronobacterium gregoryi* Argonaute; PfAgo, *Pyrococcus furiosus* Argonaute; RsAgo, *Rhodobacter sphaeroides* Argonaute; TpAgo, *Thermotoga profunda* Argonaute; WAGO, worm-specific Argonaute; PIWI-RE, PIWI domain-containing protein with conserved R and E residues. b, Crystal structures of eAgo (human Ago2, hAgo2. PDB: 4Z4C) and pAgo (*Thermus thermophilus* Argonaute, TtAgo. PDB: 3F73). Ternary complex of TtAgo (rotated 90°) with a DNA guide (red) bound to an RNA target (blue). Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Microbiology (2018)[24].

2.1.1 Current available prokaryotic Argonautes for programmable cleavage and their limitations

Long pAgos are predicted to serve as a form of adaptive defense mechanism against invading nucleic acids such as phage/viral DNA and RNA[28,30] (Figure 2-3). The cleavage activities of pAgos may be guide-independent or guide-dependent. Guide-independent cleavage or "DNA chopping" was first identified in TtAgo and is believed to be the mechanism by which pAgos generate guides for targeted or guide-dependent cleavage[31] (Figure 2-3). The chopped nucleic acids are then bound as a guide, converting the apo-pAgo to a targeted and more active form for guide-dependent cleavage of complementary target DNA, RNA, or both via the conserved catalytic tetrad, DEDX[28] (Figure 2-3). To create a double-stranded DNA break, long pAgos require two guides. Target recognition and cleavage are enabled by four canonical domains[22]: N (N-terminal), PAZ (PIWI-Argonaute-Zwille), MID (middle), and PIWI (P element-induced wimpy testis) (Figure 2-2). The N-terminal domain is essential in target cleavage[32,33] and dissociation of cleaved strands[33,34], though the detailed mechanism remains poorly understood. The MID domain interacts with the 5'-end of the guide[35] and promotes binding of the guide to its target nucleic acids[36]. The PAZ domain interacts with the 3' end of a guide[37–40], protecting it from degradation. The PIWI domain plays a pivotal role in nucleic acid cleavage via the conserved catalytic tetrad, DEDX (D: aspartate, E: glutamate, X: histidine, aspartate, or asparagine)[28].

Although pAgos are characterized as individual endonucleases, several studies indicate accessory proteins also play a vital role in pAgo-mediated DNA function. Genomic organization analyses of pAgos-containing species revealed other proteins such as helicase and single-stranded DNA binding protein are organized in an operon, suggesting a concerted role in the hosts[28]. Moreover, supplementation of helicase and single-stranded DNA binding protein enhance DNA cleavage activity of TtAgo, even at the high GC-content DNA regions[41]. This is conceivable as the known domains of pAgos do not have helicase activity, so it is reasonable to recruit other proteins to unwind DNA before DNA cleavage. Thus, the need for accessory proteins must be taken into consideration for pAgo-mediated gene-editing tool development. Despite the demonstration of programmable DNA cleavage of many long pAgos *in vitro*, currently characterized pAgos including TtAgo[23], MpAgo[42], PfAgo[29], and

MjAgo[22] work at very high temperatures (>55 °C), making them infeasible for gene editing in common mesophilic organisms relevant to biotechnology.

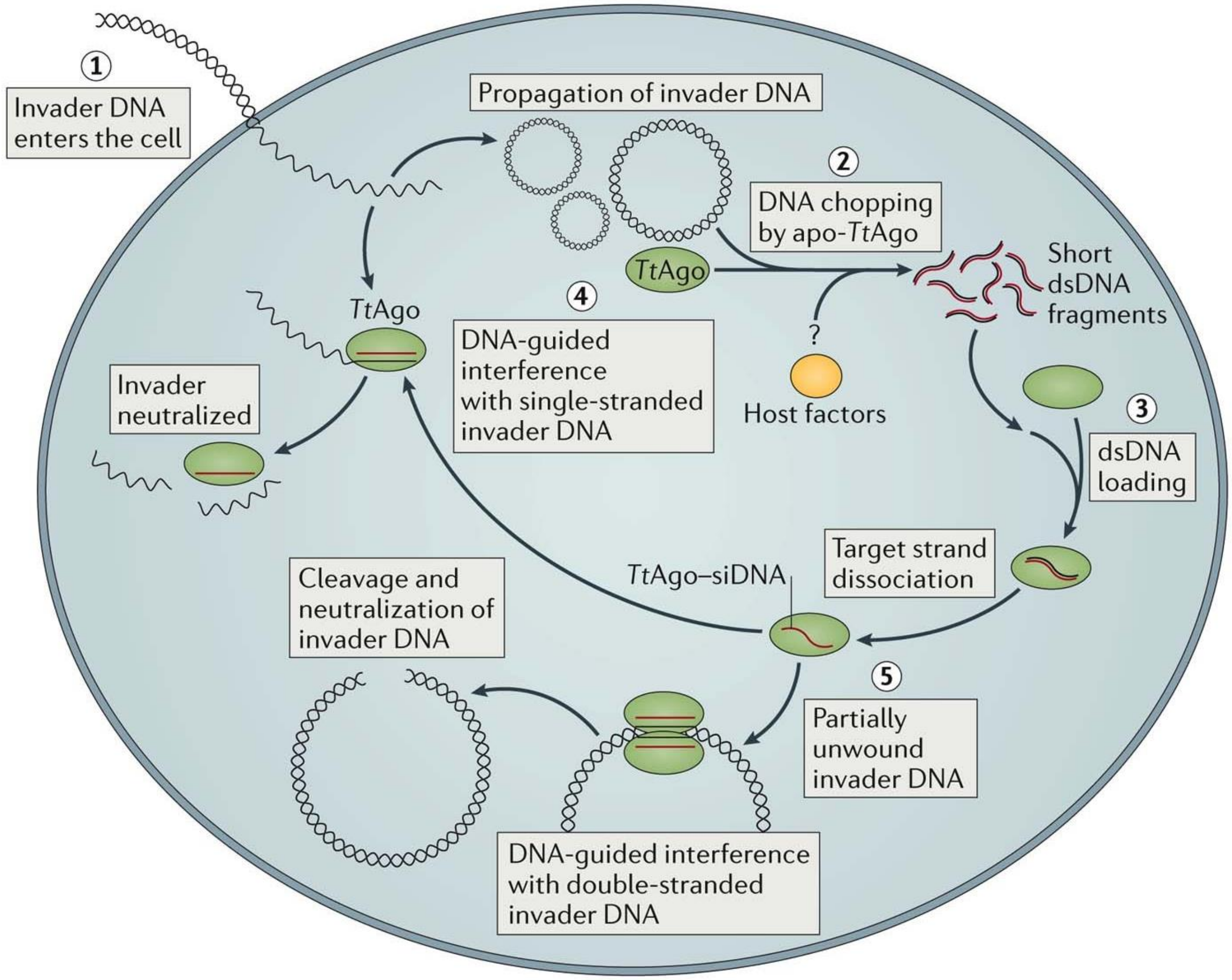

1
Invader DNA enters the cell
Propagation of invader DNA
2
DNA chopping by apo-TtAgo
TtAgo
Short dsDNA fragments
?
Host factors
3
dsDNA loading
Target strand dissociation
TtAgo
4
DNA-guided interference with single-stranded invader DNA
Invader neutralized
TtAgo–siDNA
Cleavage and neutralization of invader DNA
5
Partially unwound invader DNA
DNA-guided interference with double-stranded invader DNA

Figure 2-3. Prokaryotic argonaute (TtAgo) serves as a defense mechanism against invading nucleic acids. Step 1: Phage invades its single-stranded DNA via infection and propagates as double-stranded DNA (dsDNA). Step 2: Guide-free TtAgo (apo-TtAgo) chops DNA into small pieces of dsDNA, possibly stimulated by host factors. Step 3: Small fragments of dsDNA are loaded to guide-free TtAgo. Step 4: After releasing one strand of the dsDNA, the TtAgo-guide complex interferes with single-stranded invader DNA via cleavage. Step 5: TtAgo-guide complex can also interfere partially unwinding dsDNA via cutting with two TtAgo-guide complexes. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Microbiology[24].

2.1.2 Mesophilic NgAgo and its unresolved issues

The halophilic Argonaute from *Natronobacterium gregoryi* (NgAgo) was recently put forth as a promising candidate for pAgo-mediated gene editing as it is believed to operate at mesophilic temperatures (~37°C)[25]. However, these claims have since been refuted due to an inability to demonstrate *in vitro* DNA cleavage or replicate these findings in many eukaryotic hosts[43–47]. NgAgo expression is poor, presumably due to its halophilic characteristics that make low salt expression challenging[48,49]. Thus, all published *in vitro* cleavage assays have relied on refolded protein[26,41], which may be non-functional, resulting in inconclusive results[50].

Halophilic proteins differ in their amino acid composition compared to analogous proteins in non-halophilic hosts to maintain protein structure and enable function in high-salt environments[48]. Typically, smaller hydrophobic amino acids such as valine (V) and alanine (A) are used (Figure 2-4), and more negatively charged amino acids such as aspartate (D) and glutamate (E) are on the protein surface[48] (Figures 2-4 and 2-5), resulting in low identity/similarity of halophilic protein compared to non-halophilic protein homologs. Due to this reason, identifying domains of halophilic protein for the function is challenging. Sequence alignment of NgAgo with other catalytically active pAgos showed low identity/similarity (less than 30%), which is lower than the commonly accepted threshold (30%) for inferring conserved function in homologs.

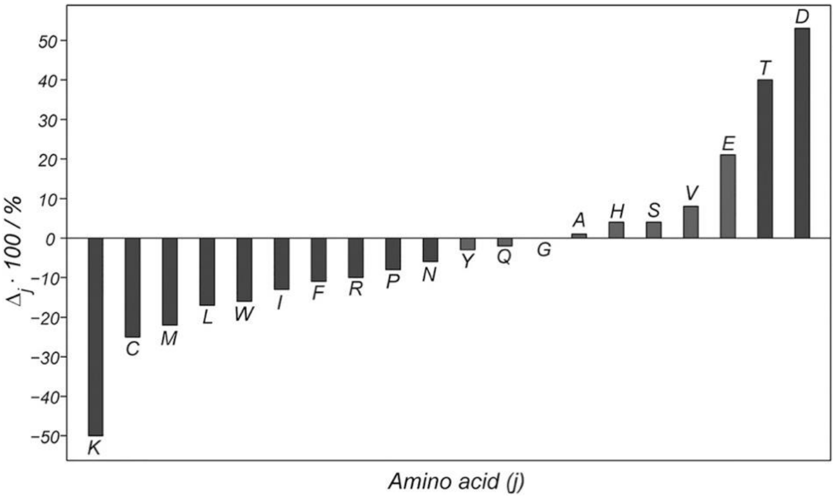
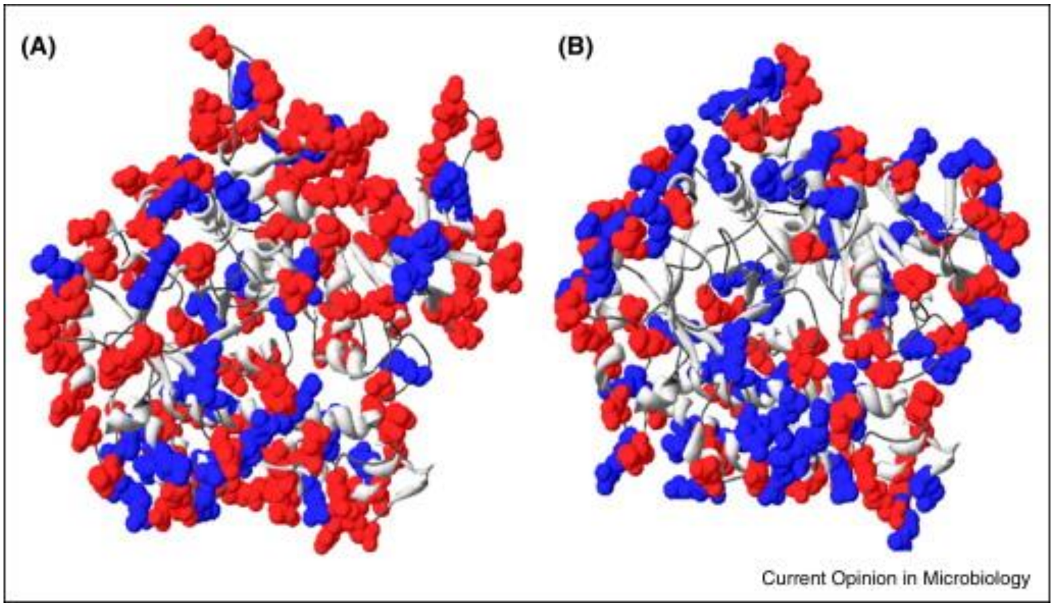


Figure 2-4. Amino acid occurrence in halophilic proteins compared to their mesophilic orthologs. A positive value ($\Delta_j > 0$) and a negative value ($\Delta_j > 0$) indicates relative enrichment and depletion of amino acid *j* in halophilic proteins, respectively). Reprinted from Chemistry & Biology, Volume 22/Issue 12, Ortega et al., Halophilic Protein Adaptation Results from Synergistic Residue-Ion Interactions in the Folded and Unfolded States, Pages 1597-1607, Copyright (2015), with permission from Elsevier.



Current Opinion in Microbiology

Figure 2-5. Surface of halophilic protein has more negatively-charged amino acids—beta-galactosidases from halophile *Halorubrum lacusprofundi* (A) and thermophile *Thermus thermophilus* (B) used as models. The surface negatively-charged and positively-charged residues of halophilic protein and thermophilic proteins are labeled in red and blue, respectively. The net surface charges are -65 (A) and -4 (B). Reprinted from Current Opinion in Microbiology, Volume 25, Shiladitya DasSarma and Priya DasSarma, Halophiles and their enzymes: negativity put to good use, Pages 120-126, Copyright (2015), with permission from Elsevier.

Moreover, negatively-charged residues on the protein surface make halophilic protein challenging to express in organisms living in low-salt environments[49] (Figure 2-5). The expressed halophilic proteins often fail to fold correctly into functional soluble proteins, forming insoluble protein aggregates instead. Scientists can refold the halophilic proteins to recover the functional activities. However, refolding may take few days, and the final yield of functional protein might be very low[50] .

The halophilic nature of NgAgo has been neglected in the literature, which might be the key to understand the knowledge gaps of NgAgo. First of all, domain analysis via sequence alignment might have missed some critical information. NgAgo (887 amino acids) is longer than all characterized pAgos (TtAgo: 685 amino acids[39]; MjAgo: 713 amino acids[51]; PfAgo: 770 amino acids; MpAgo: 639 amino acids), and the additional ~100 amino acids might contain another domain. This putative domain might have been missed during sequence alignment due to low identity/similarity.

Second, similar to other halophilic proteins, NgAgo has very low solubility when expressing in low-salt environments such as in protein expression host, *E.coli*. Scientists have refolded NgAgo to test its function *in vitro*[26,41]. Although refolded NgAgo cuts RNA but not DNA[26], mutation of putative catalytic tetrad does not abolish cleavage activity. It is doubtful whether the cleavage activity is due to NgAgo or other endonuclease contamination. While another independent study demonstrated the gene-editing ability of NgAgo in bacteria, these conflicting results remain unclear, emphasizing the need to reanalyze and characterize NgAgo with caution.

2.2 Current methods for pAgo characterization

Researchers use various tools and methods to characterize the activities of pAgos (Figure 2-6). These tools and methods include sequence/structural alignments, protein expression/purification/*in vitro* characterization, and function validation in hosts. I will discuss their fundamental principle, advantages, and limitations in the following sections, summarized in table 2-1.

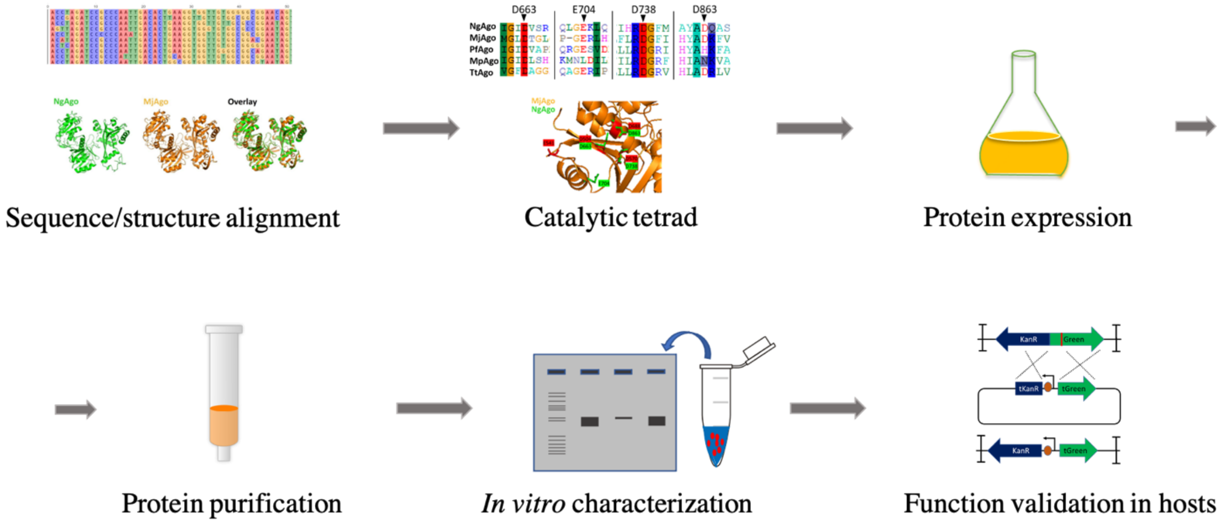


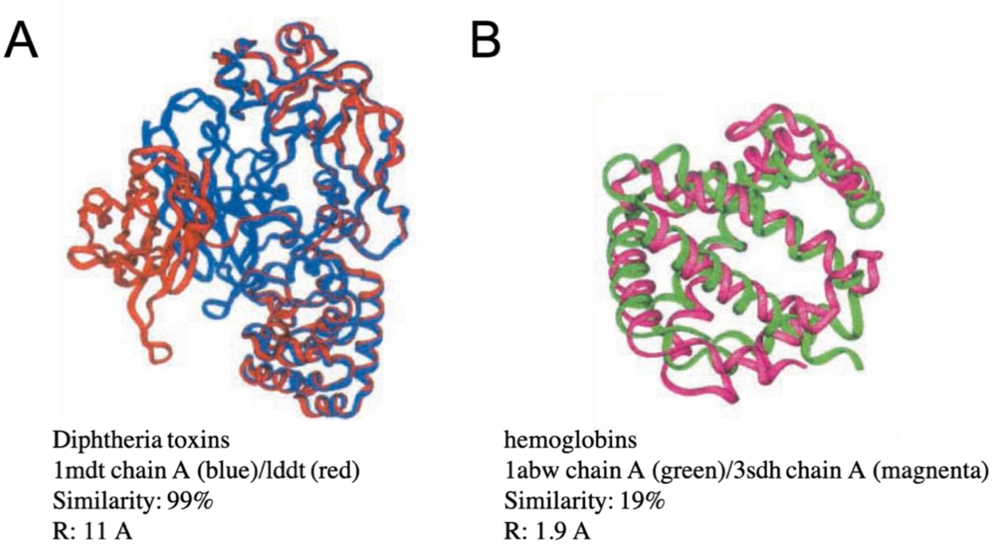
Figure 2-6. Overview of pAgo characterization. Sequence alignment or structure alignment is generally used to predict the homologs, informing the functions of pAgo of interest. Catalytic tetrad, DEDX, is then checked if it is intact, which determines nuclease activity. Interested pAgo was cloned and expressed in a host such as bacteria. The bacterial host was lysed for subsequent protein purification. Purified pAgo of interest can be carried in vitro characterization for assessing pAgo activity. Lastly, pAgo activity such as gene-editing ability can be evaluated in a host of interest.

2.2.1 Sequence and structural alignments

Researchers usually use the protein sequence of interested pAgo to do a sequence alignment against the protein database using BLAST[52] to identify homologs as their first step. The alignment identifies homologs based on the identity and similarity (acidic, basic, aliphatic, etc.) of the protein sequence. As homologs evolved from the same origin likely to carry out a similar function, sequence alignment is one of the most informative analyses to infer the function of interested pAgo.

The excess similarity of homologs is supported by the statistical estimate, E-values, with $<10^{-6}$ as a threshold for high similarity[53]. However, researchers are more comfortable using 30% of identity as a rule of thumb for identifying homologs[53]. This rule, in fact, underestimates the number of homologs by at least 33%[53]. As the missed homologs might be more distant homologs with modified functions, using 30% identity might miss critical information for comprehensive understanding. Indeed, diphtheria toxins have very high sequence similarity but low structural similarity; conversely, hemoglobins have low sequence identity and high structural similarity to carry out the same function, as illustrated in Figure 2-7.

The major advantage of using sequence alignment is that it gives us the most valuable information regarding the pAgo domains and predicted functions in a relatively short time, usually less than twenty minutes[53]. As domains dictate the protein function, researchers can infer the function of pAgos depending on the known characterized pAgos in a relatively short time and less effort investment. However, when similarity and E-values are low or at the boundary of threshold, the identified domains and homolog become less reliable[53]. This is when alternative approaches should be considered.



Diphtheria toxins
1mdt chain A (blue)/lddt (red)
Similarity: 99%
R: 11 A

hemoglobins
1abw chain A (green)/3sdh chain A (magenta)
Similarity: 19%
R: 1.9 A

Figure 2-7. Examples of diverse sequence/structure relationships. Superpositions of selected protein pairs diphtheria toxins (A) and hemoglobin (B) displayed. A, The diphtheria toxins have high sequence identity (99%) but have very different protein structures, with root-mean-square =11 angstrom. B, The hemoglobin pair has low sequence identity but with high structure similarity, with root-mean-square =1.9 angstrom. Reprinted from Biophysical Journal, Volume 83, Issue 5, Gan et al., Analysis of Protein Sequence/Structure Similarity Relationships, Pages 2781-2791, Copyright (2002), with permission from Elsevier.

An alternative approach is to use more sensitive structure prediction/alignment methods, such as Phyre2[54]. Phyre2 consists of four stages, including gathering homologous sequences, fold library scanning, loop modeling, and side-chain placement[54] (Figure 2-8). Gathering homologous sequences looks for homologs and uses them to predict secondary structures with PSIPRED[55]. Both alignment and secondary structure prediction integrate into a query hidden Markov model[56], which describes the evolution of observable events that depend on non-observable internal factors (hidden). In the next stage, fold library scanning, the query hidden Markov model would be scanned against the hidden Markov model of proteins with known structure. The top-scoring alignments are used to construct the backbone of the predicted structure. Then, loop modeling corrects the insertion and deletions in these models. Lastly, the amino acids are added to the final predicted structure. A major difference between Phyre2 and sequence alignment is that Phyre2 includes structure information. As seen in many homologous proteins, proteins with low identities can still form similar structures, having similar functions (Figure 2-7). This feature of Phyre2 overcomes the low identity/E-value issue of the sequence alignment. As this method includes the structure information, structure alignment-based homolog prediction is more reliable, especially when the protein of interest has low identity and similarity[57]. However, as structure alignment integrates structural information, which consumes higher computational costs to analyze the data, it takes a much longer time to predict homologs (0.5 to 2 hours). Moreover, the predicted structure and identified homolog are likely to be less accurate if the structure of the protein of interest has a very different structure (Figure2-7A) since the process integrates structural information.

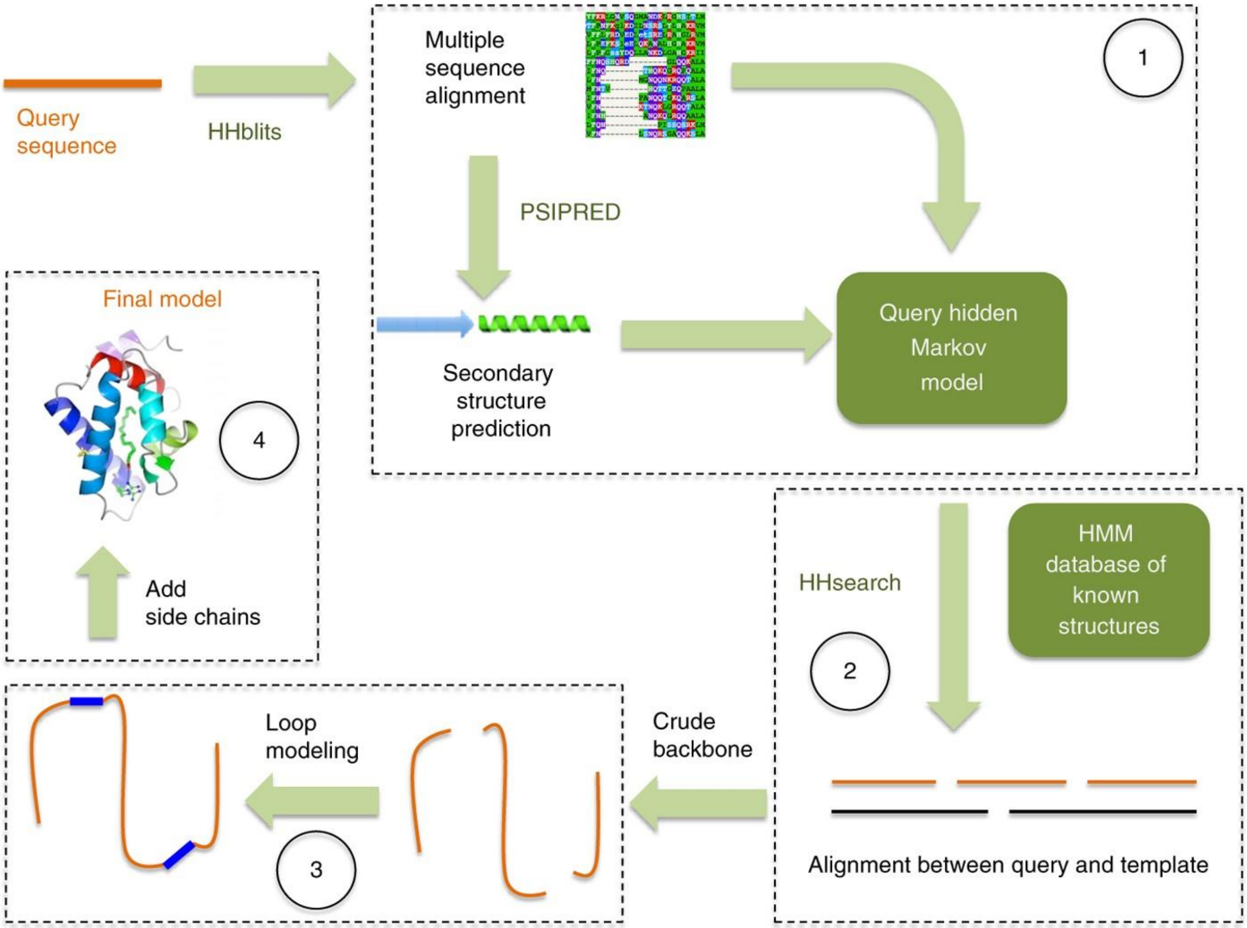

Query sequence
HHblits
Multiple sequence alignment
1
PSIPRED
Secondary structure prediction
Query hidden Markov model
Final model
4
Add side chains
HHsearch
HMM database of known structures
2
Crude backbone
Alignment between query and template
Loop modeling
3

Figure 2-8. Pipelines for structural alignment of Phyre 2. Stage 1 (Gathering homologous sequences): a query sequence is aligned against the curated nr20 (no sequences with >20% mutual sequence identity) protein sequence database with heuristic HHblits. The sequence alignment is then used to predict secondary structure. Both sequence alignment and predicted secondary structure are integrated into a query hidden Markov model. Stage 2 (fold library scanning): The query hidden Markov model is scanned against known hidden Markov model of known structure in the database. The top-scoring alignments are then used to construct the backbone of the predicted structure. Stage 3 (Loop modeling): insertion and deletions of the models are corrected via loop modeling. Stage 4 (Side-chain placement): amino acid side chains are added to complete the final predicted structure. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Protocols (2015)[54].

After getting the alignment, researchers also check if the catalytic tetrad[28], DEDX, of pAgo is intact. This would inform whether the candidate pAgo is capable of cleaving nucleic acids. Catalytic tetrad identification can usually be done with sequence alignment due to the strict requirement of the catalytic site for function. This approach is advantageous as it informs the researcher whether pAgo can cut nucleic acids. However, it does not tell us what type of guide pAgo binds or what type of nucleic acids pAgo cuts. These bioinformatic predictions, however, must be verified by experiment.

2.2.2 Protein expression, purification, and *in vitro* characterization

After sequence or structure alignment, the interested pAgo would be cloned with a purification tag and expressed in bacterial hosts such as *E. coli*[22,23,29]. The pAgo would then be purified through an affinity column and tested for purity by running SDS-PAGE. Purified pAgo with high purity would then move on to *in vitro* characterization. Standard *in vitro* characterization includes guide preference (DNA, RNA, or both), target nucleic acid preference (DNA, RNA, or both), guide 5' preference (phosphorylated or hydroxylated), and cleavage activity[23,26,29,42].

Guides can either be DNA, RNA, or both, depending on the nature of pAgos[28]. Researchers usually identify copurified guides when expressing them in bacterial host[23]. By treating with DNase or RNase, we can know the identity of the copurified guides[22,23]. Researchers can also sequence the guides to know the length and the GC content of the guides[23]. As pAgos tend to degrade partially unwound double-stranded DNA, low GC double-stranded DNA are more susceptible to degradation by pAgos, generating guides with low GC content[31]. Nonetheless, the characterization of copurified guides reveals the guide preference of pAgos, length of the guides, and GC content of the guides.

Once identified, researchers would test what nucleic acids does the candidate pAgo cleave. The guides can be either 5' phosphorylated or hydroxylated[23,42]. Researchers usually incubate candidate pAgo with guides complementary to target single-stranded DNA or RNA in appropriate buffer conditions and see what type of nucleic acids can be cleaved by candidate pAgo. If the candidate pAgo cleaves DNA, researchers will test if it can cleave double-stranded DNA. They usually select regions with different GC-content as some pAgos have a preference for low-GC regions[23,58,59].

In vitro characterization tells us the guide preference, 5' guide preference, and target nucleic acids in very conclusive ways. Researchers typically include catalytic tetrad mutant as a negative control to show that the cleavage activity is due to pAgo itself, instead of DNase contamination. This information can then compare with the predicted function of pAgo (e.g., Nucleic acid cleavage depends on catalytic tetrad.). Despite the advantages, *in vitro* characterization also has its limitations.

Although pAgo with high purity is a standard for *in vitro* characterization, not all pAgos can achieve that, at least for NgAgo due to low solubility[60]. Insoluble proteins can be refolded as an alternative to recover functional protein. However, refolding protein may take few days, and the final product may still be non-functional[50]. Thus, protein expression and purification are bottlenecks for proteins that are insoluble or difficult to purify. Furthermore, studies have shown that pAgos have physically associated with other proteins such as helicase, recA, and single-stranded binding protein, suggesting the concerted function of accessory protein in hosts[27,41,61]. The presence of the accessory proteins is important for characterization as they can enhance DNA cleavage, even with high GC-content DNA[41]. These evidence points out the limitation of purified protein for functional validation. If the tested pAgo does not show any cleavage activity, researchers should consider whether the protein is functional or needs other accessory proteins for its activity. Lastly, *in vitro* activity does not always translate to *in vivo* activity as *in vivo* is more complex than the environment in the test tube.

2.2.3 Function validation in hosts

Since *in vitro* activity does not always translate to *in vivo* activity, pAgo function should be validated in hosts. Although the DNA cleavage activity of pAgos has not been demonstrated in hosts, several CRISPR studies provide precedence to demonstrate DNA cleavage activity and gene-editing activity in hosts. To show the cleaving activity is also present in hosts, researchers use survival assays and gene-editing assays to validate the function of the interested endonuclease. As DNA break in bacterial is lethal, creating a DNA break by endonuclease such as CRISPR/Cas9 would result in death[62]. Thus, this reduced survival outcome can serve as a method, survival assay, to validate DNA cleavage activity in bacteria.

Gene-editing can be achieved by DNA repair via homologous recombination after DNA cleavage induced by an endonuclease, known as gene-editing assay[63,64]. Researchers express endonuclease and test if it can enhance homologous recombination with a DNA template and lambda-red recombinase after DNA cleavage[63]. The recombinant would then contain a modified DNA. Typically, the target gene for editing is a nonfunctional antibiotic-resistant gene[63]. The host becomes antibiotic-resistant after gene-editing. The number of antibiotic-resistant colonies can then be calculated for the gene-editing efficiency of interested endonuclease[63,64].

Conclusions of DNA cleavage and gene-editing derived from the host are very encouraging as they show activity in a more complex environment compared to a test tube, which can be viewed as a successful gene-editing application in the host. Despite the advantages, there are challenges for pAgos validation. Like CRISPRs, pAgos need to have guides to execute its on-target function. Several pAgos have been shown to have “DNA chopping” abilities[31], which cuts DNA non-specifically when there is no guide. As there is no way to constantly make targeted single-stranded DNA or RNA guides in hosts currently, guides must be transformed into the hosts. As guides can be subjected to degradation by host endonuclease and dilution via cell division, the on-target activity can only be detected before the guides are lost, making function validation in hosts challenging.

Second, the outcome of cell death or gene-editing can also be promoted by other mechanisms rather than DNA cleavage. Researchers need to cross-validate the conclusions from the *in vitro* data as well. If the DNA does not cut by pAgo *in vitro*, there is a possibility that other mechanisms may promote cell death and gene-editing. Researchers need to be cautious when drawing a conclusion from function validation in hosts.

Table 2-1. Comparison of methods for pAgo characterization

Methods	Advantages	Disadvantages and limitations	Purpose/Role
Sequence alignment	• most informative analysis • Quick: <20 minutes	• less accurate when the identity/similarity of pAgo of interest is low • needs experimental verification	• identifies pAgo homologs • suggests cleavage activity based on catalytic tetrad, DEDX
Structure alignment	• more accurate than sequence alignment	• Slow: 20 minutes – 2 hours • needs experimental verifications	• identifies pAgo homologs • suggests cleavage activity based on catalytic tetrad, DEDX
Protein expression/ purification	• well-established protein expression genetic tools and hosts	• Time-consuming (2-5 days) • It may lack accessory proteins that enhance pAgo activity • Some proteins are difficult to purify due to solubility	• Not applicable
Protein refolding	• Recovery of proteins that are challenging to purify	• The refolded protein might not be functional • It is time-consuming (hours to days)	• Not applicable
Cell-free system protein production	• It can be done in a short time (1-2 days) • expression conditions (temperature, salt concentrations, additive, etc.) can be easily modified • High-throughput expression	• It may lack accessory proteins that enhance pAgo activity	• Not applicable

Table 2-1. continued

***In vitro* characterization**	• **conditions (temperature, salt concentrations, additive, etc.) can be easily modified** • **generates very conclusive results**	• **The results from *in vitro* characterization may not work in a host as the conditions might not be ideal**	**verifies:** • **cleavage activity** • **target nucleic preference** • **guide preference** • **5' guide preference** • **GC content preference of guide**
Function validation in hosts	• host testing	• short validation window due to no guide production in hosts • availability of unwound double-stranded DNA may determine the outcome of the validation • Other mechanisms other than DNA cleavage may also promote gene-editing	• validates DNA cleavage and gene-editing activities in hosts

2.3 Future directions for pAgo characterization and development for gene-editing

Guide preference, 5' guide preference, and target nucleic acid preference can only be known through *in vitro* characterization currently. It will be valuable to develop a prediction tool based on artificial intelligence. Indeed, an artificial intelligence-based method, AlphaFold, has shown critical progress in the recent Critical Assessment of Protein Structure Prediction (CASP) competition[65]. AlphaFold2 showed that two-thirds of the predicted protein structures were within experimental error. In other words, AlphaFold2 can predict protein structures as good as the ones obtained from experiments. The accurate prediction of the protein structure can facilitate the selection of pAgos of interest with desirable properties including guide preference, 5' guide preference, and target nucleic acid preference, accelerating the development of pAgo-based gene-editing tools. As the currently available pAgo structures are limited, it is challenging to develop accurate algorithms to predict the pAgos properties as artificial intelligence relies on training based on feeding a large number of accurate inputs as pointed up in ligand binding prediction[66].

Nonetheless, AlphaFold2 or other artificial intelligence-based methods will be valuable for pAgos prediction, accelerating the pAgos characterization and development for gene-editing tool developments.

For pAgo proteins that are difficult to express and purify, researchers have refolded pAgos for subsequent *in vitro* analysis[26,41]. However, refolding proteins is challenging as they require proper conditions and sometimes take days to refold. The refolded protein may also not be functional[67]. Therefore, there is a critical need to find an alternative method that circumvents host protein expression and purification. An alternative solution is to use a bacterial cell-free system composed of bacterial lysate containing all transcription and translation machinery for protein production[68]. The bacterial cell-free system is made by lysis of bacteria and supplemented with an energy mix, amino acid mix, cofactors, ions, and DNA that encodes the protein of interest. The commercialized bacterial cell-free lysate is also a convenient option as researchers will only need to supplement with DNA encoding protein of interest. The mixture is incubated at 29-37 °C for 16-20 hours for protein production. As the protein expression is carried out in a test tube, we can modify the conditions to increase the soluble protein fraction. Moreover, as a cell-free system can produce protein in a minimal volume (10-30 microliters), it can prototype the activity of many pAgos or other endonucleases in a high throughput manner[69].

For function validation in hosts, as there is no way to produce short and programmable DNA or RNA guides, we can only detect the on-target activity in a short time-window before the guide degradation or dilution through cell division. Although there are some *in vivo* ssDNA production methods[70,71], the produced ssDNA contain additional nucleotides that interfere with pAgo function, limiting the programmability of the guide. Therefore, methods for making short and programmable DNA or RNA guides are needed to facilitate function validation of pAgos in hosts.

Last but not least, additional proteins such as helicase and single-stranded DNA binding protein may be needed for pAgo-mediated gene-editing in eukaryotic systems as pAgos do not have a helicase domain. To date, there is no gene-editing evidence with pAgos in eukaryotic systems. This may be attributed to the slower doubling time of eukaryotic systems, which means a shorter single-stranded DNA exposure time window. Indeed, a supplement of helicase and single-stranded

DNA binding protein improve TtAgo activity *in vitro*, even with high GC-content[41]. Additional support is that helicase and single-stranded DNA binding protein are often organized in an operon with pAgos, suggesting a concerted function of helicase and single-stranded DNA binding protein[28,72]. Thus, systematic screening for functional helicase and single-stranded DNA binding proteins may assist pAgo-mediated gene-editing in eukaryotic systems.

2.4 Conclusion

Gene-editing has impacted many fields attributed to the advance of gene-editing tools. The scientific community has been searching for new programmable endonucleases for gene-editing tool development as current gene-editing tools, CRISPRs, relies on sequence-specific motif to edit genes. So far, prokaryotic Argonautes have been put forward as a next-generation gene-editor without motif restriction. Despite the presence of challenges, we anticipated artificial intelligence and synthetic biology communities would bring in or develop new technologies and methodologies to accelerate the development of pAgo-based gene-editing tools.

CHAPTER 3. CHARACTERIZATION OF NGAGO ACTIVITIES

This chapter is published as preprint (Lee et al, NgAgo DNA endonuclease activity enhances homologous recombination in E. coli, biorxiv, doi:10.1101/597237, 2020) and is undergoing peer review. The numbering of the figures, tables, and references have been modified to fit the format of this thesis.

3.1 Abstract

Prokaryotic Argonautes (pAgos) have been proposed as more flexible tools for gene-editing as they do not require sequence motifs adjacent to their targets for function, unlike popular CRISPR/Cas systems. One promising pAgo candidate, from the halophilic archaeon *Natronobacterium gregoryi* (NgAgo), however, has been the subject of intense debate regarding its potential in eukaryotic systems. Here, we revisit this enzyme and characterize its function in prokaryotes. NgAgo expresses poorly in non-halophilic hosts, with the majority of protein being insoluble and inactive even after refolding. However, we report that the soluble fraction does indeed act as a DNA endonuclease. Structural homology modeling revealed that NgAgo shares canonical domains with other catalytically active pAgos and contains a previously unrecognized single-stranded DNA binding domain (repA). Both repA and the canonical PIWI domains participate in the DNA cleavage activities of NgAgo. We showed that NgAgo can be programmed with guides to cleave specific DNA *in vitro* and in *E.coli*. We also found that these endonuclease activities are essential for enhanced NgAgo-guided homologous recombination, or gene-editing, in *E. coli*. Collectively, our results demonstrate the potential of NgAgo for gene-editing and reconcile seemingly contradictory reports.

3.2 Introduction

Long prokaryotic Argonaute proteins (pAgos) are programmable endonucleases that have recently been proposed as flexible tools for genome editing[24]. These enzymes bind single-stranded DNA and/or RNA molecules as guides, which then prime the enzyme for nicking of complementary target DNA, RNA, or both[24]. Double stranded DNA cleavage requires two complementary guides, which may induce DNA repair and editing and form an alternative gene editing platform to standard CRISPR-based tools. Unlike Cas9-based gene editing strategies, however, pAgos have the distinct advantage of not requiring a protospacer adjacent motif (PAM) for function[22,23,42,73]. Thus, pAgos are not limited to targets flanked by PAM sites and can potentially cut any DNA target regardless of composition. Despite this potential, no pAgo has been developed that rivals the simplicity and function of Cas9-based strategies.

Target recognition and cleavage is enabled by four canonical domains[22]: N (N-terminal), PAZ (PIWI-Argonaute-Zwille), MID (middle), and PIWI (P element-induced wimpy testis) domains. The N-terminal domain is essential for target cleavage[32,33] and dissociation of cleaved strands[33,34], although the detailed mechanism remains poorly understood. The MID domain interacts with the 5'-end of the guide[35] and promotes binding to its target[36]. The PAZ domain interacts with the 3' end of the guide[37–40], protecting it from degradation[74]. Finally, the PIWI domain plays a pivotal role in nucleic acid cleavage via the conserved catalytic tetrad, DEDX (D: aspartate, E: glutamate, X: histidine, aspartate or asparagine)[28].

Recent emerging evidence also suggests a role for accessory proteins in pAgo activity. Within prokaryote genomes, pAgos are often organized in operons with ssDNA binding proteins and helicases among other DNA modifying proteins[72] hinting at concerted function *in vivo*. Supplementing a pAgo with these proteins *in vitro* enhances reaction rates and target specificity, reduces biases in substrate composition preferences, and enables activity on more topologically diverse substrates[41]. These effects are observed with several homologs of these accessory proteins for multiple pAgos. Moreover, pAgos also copurify with helicases, ssDNA binding proteins, and recombinases from both native and heterologous hosts[27,61] indicating conserved physical interactions in different prokaryotes. Given the need for these and potentially other unrecognized accessory proteins, *in vivo* evaluation of pAgos may more accurately reflect their activity.

Despite the potential for programmable cleavage activities by long pAgos, currently characterized pAgos including TtAgo[23], MpAgo[42], PfAgo[29] and MjAgo[22,75] work at very high temperatures (>55 °C), making them infeasible for gene editing and *in vivo* testing in common mesophilic organisms. The halophilic Argonaute from the archaeon *Natronobacterium gregoryi* (NgAgo) was recently put forth as a promising candidate for pAgo-mediated gene editing, as it was believed to be active at mesophilic (~37°C) temperatures[25]. However, these claims have since been refuted due to an inability to demonstrate *in vitro* DNA cleavage or to replicate these findings in a number of eukaryotic hosts [43–47]. NgAgo expression is poor, presumably due to its halophilic characteristics that make low salt expression challenging[48,49]. Thus, all published *in vitro* cleavage assays have relied on refolded protein[26,41], which may be non-functional, resulting in the inconclusive results. Nonetheless, recent work by Fu and colleagues demonstrated that NgAgo may still have potential as a gene editor for prokaryotic hosts[27]. While the authors were able to confirm that gene-editing was mediated by homologous recombination via RecA[27], which physically associated with NgAgo in an unanticipated manner, the specific role of NgAgo remained unclear. Here, we demonstrate that NgAgo is indeed a DNA endonuclease by identifying residues that are required for DNA cleavage, and we provide evidence that this activity is essential for NgAgo-mediated gene editing via homologous recombination repair.

3.3 Material and methods

3.3.1 Strains and plasmids

E. coli strains and plasmids used in this study are listed in Table 3-1. Cloning was carried out according to standard practices[76] with primers, template, and purpose listed in Table 3-2. Plasmids were maintained in *E. coli* DH5α. NgAgo variants (wildtype, D663A/D738A, N-del, and repA with GST or His tag) that were used for *in vitro* activity assays were cloned into an IPTG-inducible T7 plasmid, pET32a-GST-ELP64. MG1655 (DE3) *atpI*::KanR-mNeonGreen was generated using recombineering[77] via donor plasmid pTKDP-KanR-mNeonGreen-hph. For gene-editing/recombination studies[64], p15-KanR-PtetRed was used as a donor plasmid (Table 4-1).

Table 3-1. Strains and Plasmids

Name	**Relevant genotype**	**Vector backbone**	**Plasmid origin**	**Source**
Strains				
BL21 (DE3)	F– ompT gal dcm lon hsdSB(rB– mB–) λ (DE3) [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)			[78]
MG1655 (DE3)	K-12 F– λ– ilvG– rfb-50 rph-1 (DE3)			[79]
MG1655 (DE3) *atpI*::KanR-mNeonGreen	K-12 F– λ– ilvG– rfb-50 rph-1 (DE3) *atpI*::KanR-mNeonGreen			This study
Plasmids				
pBSI-SceI(E/H)	*bla*		ColE1 derivative	[80]
pTXTL-p70a-T7RNAP	*Bla*, P_{70}-T7RNAP		unknown	Arbor Biosciences
pET32a-GST-ELP64	*bla*, lacI, P_{T7}-GST-ELP64		pBR322	Professor Xin Ge (University of California, Riverside)
pTKDP-hph	*bla*, *hph*, sacB		pMB1	[77]
pCas9-CR4	*cat*, P_{Tet}-Cas9		p15A	[81]
pET-GST-Ago-His	*bla*, lacI, P_{T7}-GST-NgAgo-His	pET32a-GST-ELP64	pBR322	This study
pET32a-His-Ago	*bla*, lacI, P_{T7}-GST-NgAgo-His	pET32a-GST-ELP64	pBR322	This study
pET32a-His-repA	*bla*, lacI, P_{T7}-His-repA	pET32a-GST-ELP64	pBR322	This study
pET-GST-N-del-His	*bla*, lacI, P_{T7}-GST-N-del-His	pET32a-GST-ELP64	pBR322	This study
pET-GST-N-del/D663A/D738A-His	*bla*, lacI, P_{T7}-GST-N-del/D663A/D738A -His	pET32a-GST-ELP64	pBR322	This study
pTKDP-KanR-mNeonGreen-hph	*bla*, *hph*, KanR-mNeonGreen	pTKDP-hph	pMB1	This study
p15-KanR-PtetRed	*cat*, KanR-mNeonGreen, P_{Tet}-gam-beta-exo	pCas9-CR4	p15A	This study
pET32-BFP	*Amp*, lacI, P_{T7}-BFP	pET32a-GST-ELP64 and pBAD-mTagBFP2	pBR322	This study
pIncw-mNeonGreen	*cat*	pN565[82] (origin of replication); pCas9-CR4[81] (*cat*)	pIncW	This study

Table 3-2. DNA primers used in this study. Restriction enzyme recognition sites are underlined and indicated in the primer name.

Name	Sequences (5’>3’)	Template	Used to construct
5' NcoI 3xG Ago	**ATCACCATGGGTGGCGGTATGGTGCCAAAAAAGAAGAG**	**Nls-NgAgo-GK**	**pET-GST-Ago-His**
3' XhoI Ago	ATCACTCGAGCTTACTTACATATGGATCCCGG		
NdeI HIS-Ago 5	TATACATATGGGTCACCATCATCATCACCATTCATCGCATCACCATCACCATCACGTGCCAAAAAAGAAGAG	Nls-NgAgo-GK	pET-His-Ago
XhoI rmNdeI Ago 3'	ATATCTCGAGTTACTTACTTACGTATGGATCCCGG	Nls-NgAgo-GK	pET-His-Ago
XhoI STOP repA 3'	CTAACTCGAGTTACTCGACGGTCGTCTGG	Nls-NgAgo-GK	pET-His-repA
D663A 3'	CGGGGTAGCTCCGAGAGACCGCAATCCCAATGAACATATC	pET-His-Ago	NgAgo mutant
D663A 5'	GATATGTTCATTGGGATTGCGGTCTCTCGGAGCTACCCCG	pET-His-Ago	
D738A 5'	CGACCCATATCGTCATCCACCGTGCGGGCTTCATGAACGAAGACCTCGAC	pET-His-Ago	NgAgo mutant
D738A 3'	GTCGAGGTCTTCGTTCATGAAGCCCGCACGGTGGATGACGATATGGGTCG	pET-His-Ago	
XbaI KanR 5'	ATGGTCTAGAATGGGATCGGCCATTG	pTKIP-neo	kanR-mNeonGreen cassette integration
BamHI KanR 3'	ATTTGGATCCTTAGAAGAACTCGTCAAGAAGGC	pTKIP-neo	
XbaI tGreen 5'	CCATTCTAGACCATGGTAGATGGCTCCG	pNCS-mNeonGreen	
XhoI Green uni 3'	TGATCTCGAGAGAGAATATAAAAAGCCAGATTATTAATCCGGCTTTTTTATTATTTTTACTTGTACAGCTCGTCCATGC	pNCS-mNeonGreen	
XbaI tKanR 5'	TAGCTCTAGAGAAAGAGGAGAAATACTAGATGGGATCGGCCATTG	pTKIP-neo	donor plasmid p15-KanR-PtetRed
EcoRI tKanR 3'	ATATGAATTCGATACTTTCTCGGCAGGAGC	pTKIP-neo	

Table 3-3. continued.

Name	Sequences (5’>3’)	Template	Used to construct
XbaI J23100 tGreen 5'	**TTTCTCTAGAGCTAGCACTGTACCTAGGACTGAGCTAGCCGTCAACCATGGGAAGCCACATC**	**pNCS-mNeonGreen**	
XhoI Green uni 3'	TGATCTCGAGAGAGAATATAAAAAGCCAGATTATTAATCCGGCTTTTTTATTATTTTTACTTGTACAGCTCGTCCATGC	pNCS-mNeonGreen	
XhoI pTet 5'	ATCACTCGAGTCCCTATCAGTGATAGAGATTGACATCCCTATCAGTGATAGAGATACTGAGCACTCTAG	pTK-Red	
XhoI DT exo 3'	TGATCTCGAGAAAAAAAAACCCCGCCGAAGCGGGGTTTTTTTTTTCATCGCCATTGCTCC	pTK-Red	

3.3.2 NgAgo expression and purification

GST-NgAgo or His-NgAgo variants were expressed in BL21 (DE3) with 100 µg/ml ampicillin. 5 mL cultures started from single colonies were grown for 16 hours before subculturing in 100 ml of LB Miller containing ampicillin. Expression was induced with 0.1 mM IPTG at OD_{600} = 0.5 for either 4 hours at 37 °C or overnight at 22 °C overnight before harvesting the cells at 7500 rpm (11,500 g) at 4 °C for 5 minutes. The cell pellet was resuspended in TN buffer (10 mM Tris and 100mM NaCl, pH 7.5) and lysed via sonication at a medium power setting (~50 W) in 10 s intervals, with intervening 10 s incubations on ice to reduce heat denaturation. Cell lysates were then clarified at 12000 rpm at 4 °C for 30 minutes. The supernatant was collected as a soluble protein fraction. Both soluble and insoluble (cell pellet) fractions were purified via His-IDA nickel column (Clontech Laboratories, Mountain View, CA. Cat. No: 635657) according to the manufacturer instructions. Insoluble NgAgo protein was refolded on the column after denaturation with guanidium chloride according to manufacturer instructions. GST-tagged NgAgo variants were purified by glutathione agarose (Thermo Fisher Scientific, Waltham, MA. Cat. No: 16100) according to the manufacturer protocol.

3.3.3 Cell-free expression of NgAgo and activity assay

Cell-free TXTL reactions contained 5' phosphorylated DNA guides, Chi6 oligos, IPTG, plasmids encoding T7RNA polymerase (pTXTL-p70a-T7RNAP) and NgAgo variants, including wildtype, D663A/D738A, repA, N-del, and N-del/D663A/D738A (Table 3-3). Reactions were incubated at 29 °C for 20 hours to promote NgAgo expression before being supplemented to 125 mM NaCl and incubating at 37 °C for folding for 24 hours. $MgCl_2$ to a final concentration of 62.5 µM was then added along with target or non-target plasmid for reaction at 37 °C for an hour. RNase A (70 ng or >490 units) (Millipore Sigma, Burlington, MA. Cat. No: R6513-10MG) was then added to each reaction to remove transcribed RNA at 37 °C for 10 minutes. The reaction mixtures were then mixed with 0.5% SDS to dissociate any proteins and 6X loading dye before gel electrophoresis. The gel was visualized under a blue light (Azure Biosystems, Dublin, CA. Azure c400).

Table 3-4. Materials for NgAgo variants production by cell-free system

	Volume (µl)	**Final concentration**	**Remarks**
Cell-free system mixture	4.5	-	
5' phosphorylated DNA guides	0.5	1 µM	
Chi6 oligos	0.5	1 µM	Protect linear DNA from recBCD degradation[83]
IPTG	0.5	0.5 mM	Induce NgAgo variants expression
pTXTL-p70a-T7RNAP	0.5	2.4 nM	Encodes T7RNA polymerase for induction of NgAgo variants
Plasmids encoding NgAgo variants or mNeonGreen control	0.5	6 nM	

3.3.4 Survival assay

BL21 (DE3) was transformed with target plasmid pIncw-mNeonGreen and NgAgo expression plasmid and made electrocompetent. Electrocompetent cells were transformed with either no guides or 1 µg total of FW, RV, both guides and plated on ampicillin and chloramphenicol selective LB Miller agar plate with 0.1 mM IPTG before 16-20 hours incubation at 37 °C. Colonies were counted to measure survival rate of transformants. The unguided control was normalized to 100% and guided-treatments were normalized to the unguided control.

3.3.5 Gene-editing assay

MG1655 (DE3) *atpI*::KanR-mNeonGreen was transformed with pET-GST-NgAgo-His (to induce DNA cleavage) and p15-KanR-PtetRed (for lambda-red recombinase expression and to provide donor DNA for repair) and made electrocompetent. Electrocompetent cells were transformed with either no guides or one 1.2 µl of 100 µM total of FW, RV, both guides and incubated in LB Miller with ampicillin, chloramphenicol, and IPTG for an hour. These cultures were then diluted ten-fold in LB Miller containing ampicillin (working concentration: 100 µg/ml), chloramphenicol (working concentration: 25 µg/ml), IPTG (working concentration: 0.1mM), and anhydrotetracycline (aTc) (working concentration: 50 µg/ml), incubated until OD_{600} = 0.2 before plating with and without kanamycin (working concentration: 50 µg/ml). Colony forming units (CFU) were counted after 16-20 hours incubation at 37 °C. The unguided control was normalized to 100% and guided-treatments were normalized to the unguided control.

3.3.6 Phyre 2 and HHpred analysis

NgAgo protein (IMG/M Gene ID: 2510572918) was analyzed via Phyre 2[54] with normal mode on 2018 November 19. The normal mode pipeline involves detecting sequence homologues, predicting secondary structure and disorder, constructing a hidden Markov model (HMM), scanning produced HMM against library of HMMs of proteins with experimentally solved structures, constructing 3D models of NgAgo, modelling insertions/deletions, modelling of amino acid sidechains, submission of the top model, and transmembrane helix and topology prediction[32]. NgAgo was analyzed via HHpred[84,85] (https://toolkit.tuebingen.mpg.de/#/tools/hhpred) on 2018 November 27. The parameters for HHpred are HHblits=>uniclust30_2018_08 for multiple sequence alignment (MSA) generation method, 3 for maximal number of MSA generation steps, 1e-3 for E-value incl. threshold for MSA generation, 0% for minimum sequence identity of MSA hits with query, 20% for minimum coverage of MSA hits, during_alignment for secondary structure scoring, local for alignment mode, off for realign with MAC, 0.3 for MAC realignment threshold, 250 for number of target sequences, and 20% for minimum probability in hit list.

3.3.7 Phylogenetic analysis

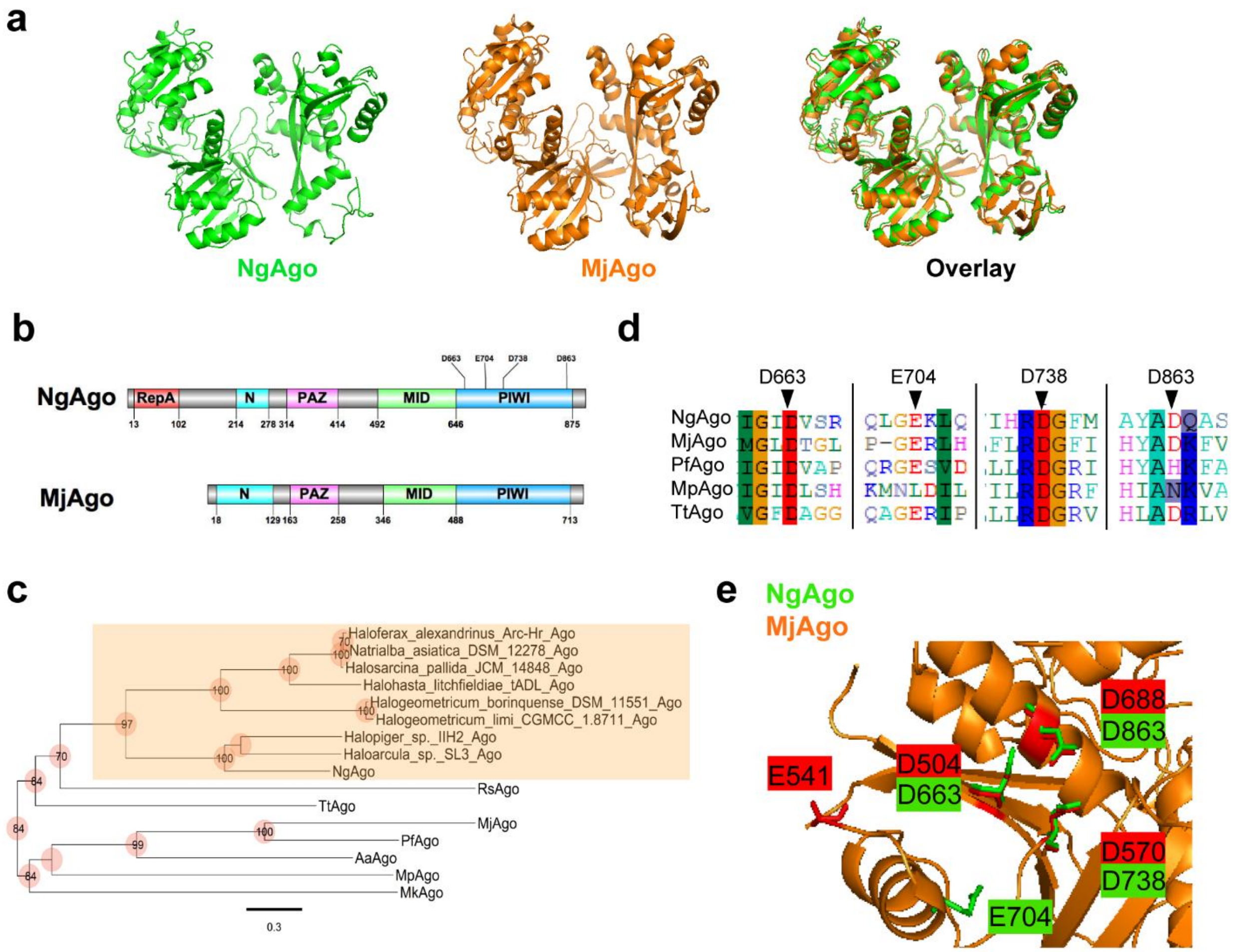
BLAST was used to compare NgAgo protein sequence with all the isolates in the database via the IMG/M server (https://img.jgi.doe.gov/). Representative full-length Argonautes with a repA domain were used to represent each species. Selected pAgos with repA domains and some well-characterized pAgos were compared, and the midpoint rooted tree was generated via the server http://www.genome.jp/tools-bin/ete with unaligned input type, mafft_default aligner, no alignment cleaner, no model tester, and fasttree_default Tree builder parameters. The nwk output file was then used for phylogenetic tree generation in R with ggtree package.

3.4 Results

3.4.1 NgAgo has canonical N-terminal, PIWI, MID, and PAZ domains, and a putative single stranded DNA binding (repA) domain.

Given the ongoing debate of the function of NgAgo, we analyzed its sequence (IMG/M Gene ID: 2510572918) with Phyre 2[54] and HHpred[84,85] to predict its structure based on characterized structural homologs. Phyre 2 and HHpred analyses found with high confidence (probability = 100%) that NgAgo shares structural features with catalytically active pAgos and eukaryotic Agos (eAgos) including archaeal MjAgo, bacterial TtAgo, and eukaryotic hAgo2 (Tables 3-4 and 3-5). Since MjAgo is the only characterized pAgo from Archaea, we used it as a template for comparative modelling. The predicted NgAgo structure is similar to the crystal structure of MjAgo, consisting of canonical N-terminal, PAZ, MID, and PIWI domains (Fig. 3-1a and b). However, the N-terminal domain of NgAgo, which plays a key role in targeted cleavage, is truncated, relative to MjAgo. This may suggest a novel mechanism for strand displacement and binding.

Structural analysis also identified an uncharacterized oligonucleotide/oligosaccharide-binding (OB) fold domain between residues 13-102 of NgAgo that commonly binds single-stranded DNA in eukaryotes and prokaryotes[86] (Fig. 3-1b). This OB domain has recently been identified as a new feature of pAgos[72]. As repA proteins were the most common matches on both Phyre 2 and HHpred, we will refer to this OB domain as repA (Tables 3-6 and 3-7). While the repA domain is absent in all characterized pAgos, at least 12 sequenced pAgo homologs share this domain. Phylogenetic analysis showed that all the repA-containing pAgos were from halophilic Archaea forming a clade that is distinct from that of the current well-characterized pAgos (Fig. 3-1c). This monophyletic group of repA-containing pAgos may represent a distinct class of pAgos that is currently unrecognized in the literature[72]. Moreover, its unique presence within halophiles may be evidence that the repA domain is required for function in high salt environments, potentially replacing the role of the canonical N-terminal domain, which was then truncated through evolution.



a
NgAgo
MjAgo
Overlay
b
NgAgo
RepA
N
PAZ
MID
PIWI
D663
E704
D738
D863
13
102
214
278
314
414
492
646
875
MjAgo
N
PAZ
MID
PIWI
18
129
163
258
346
488
713
c
Haloferax_alexandrinus_Arc-Hr_Ago
Natrialba_asiatica_DSM_12278_Ago
Halosarcina_pallida_JCM_14848_Ago
Halohasta_litchfieldiae_tADL_Ago
Halogeometricum_borinquense_DSM_11551_Ago
Halogeometricum_limi_CGMCC_1.8711_Ago
Halopiger_sp._IIH2_Ago
Haloarcula_sp._SL3_Ago
NgAgo
RsAgo
TtAgo
MjAgo
PfAgo
AaAgo
MpAgo
MkAgo
70
100
100
100
100
97
100
70
84
84
100
99
84
0.3
d
D663
E704
D738
D863
NgAgo IGIDVSR QLGEKLQ IHRDGFM AYADQAS
MjAgo MGLDTGL P-GERLH FLRDGFI HYADKFV
PfAgo IGIDVAP QRGESVD LLRDGRI HYAHKFA
MpAgo IGIDLSH KMNLDIL ILRDGRF HIANKVA
TtAgo VGFDAGG QAGERIP LLRDGRV HLADRLV
e
NgAgo
MjAgo
D688
D863
E541
D504
D663
D570
D738
E704

Figure 3-1. NgAgo belongs to a distinct clade of pAgos with a catalytic DEDX tetrad and novel repA domain. a, Phyre 2 simulation 3D structure based on MjAgo structure (PDB: 5G5T). NgAgo structure is similar to MjAgo structure except for at the N-terminal domain. b, Domain architecture analysis of NgAgo based on Phyre2 and HHpred reveals that NgAgo has an uncharacterized repA domain, a truncated N-terminal domain, a MID domain, and a PIWI domain. c, Phylogenetic analysis of repA-containing pAgos (orange shaded) found from BLASTP against all isolates via JGI-IMG portal and other characterized pAgos. d, The catalytic tetrad of NgAgo is conserved with catalytically active pAgos including MjAgo, PfAgo, MpAgo, and TtAgo in sequence alignment. e, All residues of the catalytic tetrad (D663, E704, D738, and D863) DEDD, except E704 are structurally colocalized with the catalytic tetrad of MjAgo (D504, E541, D570, and D688).

Table 3-5. Top 10 hits of NgAgo in Phyre 2 search

Structure ID	Structure source	Protein	Probability	Identity with NgAgo
5GUH	PDB	Silkworm PIWI-clade Argonaute Siwi	100%	15%
4EI3	PDB	Homo sapiens Argonaute2	100%	18%
3HO1	PDB	Thermus thermophilus Argonaute N546 mutant	100%	19%
4F1N	PDB	Kluyveromyces polysporus Argonaute	100%	14%
3DLB	PDB	Thermus thermophilus Argonaute	100%	19%
2F8S	PDB	Aquifex aeolicus Argonaute	100%	16%
5G5T	PDB	Methanocaldcoccus janaschii Argonaute	100%	15%
1U04	PDB	Pyrococcus furiosus Argonaute	100%	12%
5AWH	PDB	Rhodobacter sphaeroides Argonaute	100%	14%
5THE	PDB	Vanderwaltozyma polyspora Argonaute	100%	17%

Table 3-6. Top 10 hits of NgAgo in HHpred search

Structure ID	Protein	Probability	E-value	Identity to NgAgo
5GUH	silkworm PIWI-clade Argonaute Siwi	100%	1e-86	15%
4Z4D	Homo sapiens Argonaute2	100%	3.4e-77	16%
4F1N	Kluyveromyces polysporus Argonaute	100%	3e-77	17%
4NCB	Thermus thermophilus Argonaute	100%	2.5e-68	17%
5G5S	Methanocaldcoccus janaschii Argonaute	100%	2.6e-68	12%
1YVU	Aquifex aeolicus Argonaute	100%	3.9e-68	16%
1U04	Pyrococcus furiosus Argonaute	100%	1.2e-66	14%
5I4A	Marinitoga piezophila Argonaute	100%	1.6e-65	14%
6D92	Rhodobacter sphaeroides Argonaute	100%	1.9e-55	14%
5THE	Vanderwaltozyma polyspora Argonaute	100%	1.7e-56	17%

Table 3-7. Top 10 hits of repA domain of NgAgo in Phyre 2 search. A non-OB fold domain match was eliminated in this table.

Structure ID	Source	Protein	Probability	Identity to NgAgo
2KEN	PDB	Methanosarcina mazei OB domain of MM0293	95.8%	12%
3DM3	PDB	Methanocaldococcus jannaschii repA	95.2%	23%
2K50	PDB	Methanobacterium thermoautotrophicum repA-related protein	94.6%	12%
1O7I	PDB	Sulfolobus solfataricus ssb	94.4%	16%
1FGU	PDB	Homo sapiens REPA	92.3%	15%
d1jmca2	SCOP	Homo sapiens RPA70	92%	15%
4OWX	PDB	Homo sapiens SOSS complex subunit B1	91.8%	15%
3E0E	PDB	Methanococcus maripaludis repA	78.2%	20%
2K75	PDB	Thermoplasma acidophilum OB domain of Ta0387	67.2%	14%
d1wjja_	SCOP	Arabidopsis thaliana hypothetical protein F20O9.120	66.0%	16%

Table 3-8. Top 10 hits of repA domain of NgAgo in HHpred search. A non-OB fold domain match was eliminated in this table.

Ranking	Structure ID	Protein	Probability	E-value
27	4OWT	Homo sapiens SOSS1 subunit B1	94.68%	0.06
28	1WJJ	Arabidopsis thaliana hypothetical protein F20O9.120	94.65%	0.086
29	1O7I	Sulfolobus solfataricus single-stranded DNA binding protein chain B	94.0%	0.28
30	2K50	Methanobacterium thermoautotrophicum repA	92.46%	0.036
31	3DM3	Methanocaldococcus jannaschii repA	91.96%	0.65
33	3E0E	Methanococcus maripaludis repA	88.18%	2.5
34	1YNX	Saccharomyces cerevisiae repA	87.6%	1.3
35	5D8F	Homo sapiens SOSS complex subunit B1	84.78%	6.7
36	1JMC	Homo sapiens RPA70	82.12%	4.7
37	4HIK	Schizosaccharomyces pombe Pot1pC	81.44%	5.1

Our analysis of NgAgo also confirmed the presence of a conserved catalytic tetrad, DEDX (X: H, D or N)[28], which is critical for nucleic acid cleavage by the PIWI domain of Argonautes. The catalytic tetrad (D663, E704, D738, and D863) of NgAgo aligns well with those from other catalytically active pAgos, including MjAgo[22], PfAgo[29], MpAgo[42], and TtAgo[23] (Fig. 3-1d). Moreover, structural alignment of NgAgo and MjAgo display good colocalization of D663, D738, and D863 within the catalytic tetrad suggesting that NgAgo may have similar nucleic acid cleavage activity (Fig3-1e).

3.4.2 Soluble, but not refolded, NgAgo exhibits DNA cleavage activity *in vitro*

As halophilic proteins tend to be insoluble in low-salt environments due to their sequence adaptations[48,49,87], we first optimized expression conditions to obtain more soluble NgAgo protein (Fig. 3-2). NgAgo was still unstable in optimal expression conditions, as evidenced by truncated peptide products (Fig. 3-2b). We purified wildtype NgAgo from both the soluble and insoluble fractions to test for 5’P-ssDNA guide-dependent DNA cleavage (Fig. 3-3). Insoluble NgAgo was refolded during purification using established methods[26]. Purified NgAgo from the soluble fraction (sNgAgo) nicks plasmid DNA and genomic DNA, independent of a guide (Fig. 3-4a), as evidenced by the presence of the nicked and linearized plasmid. However, refolded NgAgo from the insoluble lysate fraction (rNgAgo) has little or no activity on DNA (Fig. 3-4b), consistent with a study by Ye and colleagues[26].

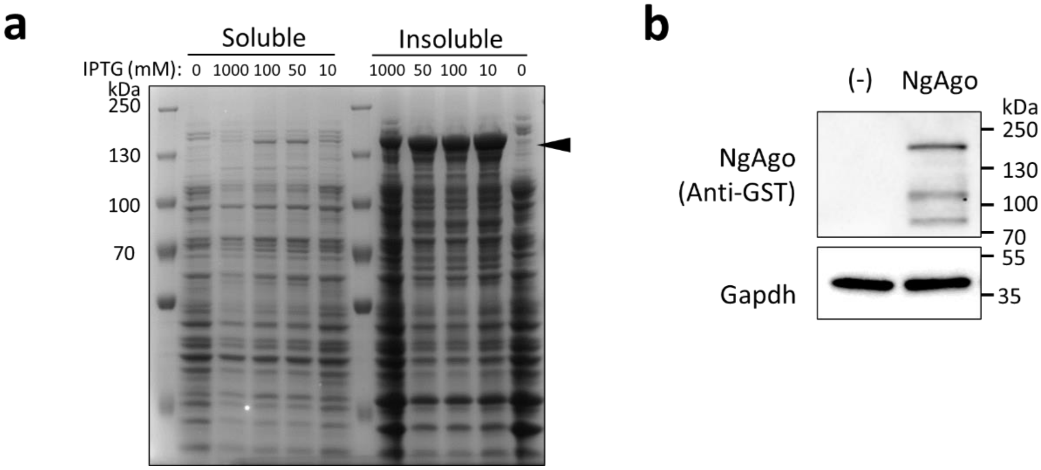


Figure 3-2. Optimization of soluble NgAgo protein expression. a, Different IPTG concentrations (1000 mM, 100 mM, 50 mM, and 10 mM) were used to induce GST-NgAgo expression in BL21 (DE3). Soluble and insoluble protein fractions were analyzed by SDS-PAGE to determine the optimal conditions for soluble NgAgo expression. b, Soluble GST-NgAgo expression with 100mM IPTG was probed with anti-GST antibody and a Gapdh internal control.

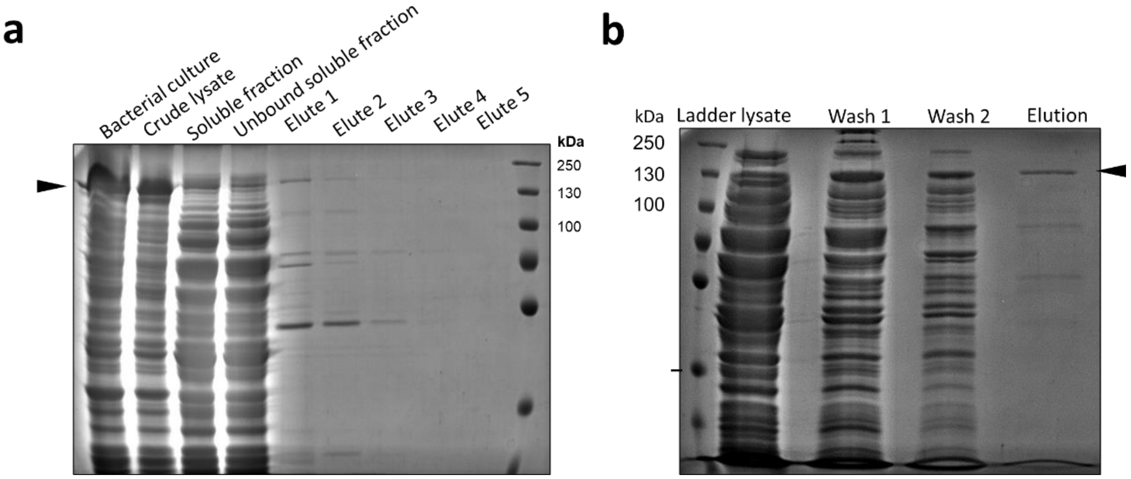


Figure 3-3. SDS-PAGE analysis of His-tag purified NgAgo variants. a, SDS-PAGE analysis of purified WT NgAgo from soluble fraction (sNgAgo). Elute 1 was used for *in vitro* assay. b, SDS-PAGE analysis of purified WT NgAgo from insoluble fraction after refolding (rNgAgo). Elution fraction was used for *in vitro* assay.

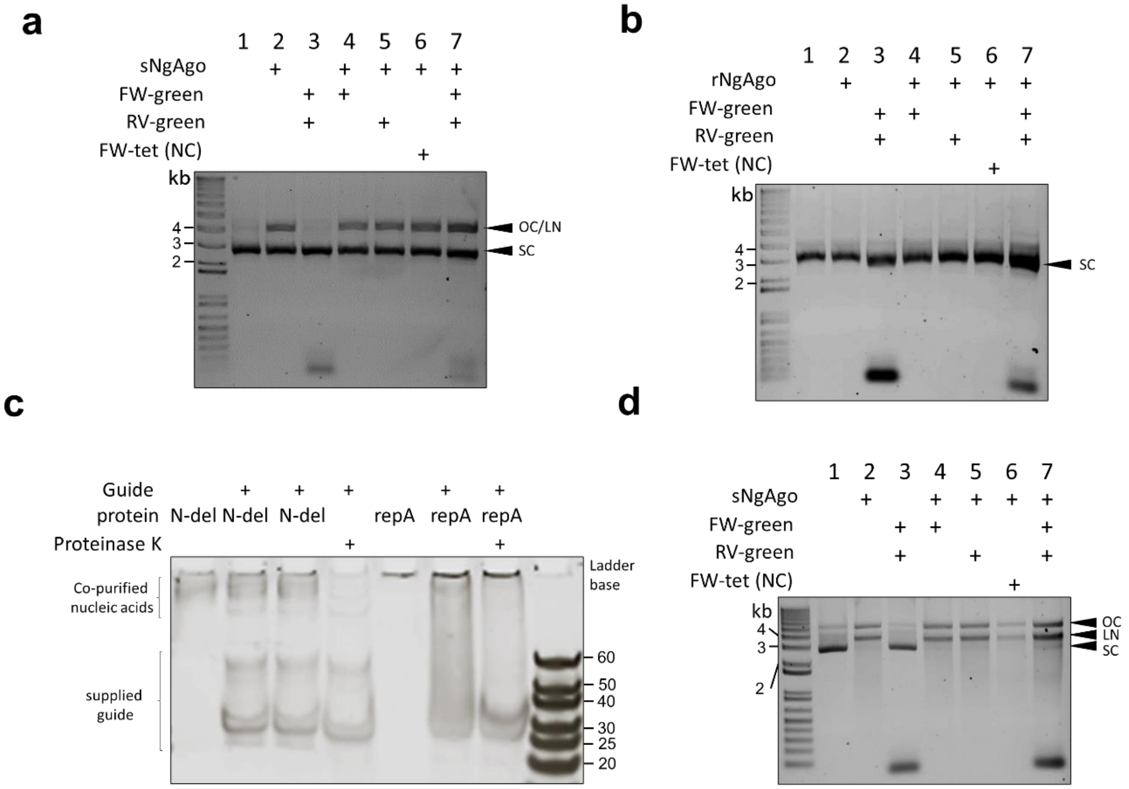


Figure 3-4. Soluble NgAgo variants nick and cut plasmids DNA in vitro. a, Soluble NgAgo (sNgAgo) nicks an cuts plasmids DNA regardless the presence of guide DNA. b, Refolded NgAgo, rNgAgo, has no effect on plasmids DNA regardless the presence of guide DNA. c, Electrophoretic mobility shift assay (EMSA) of N-del and repA domain with guides. N-del does not show band shifting while repA treatment shifts the bands. d, Soluble NgAgo (sNgAgo) nicks and cuts the plasmids DNA regardless the presence of guide DNA with Han's guide-reloading protocol. OC: open circular; LN: linear; SC: supercoiled.

3.4.3 RepA and PIWI domains of NgAgo are required for DNA cleavage

To rule out the possibility of non-specific host nuclease impurities (Fig. 3-5), we pursued cell-free expression of NgAgo. This approach has successfully been used to rapidly prototype other endonucleases including CRISPR-Cas endonuclease[69]. NgAgo expression was induced in the presence of 5' phosphorylated guides that targeted a plasmid substrate, pNCS-mNeonGreen (Figs 3-6a and b). NaCl was supplemented after expression to promote proper folding of the halophilic enzyme (Fig. 3-6c, materials and methods). To identify regions critical for DNA cleavage, we constructed and expressed the repA domain of NgAgo (residues 1-102), a truncated NgAgo without the repA domain (residues 105-887, referred to as N-del) and D663A/D738A point mutations in the full-length protein and N-del variant (Fig. 3-6d). D663A/D738A is a double mutant within the catalytic tetrad that corresponds to the catalytic double mutant D478A/D546A of TtAgo[23], which lost all cleavage activities[23,31].

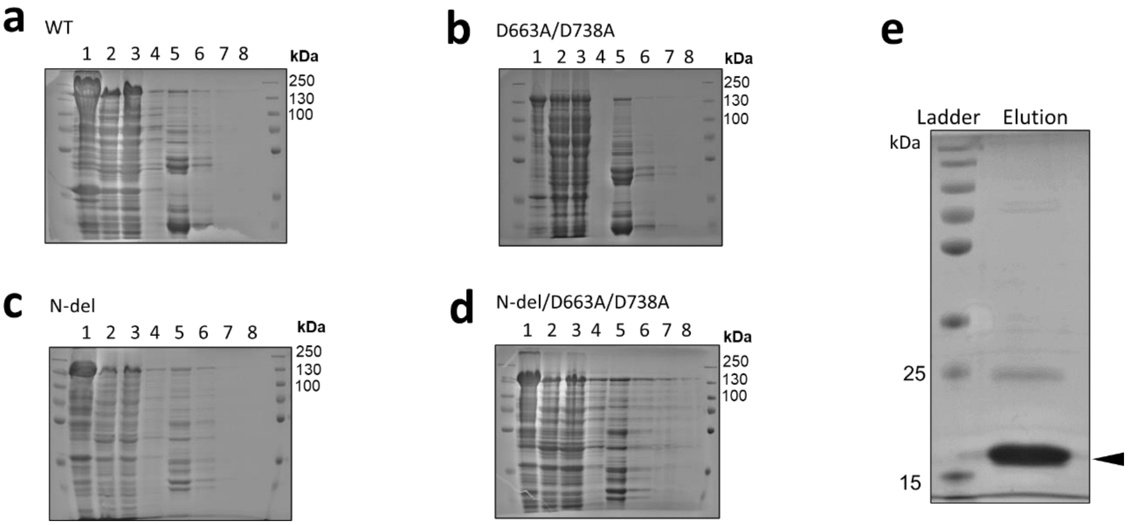


Figure 3-5. SDS-PAGE analysis of GST-tag purified soluble NgAgo variants. a, SDS-PAGE analysis of GST-tag purified WT NgAgo. b, SDS-PAGE analysis of GST-tag purified D663A/D738A. c, SDS-PAGE analysis of GST-tag purified N-del. d, SDS-PAGE analysis of GST-tag purified N-del/D663A/D738A. 1: whole cell lysate; 2: soluble fraction; 3: unbound soluble fraction; 4 washed fraction; 5-8: eluted fraction 1-4. e. His-tagged soluble repA

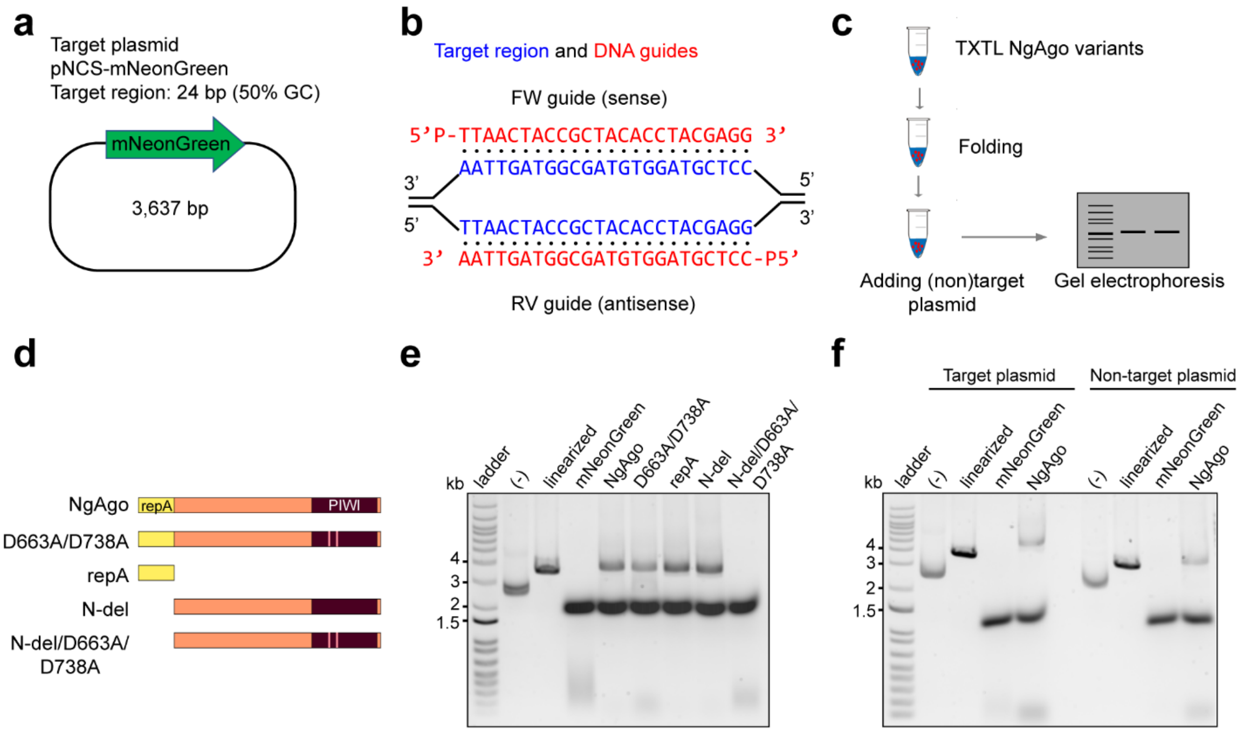


Figure 3-6. NgAgo variants degrade plasmid DNA *in vitro* via the repA domain and D663/D738 residues in the PIWI domain. a, Target plasmid pNCS-mNeonGreen contains a 24-base pair target site with 50% GC content. b, 5' phosphorylated DNA guides binds to target sequence in pNCS-mNeonGreen. c, Procedure for bacterial cell-free-system production of NgAgo and DNA degradation assessment. d, NgAgo variants used in the *in vitro* assay to identify which domain is essential for nicking and cleaving activity. e, Plasmids were treated with NgAgo variants or mNeonGreen as a endonuclease negative control for an hour before analysis on an agarose gel. Wildtype and D663A/D738A degrades plasmids DNA while N-del degrades plasmid DNA with compromised activity. N-del/D663A/D738A loses the ability to degrade plasmid DNA. f, NgAgo degrades both target plasmid pNCS-mNeonGreen and non-target plasmid pBSI-SceI(E/H). Negative controls (-) are plasmids without any treatments.

Not all NgAgo variants displayed DNA cleavage activity, confirming that previously observed DNA cleavage could be attributed to NgAgo activity (Fig. 3-6e). Both wildtype NgAgo and D663A/D738A linearized substrate DNA suggesting catalytic activity beyond the PIWI domain[26] or rescue of functionality by other domains even in the presence of a PIWI mutation. Both repA and PIWI domains participate in DNA cleavage and with each being sufficient for activity as cleavage was retained in both repA and N-del mutants. While it is unclear how the repA domain might lead to DNA damage, its single-stranded DNA binding activity in isolation may be weak (Fig 3-4c), leaving exposed ssDNA susceptible to oxidative degradation[88]. Nonetheless, only in the presence of both a repA deletion and PIWI mutation, N-del/D663A/D738A, is DNA degradation completely lost. When a non-target plasmid with no complementarity to the supplied guides was incubated with the enzymes, fewer lower molecular weight products were generated by NgAgo relative to that when incubated with target plasmid containing a. While this result suggests off-target or guide-independent activity, this activity is reduced relative to guided cleavage as evidenced by fewer degradation products (Fig. 3-6f). That is, NgAgo-induced DNA degradation was also both target specific and non-specific, consistent with proposed pAgo models of non-specific DNA ‘chopping’ for guide acquisition and enhanced specific cleavage of complementary sequences[31].

3.4.4 NgAgo has specific *in vivo* activity at plasmid and genomic loci in bacteria

Next, we tested whether NgAgo can be programmed to target DNA *in vivo*. We chose *E.coli* instead of mammalian cells as our model because NgAgo, like most pAgos, lacks helicase activity needed to separate DNA strands for pAgo recognition and nicking of complementary sequences[41,]. The rapid rate of bacterial DNA replication increases the abundance of accessible unpaired DNA targets for NgAgo actitivy. Additionally, *E.coli* lack histones, which are known to inhibit pAgo activity[75].

Studies have reproducibly demonstrated an ability of NgAgo to reduce gene expression[44,47] and have suggested RNA cleavage as a possible mechanism. However, two alternative hypotheses could also explain this phenomenon: (i) NgAgo cuts DNA leading to poor expression, and (ii) NgAgo inhibits transcription by tightly binding DNA. To distinguish between these three hypotheses, we created a two-plasmid system that harbors an inducible NgAgo expression cassette on one plasmid and another that serves as a target harboring a transcriptionally inactive pseudogene target, *mNeonGreen*, and a selectable marker or essential gene under selective conditions, *cat* (Fig. 3-7a). NgAgo was expressed in cells with both these plasmids and transformed with phosphorylated guide ssDNA (P-ssDNA) targeting different strands of *mNeonGreen*, including forward (FW, sense/coding), reverse (RV, antisense/non-coding), both FW and RV, or without a guide. After transformation, these cells were streaked on selective media (Fig. 3-7b). When guides were targeted to the transcriptionally silent *mNeonGreen* (Fig.

3-8), fewer than half the colony forming units were observed relative to unguided controls (Fig. 3-7c).

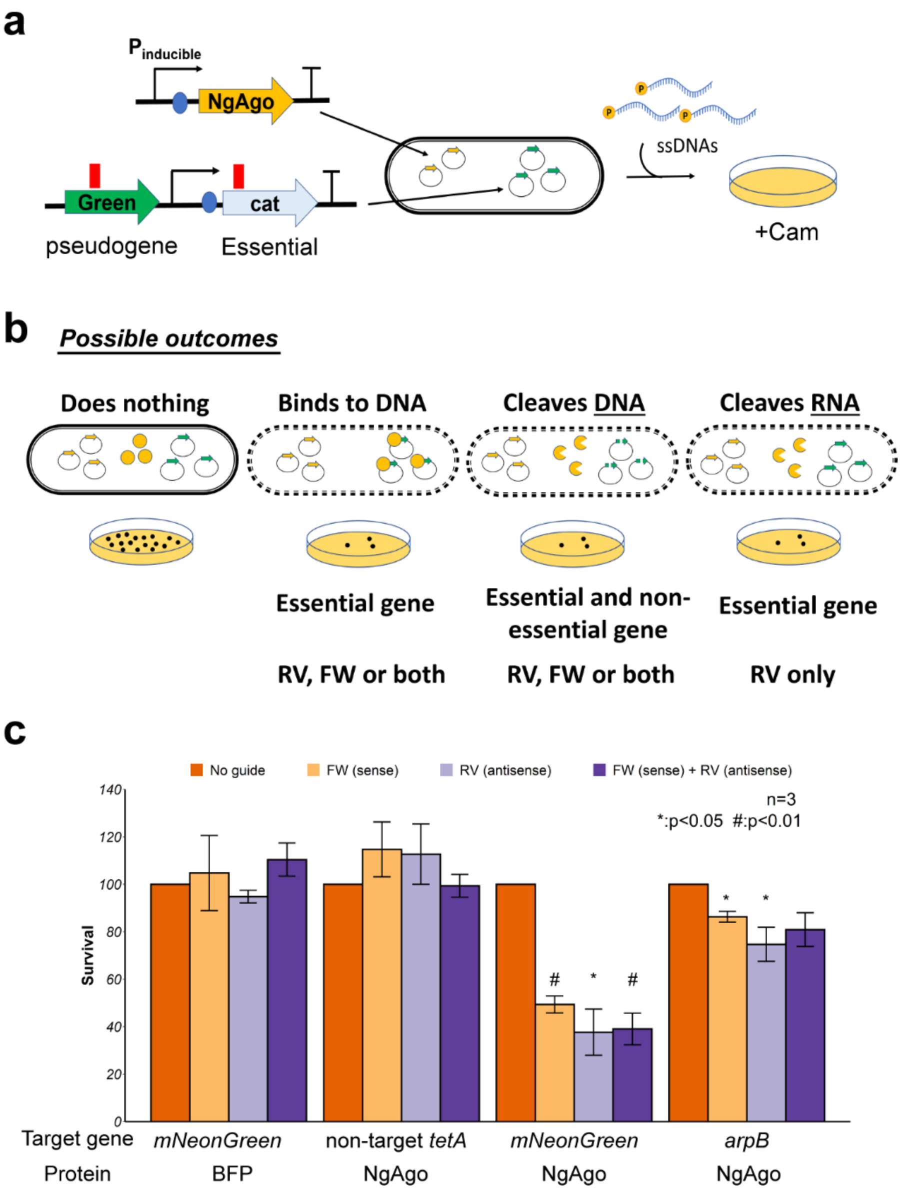


Figure 3-7. NgAgo can be programmed to target DNA in *E. coli*. a, Workflow of testing NgAgo function in *E. coli*. Two plasmids system used to test the function of NgAgo. One plasmid harbors NgAgo driven by T7 inducible promoter while the other low-copy plasmid serves as the target of NgAgo, including an untranscribed pseudogene, mNeonGreen. b. Four possible outcomes relative to an unguided control including no interaction, DNA binding, DNA cleaving, and RNA binding/cleaving, reveal the function of NgAgo. c, Survival rate targeting a pseudogene (mNeonGreen) on the plasmid or targeting a nonessential gene (arpB) in the genome with NgAgo or BFP control.

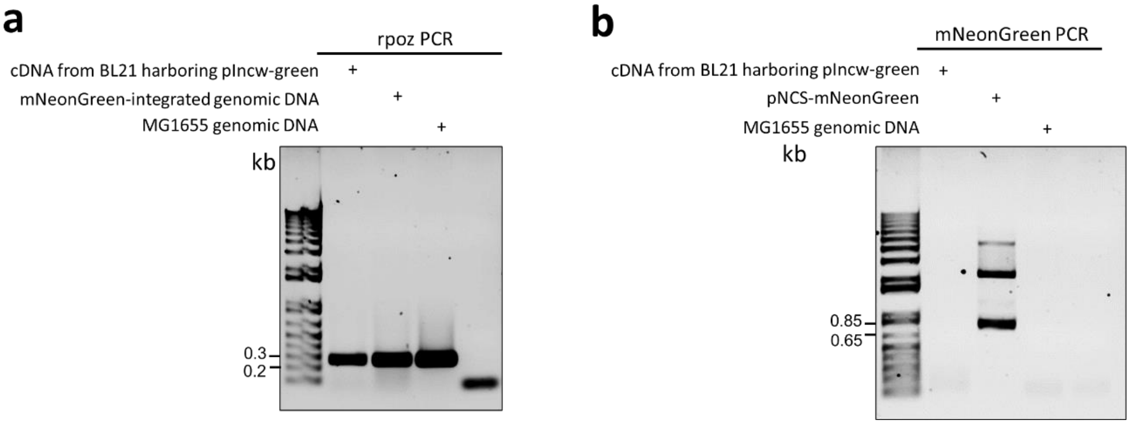

a
rpoz PCR
cDNA from BL21 harboring pIncw-green +
mNeonGreen-integrated genomic DNA +
MG1655 genomic DNA +
kb
0.3
0.2
b
mNeonGreen PCR
cDNA from BL21 harboring pIncw-green +
pNCS-mNeonGreen +
MG1655 genomic DNA +
kb
0.85
0.65

Figure 3-8. Pseudogene mNeonGreen of pIncw-green is transcriptionally silent. a, RNA polymerase subunit, *rpoz*, was amplified with cDNA from BL21 harboring pIncw-mNeonGreen, mNeonGreen-integrated genomic DNA, and WT genomic DNA. b, mNeonGreen was amplified with cDNA from BL21 harboring pIncw-mNeonGreen, pNCS-mNeonGreen plasmid DNA, and WT genomic DNA.

Control studies with either guides alone or NgAgo alone did not identify any cell toxicity, suggesting that the reduction in survival was due to NgAgo activity (Figs.3-9 and 3-10). As similar results were obtained regardless of strand targeted and the target produced no RNA, NgAgo must interact at the DNA level. One possible mechanism is plasmid curing and loss of the selective marker through cleavage of the test plasmid, in agreement with our *in vitro* (Figs 3-4 and 3-11) and cell-free studies (Fig 3-6). Using BFP in place of NgAgo does not reduce survival when incubated with guides complementary to the pseudogene *mNeonGreen* (Fig. 2-7c), confirming the survival reduction effect requires NgAgo expression. Finally, this effect is target specific. When targeted to an absent locus (*tetA*), there were no significant changes in the number of surviving colonies relative to unguided controls (Fig. 3-7c). This assay only quantifies activity relative to an unguided control and as such cannot measure off-target activity present in unguided controls. However, the reduction of survival in a guide- and target-dependent manner suggests that NgAgo has the capacity for targeted DNA endonuclease activity *in vivo* in *E. coli*.

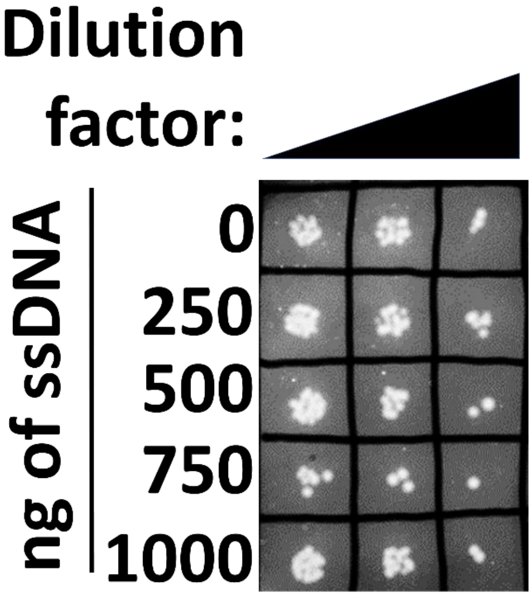


Figure 3-9. ssDNA guides are non-toxic. Different amounts of ssDNAs were transformed into BL21 and spot plated at varying dilutions (1000x, 2000x and 5000x). Ten microliters of each diluted sample was plated on LB.

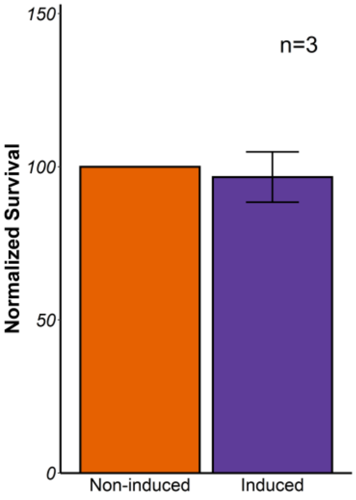


Figure 3-10. Off-target activity assessment with the two-plasmid system. The host strain harboring NgAgo expression plasmid and target plasmid are plated in the presence or absence of IPTG inducer and the number of colony forming units relative to the non-induced control was calculated as ‘survival’. No significant change of survival is observed. Error bars are the standard errors generated from three replicates. Statistically significant results are indicated with * (p-value< 0.05, paired t-test).

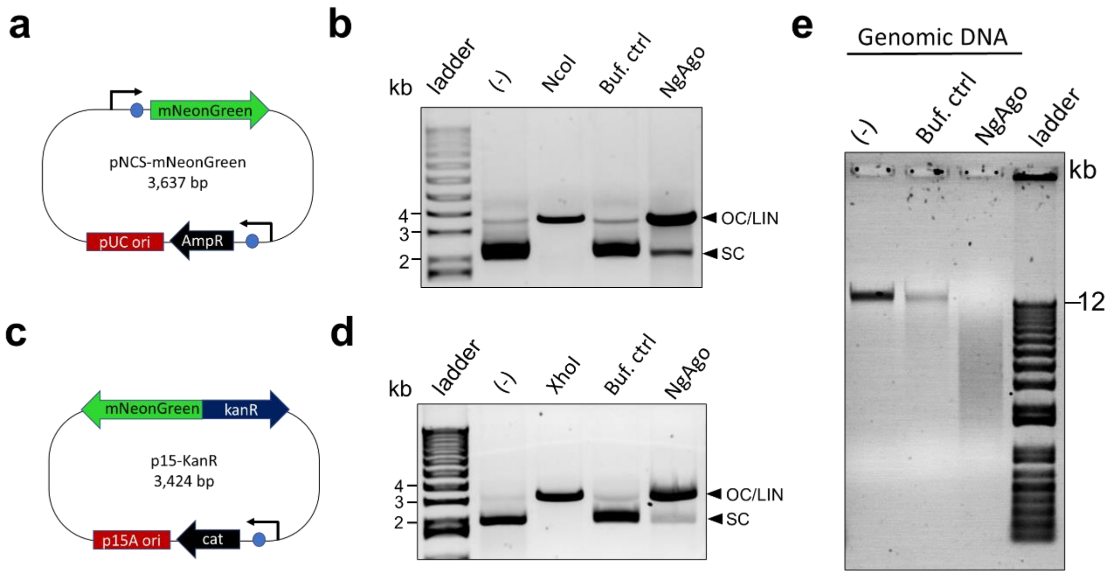


Figure 3-11. Soluble wildtype NgAgo nicks and cuts DNA in the absence of guide DNA. a, Plasmid map of the related plasmid, pNCS-mNeonGreen. This plasmid shares the same ampicillin antibiotic resistant gene with the NgAgo expression plasmid. b, Agarose gel analysis of wildtype NgAgo-treated pNCS-mNeonGreen, showing the degraded DNA product. c, Plasmid map of the unrelated plasmid, p15-KanR. This plasmid does not share genetic elements with the NgAgo expression plasmid. d, Agarose gel analysis of wildtype NgAgo-treated p15-KanR, showing the degraded DNA product. e, Agarose gel analysis of wildtype NgAgo-treated MG1655 genomic DNA, showing the degraded DNA product.

To confirm that the reduced survival is not limited to targets on the plasmid, we also targeted a genomic locus, *arpB*. *arpB* is a non-essential pseudogene that is interrupted by a stop codon[89]. Since *arpB* RNA is not required for survival (i.e., the arpB mutant is nonlethal), RNA cleavage would not reduce survival. However, double stranded DNA breaks in *E. coli* are lethal due to inhibited genome replication[90]. As targeting *arpB* did reduce survival (Fig. 3-7c), this suggests NgAgo also cleaves genomic DNA, consistent with our plasmid cleavage results.

Next, we asked if repA and PIWI domains are required for targeting in *E.coli* by evaluating the ability of different variants to target *mNeonGreen*. Our results showed that the PIWI mutant (D663A/D738A) and truncated repA deletion (N-del) lost the ability to reduce survival (Fig. 3-12), suggesting the process of targeting and DNA cleavage was disrupted. Moreover, PIWI mutation enhanced survival activity via unknown mechanisms (Fig. 3-12), potentially via its interactions with guide and other proteins[27]. Nonetheless, both intact repA and PIWI domains were required for targeted NgAgo activity.

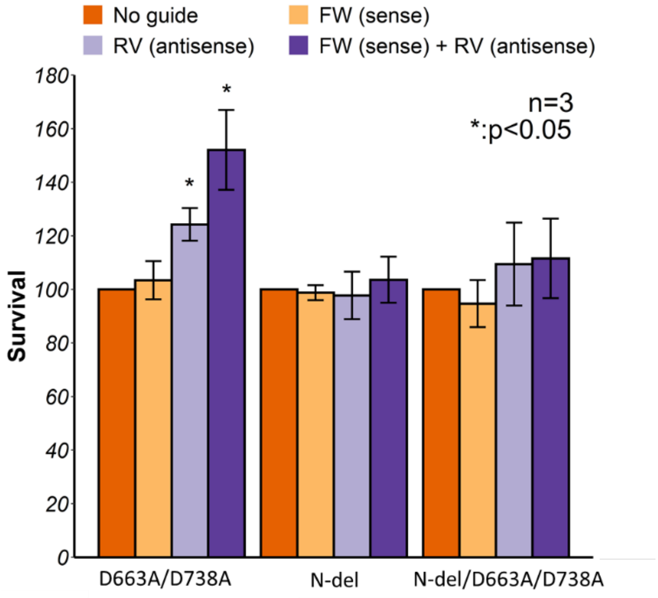


Figure 3-12. NgAgo variants lose the ability to reduce survival. Survival rate targeting a pseudogene (mNeonGreen) on the plasmid with NgAgo variants, including D663A/D738A, N-del, and N-del/D663A/D738A.

3.4.5 DNA-cleaving domains are needed for NgAgo programmable genome editing in bacteria

Since we have shown that NgAgo can cleave DNA *in vitro* and in *E.coli*, we asked whether this activity was essential for the reproducible gene editing by NgAgo observed in other prokaryotes[27]. To test for NgAgo gene editing activity, we created a kanamycin sensitive MG1655 (DE3) strain harboring a cassette composed of a *kanR* resistance gene lacking an RBS and promoter and a *mNeonGreen* gene flanked by two double terminators (Fig. 3-13a). This arrangement prevented any KanR/mNeonGreen expression from transcription read-through and translation from upstream and downstream genes. We then provided a donor plasmid with a truncated *mNeonGreen*, a constitutive promoter, an RBS and a truncated *kanR*, which is also KanR$^-$ but can recombine with our locus to create a KanR$^+$ phenotype (Fig. 3-13a). As DNA breaks in *E.coli* are lethal, repair via recombination should increase the number of KanR$^+$ transformants if NgAgo induces DNA cleavage. We validated this system with CRISPR/Cas9, which showed a 4-fold enhancement in recombination efficiency (Fig. 3-14).

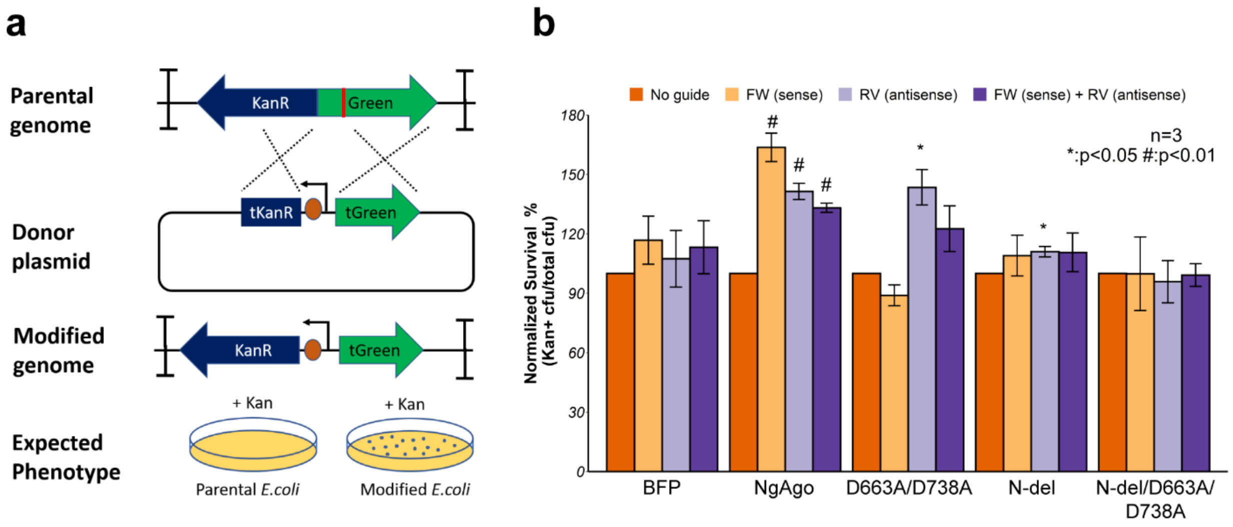


Figure 3-13. NgAgo enhances gene-editing via ʎ-red-mediated homologous recombination in *E.coli*. a, Design of gene-editing assay in MG1655 (DE3). *KanR* and *mNeonGreen* (Green) cassette without promoter and RBS, flanked by two double terminators, is integrated in MG1655 (DE3). Donor plasmid with truncated *mNeonGreen* (tGreen) encodes a nonfunctional truncated *KanR* (tKanR). Guide was transformed to target the *mNeonGreen* (red line). After successful gene editing, modified genome has a functional KanR cassette, enabling survival in Kan selective plate. b, NgAgo variants enhance gene editing efficiency with ~1 microgram of guide(s) relative to an unguided control while blue fluorescent protein (BFP) control has no enhancement with guides. Error bars are the standard errors generated from three replicates. Statistically significant results are indicated with * (p-value< 0.05, paired t-test).

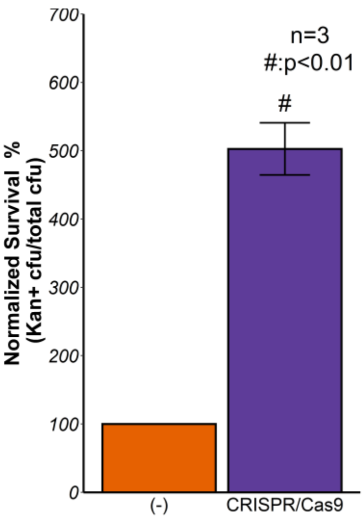


Figure 3-14. CRISPR/Cas9 enhances homologous recombination in *E.coli*. Induction of CRISPR/Cas9 enhance homologous recombination for 4-fold. Kanamycin positive colonies are normalized by total colony forming units. Error bars are the standard errors generated from three replicates. Statistically significant results are indicated with * (p-value< 0.05, paired t-test).

Wildtype NgAgo increased homologous recombination efficiency when provided with FW, RV, and both guides compared with an unguided control (Fig. 3-13b), demonstrating that guide-dependent NgAgo activity can enhance gene editing. In contrast, a BFP protein control showed no statistically significant enhancement in recombination compared to the unguided control (Fig. 3-13b). The PIWI mutant of NgAgo, D663A/D738A, displayed reduced but some statistically significant enhancement in homologous recombination; however, this was only true for one of the guides tested. The PIWI mutant displayed no significant enhancement of recombination with the FW or both guides (Fig. 3-13b). While the mechanism behind this pattern is unclear, these data suggest that the catalytic tetrad within the PIWI domain is not essential for enhanced homologous recombination under some conditions, in agreement with other published studies[27]. The N-del mutant of NgAgo lacking the repA domain displayed even weaker statistically significant enhancement in homologous recombination above unguided controls (11%) in the presence of the RV guide only (Fig. 3-13b). The N-del/D663A/D738A catalytic mutant showed no increase in gene editing activity in the presence of FW, RV, or both guides compared to an unguided control. This trend in homologous recombination enhancement is consistent with our observed DNA endonuclease activities (Fig 3-6e) suggesting that the DNA endonuclease activity mediated by the repA and PIWI domains is essential for enhanced homologous recombination and gene editing.

3.5 Discussion

NgAgo has been subject to intense debate in the literature in recent years[25,45–47,91]. Although previous studies suggested that refolded NgAgo does not cut DNA *in vitro*[26,41], consistent with our findings, we establish that soluble NgAgo can, in fact, cleave DNA *in vitro*. That is, refolded NgAgo, which has been historically studied due to the poor soluble expression of this halophilic enzyme, may not be an accurate assessment of NgAgo activities. However, when soluble protein is concentrated and isolated, there is indeed some capacity for nonspecific or guide-independent DNA cleavage as we have demonstrated *in vitro*. Moreover, this behavior may be salt dependent, reflecting the halophilic lifestyle of the native host; NgAgo expressed from cells grown with LB Lennox showed no activity in our hands (data not shown) relative to that produced from cells grown on LB Miller (this work). Our parallel studies in cell-free expression systems that allow for control of salt conditions and lack potentially contaminating endonuclease expression confirm this observation. Most importantly, we generated a catalytically dead N-del/D663A/D738A mutant making it unlikely that the detected activity is the result of sample contamination.

NgAgo activity is mediated not only by the PIWI domain, like canonical pAgos, but also an uncharacterized and previously unrecognized accessory repA or single-stranded DNA binding domain fused to the N-terminus that appears common among halophilic pAgos (Fig 3-1c). Our work is the first report to suggest a role for this domain in NgAgo function and may be another source of the ongoing literature debate. Previously studied 'catalytic' mutants left this domain intact and were unable to detect a change in NgAgo function suggesting sample contamination or inactivity[26]. However, this and growing evidence from the literature[27,41,61] suggest that accessory proteins and domains may be essential for pAgo function. As homologous accessory proteins from heterologous hosts can mediate function[27,41], we investigated whether *in vivo* cleavage, as observed via cell survival and DNA recombination efficiency, would be induced by NgAgo and its mutants. Not only were these assay results consistent with DNA cleavage, but they also importantly suggested an ability to target specific gene loci via single-stranded 5'P DNA guides. Our work here underscores the role of unrecognized accessory proteins, supplied via the expression host, and a need to characterize these proteins to more accurately assess pAgo activity.

Finally, our results provide supporting evidence to encourage the development of NgAgo for gene-editing. When provided with homologous target and donor sequences, NgAgo can enhance homologous recombination. Much like other pAgos, the PIWI domain participates in DNA editing in prokaryotes as shown here and by Fu *et al*[27]. Moreover, without repA, PIWI mutants of NgAgo exhibit reduced cleavage activity with a concomitant reduction in homologous recombination efficiency. Both the repA deletion and the PIWI mutation (N-del/D663A/D738) are needed to fully abolish catalytic and gene-editing functions. In the presence of both functional domains, NgAgo can effectively enhance homologous recombination by inducing a double stranded break at a targeted region. Despite the programmable DNA-cleaving ability of NgAgo, there remain several challenges to its development as a robust tool for gene-editing applications: guide-independent or off-target cleavage, unknown accessory proteins needed for function, poor expression, salt dependence, and potentially low activity in eukaryotic hosts. Nonetheless, further insight may lead to protein engineering strategies to overcome these hurdles and develop NgAgo as a robust tool for gene-editing.

3.6 Conclusion

Based on the above findings, we conclude that NgAgo is a novel DNA endonuclease that belongs to an unrecognized class of pAgos defined by a characteristic repA domain. NgAgo uses both a well-conserved catalytic tetrad in PIWI and a novel uncharacterized repA domain to cleave DNA. This cleavage activity is essential to enhancing gene-editing efficiency in prokaryotes. Despite the challenges of NgAgo, our work establishes innovative approaches to probe NgAgo activity (and that of other pAgos) and identifies critical protein features for its development as a next generation synthetic biology tool.

CHAPTER 4. SOLUBILITY ENGINEERING OF HALOPHILIC NGAGO

4.1 Introduction

Halophilic proteins are proteins isolated from (salt-loving organisms) halophiles, which live in high-salt environments containing salts ranging from 0.5 M to 5.2 M[92]. Compared to their non-halophilic homologs, halophilic proteins have evolved multiple features to adapt to high-salt environments. First, halophilic proteins substitute negatively-charged amino acids for neutral amino acids on their surface to maintain protein structure and function in high-salt conditions[48,49]. These negatively-charged residues bind to salt and water, contributing to the stability and folding process of halophilic proteins. This process is recognized as a key determinant of protein adaptation in high-salt environments[49,87]. Second, halophilic proteins have less hydrophobic amino acids, resulting in a greater propensity to form random coils rather than alpha-helix structures[93]. Although hydrophobic amino acids contribute to protein stability in non-halophilic proteins via hydrophobic interactions, strong hydrophobic interactions in halophilic proteins may make them prone to aggregation in hypersaline conditions, resulting in loss of function. Decreasing the hydrophobic residues allows the halophilic proteins to weaken the hydrophobic interactions, preventing aggregation and loss of protein function[94].

These adaptations have made halophilic proteins stable in high-salt conditions; however, halophilic proteins tend to misfold when expressed in low-salt conditions due to loss of support of salt. Due to this characteristic, the activity of halophilic proteins is greatly compromised. In Chapter 3, I showed that Argonaute from halophilic archaeon *Natronobacterium gregoryi* (NgAgo) had DNA cleavage and gene-editing activities[60]. However, most of the NgAgo proteins (>90%) were misfolded, limiting their activity in bacterial hosts for gene-editing. Therefore, there is a critical need to increase the solubility of NgAgo to improve its activity.

Although there has been no previous effort to improve the solubility of the halophilic protein in low-salt conditions, engineering of a non-halophilic protein to adapt in high-salt conditions provides valuable insight. By increasing the negatively-charged residues on the protein surface, a non- halophilic protein can function in high-salt conditions without significant compromise of its function[95]. With this inspiration, I hypothesized that by replacing the negatively-charged residues with neutral or positively-charged residues on the protein surface, halophilic proteins can increase solubility when expressed under low-salt conditions.

I systematically identified the negatively-charged residues on the surface and compared them with the catalytically active pAgos. I then narrowed down my focus on replacing the negatively-charged residues that are not conserved among homologs with neutral or positively charged residues as non-conserved residues are less likely to be critical to function. Moreover, a computational program, Suspect[96], was used to predict whether the mutation would disrupt the protein structure, increasing the probability of success. With these criteria, I identified seven residues for mutation and validation. Neutral and/or positively-charged residues then replaced the amino acids located at these positions. The mutants were then subjected to SDS-PAGE/Western Blot to test their solubility. Mutants with increased solubility were then tested if they can improve their ability to induce DNA break in bacteria.

4.2 Materials and Methods

4.2.1 Strains and plasmids

E. coli strains and plasmids used in this study are listed in Table 4-1. Primers used in construction are listed in Table 4-2. All molecular biology manipulations were carried out according to standard practices[76]. NgAgo and its mutants were transformed and expressed in BL21 (DE3) with Miller LB.

Plasmid pET32-GST-NgAgo was used to generate NgAgo-E249Q, NgAgo-E249R, NgAgo-E252Q, NgAgo-E252R, NgAgo-D290N, NgAgo-D290K, NgAgo-E463Q, NgAgo-D533N, NgAgo-D537N, and NgAgo-E550Q. The template DNA was subjected to site-directed mutagenesis with Phusion DNA polymerase (Thermo Fisher Scientific, Waltham, MA. Cat. No: F530S) and primers listed in Table 4-2. The clones were sequenced to confirm their identity before verifying their solubility.

Table 4-1. Strains and plasmids for solubility mutants

Name	**Relevant genotype**	**Vector backbone**	**Plasmid origin**	**Source**
Strains				
BL21 (DE3)	F– ompT gal dcm lon hsdSB(rB–mB–) λ (DE3) [lacI lacUV5-T7p07 ind1 sam7 nin5]) [malB+]K-12(λS)			[44]
Plasmids				
pET-GST-NgAgo	*bla*, lacI, P_{T7}-GST-NgAgo	pET-GST-Ago-His	pBR322	This study
pET-GST-NgAgo-E249Q	*bla*, lacI, P_{T7}-GST-NgAgo-E249Q			This study
pET-GST-NgAgo-E249R	*bla*, lacI, P_{T7}-GST-NgAgo-E249R			This study
pET-GST-NgAgo-E252Q	*bla*, lacI, P_{T7}-GST-NgAgo-E252Q			This study
pET-GST-NgAgo-E252R	*bla*, lacI, P_{T7}-GST-NgAgo-E252R			This study
pET-GST-NgAgo-D290N	*bla*, lacI, P_{T7}-GST-NgAgo-D290N			This study
pET-GST-NgAgo-D290K	*bla*, lacI, P_{T7}-GST-NgAgo-D290K			This study
pET-GST-NgAgo-E463Q	*bla*, lacI, P_{T7}-GST-NgAgo-E463Q			This study
pET-GST-NgAgo-D533N	*bla*, lacI, P_{T7}-GST-NgAgo-D533N			This study
pET-GST-NgAgo-D537N	*bla*, lacI, P_{T7}-GST-NgAgo-D537N			This study
pET-GST-NgAgo-E550Q	*bla*, lacI, P_{T7}-GST-NgAgo-E550Q			This study

Table 4-2. DNA primers used in this study. The modified nucleotides corresponding to the modified codons are in lower case.

Name	**Sequences (5'>3')**	**Template**	**Used to construct**
E249Q 5'	CGGTTACTCGCCCGCcAgCTCGTCGAAGAGGGG	pET-GST-NgAgo	pET-GST-NgAgo-E249Q
E249Q 3'	CCCCTCTTCGACGAGcTgGCGGGCGAGTAACCG		
E249R 5'	CGGTTACTCGCCCGCcgtCTCGTCGAAGAGGGG	pET-GST-NgAgo	pET-GST-NgAgo-E249R
E249R 3'	CCCCTCTTCGACGAGacgGCGGGCGAGTAACCG		
E252Q 5'	GCCCGCGAACTCGTCcAgGAGGGGCTCAAACGC	pET-GST-NgAgo	pET-GST-NgAgo-E252Q
E252Q 3'	GCGTTTGAGCCCCTCcTgGACGAGTTCGCGGGC		
E252R 5'	GCCCGCGAACTCGTCcgtGAGGGGCTCAAACGC	pET-GST-NgAgo	pET-GST-NgAgo-E252R
E252R 3'	GCGTTTGAGCCCCTCacgGACGAGTTCGCGGGC		
D290N 5'	CTACATGAGCGGTATaACCTCTCTGTCGAAG	pET-GST-NgAgo	pET-GST-NgAgo-D290N
D290N 3'	CTTCGACAGAGAGGTtATACCGCTCATGTAG		
D290K 5'	CTACATGAGCGGTATaAaCTCTCTGTCGAAGTC	pET-GST-NgAgo	pET-GST-NgAgo-D290K
D290K 3'	GACTTCGACAGAGAGtTtATACCGCTCATGTAG		
E463Q 5'	AATGGGTCCACGGTAcAGTTCTCCTCGGAGT	pET-GST-NgAgo	pET-GST-NgAgo-E463Q
E463Q 3'	ACTCCGAGGAGAACTgTACCGTGGACCCATT		
D533N 5'	TGCAAAGCGCAGTGGaACACGATGGCTGACC	pET-GST-NgAgo	pET-GST-NgAgo-D533N
D533N 3'	GGTCAGCCATCGTGTtCCACTGCGCTTTGCA		
D537N 5'	TGGGACACGATGGCTaACCTCCTCAACCAAG	pET-GST-NgAgo	pET-GST-NgAgo-D537N
D537N 3'	CTTGGTTGAGGAGGTtAGCCATCGTGTCCCA		
E550Q 5'	CCACCGACACGGAGCcAGACCGTCCAATATG	pET-GST-NgAgo	pET-GST-NgAgo-E550Q
E550Q 3'	CATATTGGACGGTCTgGCTCCGTGTCGGTGG		

4.2.2 Expression of NgAgo and its mutants

After transformation and inoculation overnight at 37 °C, the liquid culture of bacteria with NgAgo expression plasmids was diluted a hundred-fold and grown until OD=0.5 for protein induction with 0.1 mM IPTG in Miller LB. The bacteria were then grown for 4 hours at 37 °C before harvesting at 11,5000 g at 4 °C for 5 minutes. The cell pellet was resuspended in TN buffer (10 mM Tris and 100mM NaCl, pH 7.5) and lysed via sonication at a medium power setting (~50 W) in 10 s intervals, with intervening 10 s incubations on ice to reduce heat denaturation. Cell lysates were then clarified at 12000 rpm at 4 °C for 30 minutes. The supernatant was collected as a soluble protein fraction.

4.2.3 SDS-PAGE and Western Blot

The soluble fraction of NgAgo protein expression lysate was quantified via Bradford assay (Thermo Fisher Scientific, Waltham, MA. Cat. No: 23200). The BL21 (DE3) without any plasmid was used as a negative control, while wildtype GST-NgAgo was used as a baseline to test if the mutants have higher solubility. An equal mass of protein lysate with the addition of water if needed was mixed with fourteen microliters of 2X Tris-glycine SDS Sample Buffer (Thermo Fisher Scientific, Waltham, MA. Cat. No.: LC2676) and supplemented with 2 μL of 1 M DTT (Thermo Fisher Scientific, Waltham, MA. Cat. No.: AC426380100). After incubation at 85 °C for 5 min to denature the proteins, samples were then placed on ice for 5 min before loading on to 8% SDS-polyacrylamide gels. PageRuler Plus Prestained Protein Ladder (Thermo Fisher Scientific, Waltham, MA. Cat. No.:26620) was used as a ladder. SDS-PAGE was stained and destained before visualization[97]. Briefly, the SDS-PAGE was stained with Coomassie blue (Fisher Scientific, Pittsburgh, PA; Cat. No.: BP101−25) for 10 min. Gels were then destained with destaining buffer (10% glacial acetic acid and 10% methanol) overnight before visualization under visible light with an Azure c400 imager (Azure Biosystems, Dublin, CA).

For Western Blot, the protocol is similar to a previous studies[98,99]. Briefly, the gel was then transferred to a PVDF membrane for 90 mins with 100v. The PVDF membrane was then blocked by 5% skim milk in Tris-buffered saline with 0.1% Tween-20 at 23°C for an hour. Then the membrane was subjected to primary antibodies hybridization for 16 hours. Primary antibodies, including anti-GAPDH (1:2000 dilution) (Invitrogen, Carlsbad, CA. Cat. No.: MA5-15738-HRP) and anti-GST (1:1000 dilution) (Santa Cruz Biotechnology, Dallas, TX. Cat. No.: sc-138), were diluted in 5% skim milk. The primary antibody-hybridized PVDF membrane was washed with Tris-buffered saline with 0.1% Tween-20 three times for 5 minutes each. The membranes were then hybridized to secondary antibody m-IgGκ BP-HRP (1:2500 dilution) (Santa Cruz Biotechnology, Dallas, TX. Cat. No.: sc-516102) at 23°C for an hour before imaging under chemiluminescent detection mode with an Azure c400 imager (Azure Biosystems, Dublin, CA).

4.2.4 Protein structure modeling

Wildtype and solubility mutants, including N-del-D290N, N-del-D290k, and N-del-E463Q, were subjected to Robetta structure prediction. A deep-learning-based TrRosetta[100] was selected as the prediction algorithm, and no other parameters are needed.

4.3 Results and discussion

4.3.1 Selection of ten mutants for increasing solubility

To identify the residues that might induce misfolding in low-salt environments, I focused on the negatively-charged residues on the NgAgo protein surface based on the predicted NgAgo structure (Fig. 4-1). The negatively charged residues were checked if they are conserved across other catalytically active pAgos. As conserved amino acids at certain positions indicate functionality, we selected the non-conserved amino acids as they are unlikely to affect the protein structure and function if mutated. The selected sites were then subjected to mutational sensitivity analysis[96], which inform the impact of the substitution of all amino acids in a given position. I selected the positions that do not significantly impact the protein structure and function when subjected to mutation (Fig. 4-1).

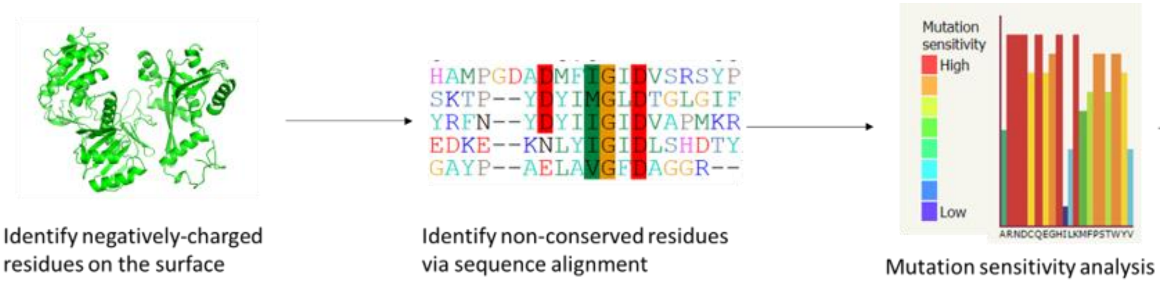


Figure 4-1. Selection pipeline for identification of negatively-charged residue candidates on the surface for solubility engineering. Negatively-charged amino acids on the predicted NgAgo structure were identified. The identified residues were then examined if they are conserved across catalytically active pAgos. The non-conserved residues were then subjected to mutational sensitivity analysis to examine the impact of all possible mutations in a given position.

After this series of selections, I narrowed down to seven candidates, including E249, E252, D290, E463, D533, E537, and E550. As these negatively-charged residues may cause misfolding in low-salt environments, I mutated the negatively-charged to neutral residues. In addition, I mutated E249, E252, and D290 to positively-charged residues to test if they further increase solubility as these three sites are located within or near the N-terminal domain, involving in the target nucleic acid dissociation after cleavage[33,34] but not as important as residues found in the guide-binding domain (PAZ and MID domains)[35–40] and nucleic acid-cleaving domain (PIWI domain)[28]. Thus, I created NgAgo-E249Q, NgAgo-E249R, NgAgo-E252Q, NgAgo-E252R, NgAgo-D290N, NgAgo-D290K, NgAgo-E463Q, NgAgo-D533N, NgAgo-D537N, and NgAgo-E550Q for solubility testing.

4.3.2 NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q have increased solubility

After making the mutants, I expressed them in bacteria and tested their solubility compared to the wildtype NgAgo. The soluble fraction of protein lysate of each mutant was collected and ran on SDS-PAGE to check for the intensity of the soluble band. However, as the band did not have significant changes (Fig. 4-2), it is difficult to assess the changes in solubility.

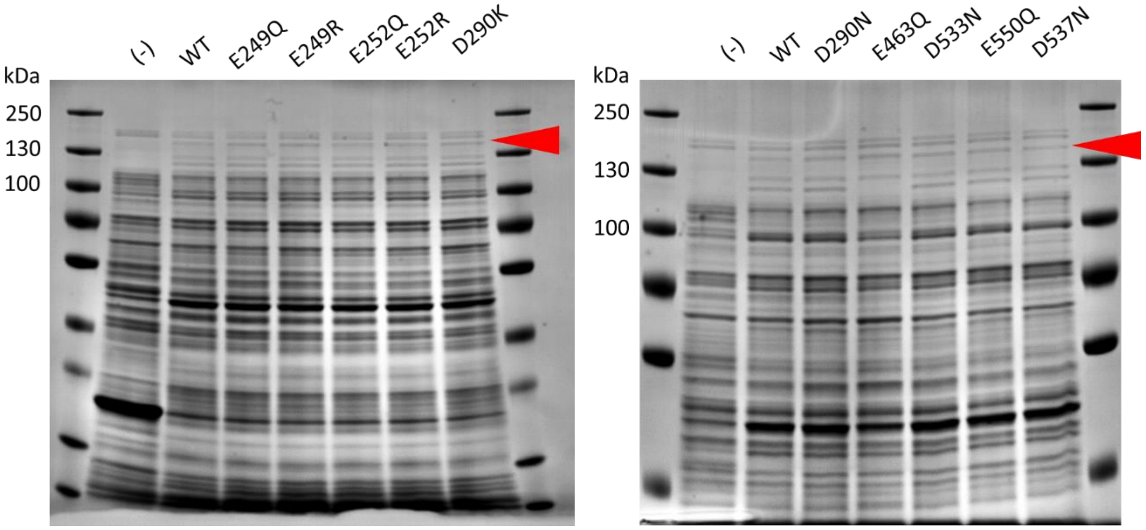


Figure 4-2. Soluble fraction of NgAgo mutants couldn't be assessed their solubility via SDS-PAGE. The soluble fraction of lysate containing individual NgAgo variants was collected and ran with SDS-PAGE to determine their solubility. BL21 (DE3) was used as a control to test if NgAgo-E249Q, NgAgo-E249R, NgAgo-E252Q, NgAgo-E252R, NgAgo-D290N, NgAgo-D290K, NgAgo-E463Q, NgAgo-D533N, NgAgo-D537N, and NgAgo-E550Q have increased soluble fraction. Arrowheads indicate the expected size of the NgAgo variants.

Then I used a more sensitive Western Blot, which can amplify the signals of the target protein, to better assess changes in solubility. A house-keeping gene, Gapdh, was used as a protein loading control. Indeed, Western Blot with wildtype and mutants showed that NgAgo-D290K, NgAgo-D290N, and NgAgo-E463Q have a significant increase in intensity compared to wildtype NgAgo (Fig. 4-3), indicating increased solubility/expression of these mutants.

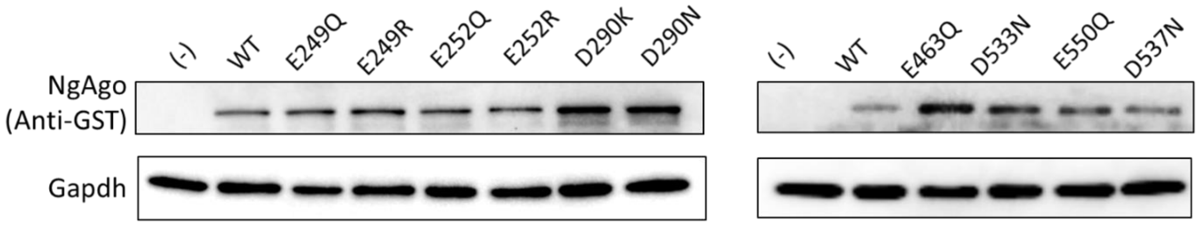


Figure 4-3. NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q have increased solubility. Soluble fractions of lysate containing NgAgo-E249Q, NgAgo-E249R, NgAgo-E252Q, NgAgo-E252R, NgAgo-D290N, NgAgo-D290K, NgAgo-E463Q, NgAgo-D533N, NgAgo-D537N, and NgAgo-E550Q were loaded with an equal amount (30 µg) to test their solubility. Antibodies probing against Gapdh and GST were used to detect Gapdh loading control and NgAgo variants, respectively.

4.3.3 NgAgo-D290K, NgAgo-D290N, and NgAgo-E463Q lose the ability to induce cell death in bacteria

Given that increased solubility does not ensure increased activity, I tested if the mutants can induce cell death in bacteria when programmed with guides. As I would like to see if the mutants are more active in reducing the survival of bacteria, I reduced the amount of guides from 1000 ng to 850 ng. As shown in figure 4-4, wildtype NgAgo can still significantly reduce survival but to a lesser extent when targeting the nonessential pseudogene, mNeonGreen. Then I tested if NgAgo mutants can reduce survival. When programmed with guides targeting a nonessential pseudogene, NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q lose the ability to reduce survival, indicating the mutations break the NgAgo functions.

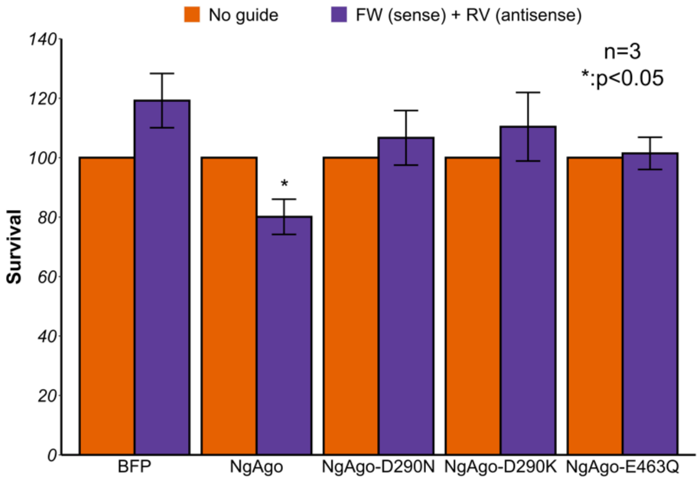


Figure 4-4. NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q lose the ability to induce cell death in bacteria. Survival assay was used to detect the activity of NgAgo variants, including NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q. BFP was used as a protein control.

As D290 is located in the region between the N-terminal and PAZ domains of NgAgo (Fig. 3-1), it should not affect the critical function of both N-terminal and PAZ domains. Still, mutation of this site removes the ability of NgAgo to impact cell survival. Similarly, although E463 is located in a region without functional domains (between PAZ and MID domains of NgAgo) (Fig. 3-1), the mutation of this site also breaks protein function. To figure out why mutation of these two sites leads to loss of protein function, I performed structural modeling of these three mutants, including NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q.

Given that the currently available protein structures of pAgos do not have a repA domain, the modeling based on the deep-learning method, TrRosetta[100], resulted in unreliable structures with low confidence (<0.7, as the structures with score below 0.7 would not be refined accurately) (data not shown). I then removed the repA domain (N-del) for modeling. Indeed, eliminating the repA domain resulted in 0.7-0.72 confidence of the modeling for subsequent analysis. Structural modeling of these three mutants showed subtle but significant changes of the structures compared to wildtype NgAgo. Compared to the N-del, N-del-D290K has significant differences. PAZ (light blue; top right) and MID (green; middle) have more extended alpha-helix structures (Fig. 4-5a' and b'). As PAZ and MID domains form a binding pocket of the guide, structure disruption is likely to negatively impact the guide binding. Moreover, the PIWI domain (red) of N-del-D290K loses an alpha-helix structure (Fig. 4-5a'' and b''), which might affect the cleavage activity.

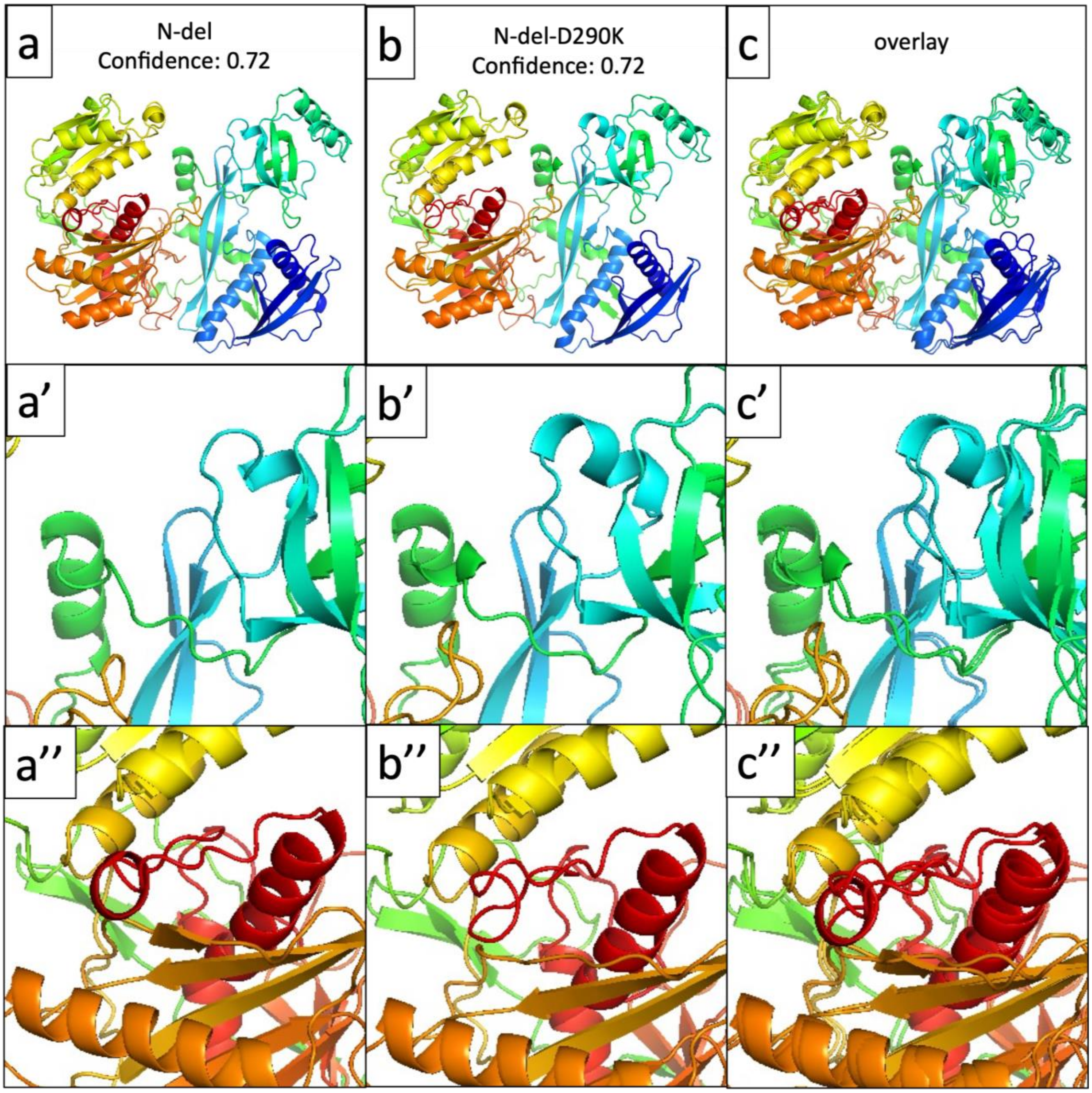


Figure 4-5. Structure modeling of N-del (a, a’, and a’’), N-del-D290K (b, b’, and b’’), and their overlay (c, c’, and c’’). Both predicted N-del and N-del-D290K have reliable structures with a confidence of 0.72.

Ndel-D290N also has significant changes compared to the N-del (Fig. 4-6). For example, the PIWI domain (orange) has an additional alpha-helix structure in the center of the guide binding/target cleaving site (Fig. 4-6a’ and b’). The alpha-helix structure in the PIWI domain (red) is also missing (Fig. 4-6a’’ and b’’). These changes in the PIWI domain might explain why NgAgo-D290N loses its activity.

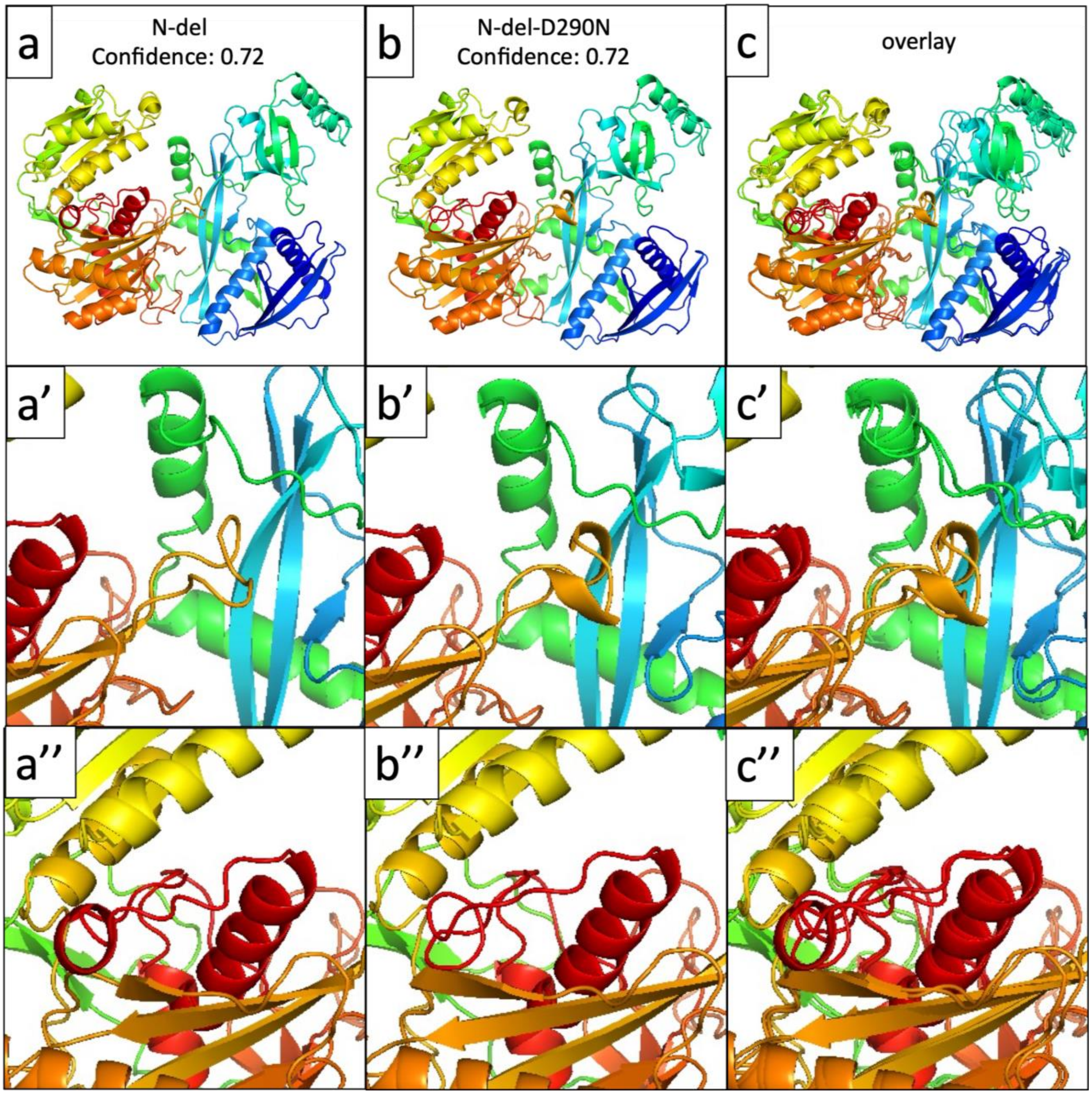


Figure 4-6. Structure modeling of N-del (a, a’, and a’’), N-del-D290N (b, b’, and b’’), and their overlay (c, c’, and c’’). Both predicted N-del and N-del-D290N have reliable structures with a confidence of 0.72.

Ndel-E463Q has dramatic changes in the structure (Fig. 4-7). The MID (green; middle) domain has a more extended alpha-helix structure (Fig. 4-7a’ and b’), while the PIWI domain (red) also has a distorted alpha-helix structure (Fig. 4-7a’’ and b’’), which provide a reasonable explanation for why NgAgo-E463Q loses its activity.

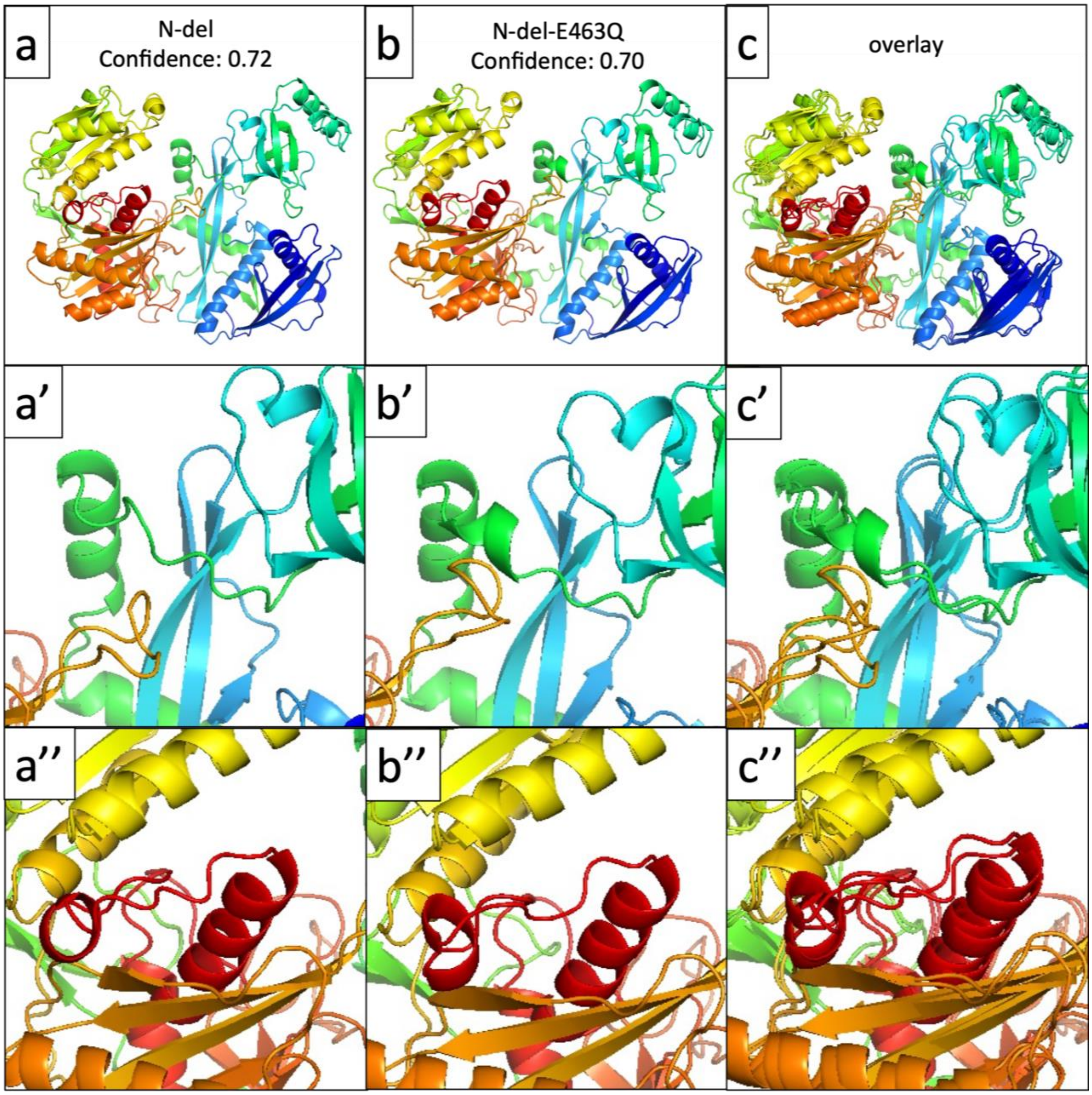


Figure 4-7. Structure modeling of N-del (a, a', and a''), N-del-D290N (b, b', and b''), and their overlay (c, c', and c''). Both predicted N-del and N-del-E463Q have reliable structures with confidence 0.72 and 0.70, respectively.

4.4 Conclusion and future work

By rational protein engineering design, I substituted the negatively-charged residues with neutral or positively-charged amino acids of NgAgo for increasing its solubility and activities. Although NgAgo-D290K, NgAgo-D290N, and NgAgo-E463Q have increased solubility, they lose the ability to induce cell death in bacteria. Structural modeling revealed the subtle but significant changes of their predicted protein structure, providing an explanation on how the mutations may break NgAgo's function.

Although the mutations disrupted protein function, they increased solubility. Failure to increase activity might be due to the neglected halophilic-specific information in the selection criteria. For example, as NgAgo is the only pAgo from halophile, it might have some adaptation specifically found in NgAgo. When I aligned NgAgo with other characterized pAgos, the halophilic adaptation is likely not conserved in other non-halophilic pAgos. As I selected the non-conserved residues, I might select the residues that are specifically adapted for halophilic pAgos, causing subsequent failure for improving NgAgo activity. Moreover, the mutational analysis software, SuSPect, might not be ideal. As SuSPect is a deep-learning-based software, which relies on the input of available pAgo structures, no available halophilic pAgo structure can be input to train SuSPect to identify the halophilic-specific adaptations. As SuSPect lacks halophilic pAgos input, I might choose residues sensitive to mutation in halophilic pAgos. In the future, the halophilic nature must be considered when using any software to identify potential sites for rational protein engineering design.

An alternative approach is to create a library containing the replacement of single negatively-charged residue with neutral or positively charged residue and use a directed evolution approach to select the library, given the enhanced solubility and activity of NgAgo variants would dominate the culture. With this approach, researchers can select NgAgo with improved solubility and activity, while NgAgo with enhanced solubility, not improved activity, will be selected out in one experiment. Other directed evolution approaches, such as PACE[101], OrthoRep[102], and VEGAS[103], can also be used to improve the activity of NgAgo or other pAgos in the future.

CHAPTER 5. A HIGHLY SENSITIVE POSITIVE SELECTION SCREEN FOR DEVELOPMENT OF GENE-EDITING TECHNOLOGIES

This chapter is adapted from material originally presented as part of the master thesis of Michael Mechikoff[104]. My specific contributions include new strains and plasmids and unpublished results in Section 5.4.3. This chapter is being prepared for publication.

5.1 Abstract

Prokaryote genomes encode diverse programmable CRISPR-associated DNA endonucleases with significant potential for biotechnology. However, these endonucleases differ significantly in their activity, specificity, and guide design preference, which all have implications for subsequent gene-editing tool development. While positive selection screens based on toxic proteins such as ccdB and barnase have been developed to evaluate such proteins, their high levels of toxicity make them unwieldy to use, and they are insensitive to candidates with modest but unique activities. Here, we develop and validate a more sensitive positive selection screen based on I-SceI to detect and enrich for programmable DNA endonuclease activity. I-SceI is a homing endonuclease that causes a deadly double-stranded break at an 18 base pair sequence inserted into an engineered *E. coli* genome. Cell death is rescued by candidate endonucleases designed to target and cure the I-SceI expression plasmid, thereby preventing cell death. By linking cell growth to programmable endonuclease activity, we amplify signal output to detect small differences in activity. We validated this assay with wild type SpCas9, xCas9, and eSpCas9 to capture endonuclease activity at short time scales not observed by traditional *in vitro* assays and demonstrated an ability to enrich for more active endonuclease variants from a mixed population. This system may be applied in high-throughput to rapidly characterize novel programmable endonucleases and be adapted for directed evolution of endonuclease function.

5.2 Introduction

Programmable endonucleases are important enzymes for modern biotechnology that can be targeted to cleave specific DNA sequences, enabling sequence modification[105]. CRISPR/Cas9 from *Streptococcus pyogenes* (SpCas9) is the most widely used and studied programmable endonucleases due to its ease of expression and high activity in a wide variety of hosts[106–111].

However, its use is restricted to regions adjacent to a defined NGG protospacer adjacent motif (PAM), which it modifies with variable efficiency dictated by the target sequence[112]. As a result, different variants of SpCas9 have been engineered to optimize its properties, and different Cas9 homologs or Cas systems have been developed to target alternative regions[17,113]. For example, xCas9, an engineered variant of SpCas9, has been developed to recognize a diversity of PAMs, including NG, GAA, and GAT, as opposed to NGG[114]. Another variant, eSpCas9, was developed with enhanced specificity, reducing the off-target activity of wildtype SpCas9[115]. While the availability of these SpCas9 variants, different Cas systems/homologs[17,113], and prokaryotic argonautes[27,60] greatly expand our ability to edit varying regions with increasing amounts of specificity, the rates, and thus efficiencies, of these enzymes, are challenging to measure.

Current methods of assessing activity rely on *in vitro* characterization, which involves protein expression, purification, and activity assays[116,117]. This method of *in vitro* characterization is laborious, time-consuming, and not suitable for high-throughput characterization of variants made via directed evolution. In addition, *in vitro* characterization excludes endonucleases that are difficult to be isolated with high purity and yield. All these limitations slow down the characterization of engineered endonucleases and newly discovered endonucleases, impeding the development of next-generation gene-editing technologies via programmable endonucleases.

In vivo characterization, however, overcomes the shortcomings of *in vitro* characterization by linking cell phenotypes to endonuclease activity, which allows users to rapidly characterize endonucleases without tedious protein purification. Ideally, an *in vivo* endonuclease activity assay should be sensitive and specific to resolve even low levels of endonuclease activity and limit the possibility for false negatives. The system should also link a visible phenotype to endonuclease activity, allowing quick identification of activity level. The signal from activity is amplified by linking enzymatic activity to cell survival. As bacterial cell growth is exponential, kinetics differences between enzymes are better highlighted. Since enzyme activity is linked to survival of the bacteria, cells harboring more active enzyme variants will be enriched in the population, which can be used for direction evolution of desirable traits such as activity and specificity, as previously demonstrated with ccdb-based selection systems[118,119]. In contrast, *in vitro* systems require constant monitoring to capture enzyme kinetics and, depending upon the turnover rate of the enzyme and amount of substrate, may only be able to identify the steady-state response of the enzyme, which is likely different *in vivo*.

Current *in vivo* assays are not ideal for assessing endonuclease enzymes due to their extreme sensitivity to environmental parameters. These assays rely on toxins encoded on a target plasmid that is cured by endonuclease activity to rescue growth. Thus, this type of system links cell survival to nuclease activity. However, common toxins used, Ccdb[120,121] and barnase[122], are highly lethal, and even low levels of leaky expression can cause cell death[120]. Reducing this leaky expression is possible but difficult to achieve. Moreover, the extreme toxicity of CcdB and Barnase prevents detection of low levels of endonuclease activity. Endonucleases with less activity may not sufficiently cleave the toxin-encoding plasmid, and the small amount of highly lethal toxins will kill the cell. Therefore, current systems falsely report a lack of endonuclease activity when activity levels are low. For the discovery and comparison of endonucleases with differing levels of activity, a less toxic yet effective and tunable system is desired.

Here, we develop a novel, *in vivo* endonuclease activity assay in *E. coli* that links cell survival to programmable endonuclease activity. Endonuclease activity rescues a lethal phenotype induced by the homing endonuclease I-SceI. As the assay output is relative growth, small differences in endonuclease activity are amplified by exponential cell growth for ease in detection. Moreover, an *in vivo* assay of this nature is faster than conventional *in vitro* assays, which requires lengthy protein expression and purification before evaluation. As a proof of concept, we validated this assay with wild type SpCas9, xCas9, and eSpCas9, demonstrating its versatility to work with an array of enzymes and rapidly quantify activity. Similarly, we demonstrate an ability to enrich for more active endonucleases confirming its potential for directed evolution.

5.3 Material and methods

5.3.1 Growth conditions

Strains were propagated in LB-Miller (LB) media at 37° C at 250 RPM unless otherwise noted. Super Optimal broth with Catabolite repression (SOC)[123] was used to recover cells after transformation. As appropriate, the antibiotic ampicillin, kanamycin, and tetracycline were supplemented at 100 µg/mL, 50 µg/mL, and 10 µg/mL, respectively. Inducers, L-arabinose, anhydrotetracycline (aTc), and rhamnose, were supplemented to a final concentration of 10mM, 200 ng/mL, and 0.2% (w/v), respectively, where indicated. Tetracycline, kanamycin, aTc, and LB are from Fisher Bioreagents, Fairlawn, NJ. Rhamnose is from Sigma-Aldrich, St. Louis, MO. Ampicillin and L-arabinose are from Acros Organics, New Jersey.

5.3.2 Strains and Plasmids

E. coli strains and plasmids used in this study are listed in Table 5-1. Primers used in construction are listed in Table 5-2. All molecular biology manipulations were carried out according to standard practices[123]. The *in vivo* endonuclease activity assays were conducted in *E. coli* strain KS165, which contains a DE3 cassette for T7 induction of endonuclease expression and an I-SceI recognition target integrated within the chromosome. *E. coli* MG1655 (DE3) was a gift from Prof. Kristala Prather (MIT, Cambridge, MA). To generate KS165, I-SceI recognition sites were integrated into the genome of MG1655 (DE3) using a standard recombineering protocol as follows[124]. *tetA* from pTKS/CS was amplified with primers containing I-SceI recognition sites and homology to the *nth* locus. The PCR product was subsequently isolated and integrated at the *nth* locus. Tetracycline-resistant clones were isolated on LB agar plates supplemented with tetracycline. Colonies were then checked via colony PCR, and Sanger sequenced to confirm correct integration at the *nth* locus.

Table 5-1. Strains and plasmids.

Name	Relevant Phenotype	Plasmid Origin of Replication	Source
Strain			
KS165	Tet^R, I-SceI recognition site x2	N/A	This study
Plasmids			
pEndoSceWT	Amp^R, ara inducible I-SceI	p15A	[121]
pEndoSce-D44S	Amp^R, ara inducible I-SceI	p15A	[121]
pColE1-ISceI	Amp^R, ara inducible I-SceI	ColE1	This study
pColE1-ISceI-D44S	Amp^R, ara inducible I-SceI	ColE1	This study
pFREE-sgRNA	Kan^R, rhamnose inducible guide targeting pColE1-ISceI, aTc inducible Cas9*	ColA	[125].The guides from the original pFree were replaced with a single guide targeting pColE1-ISceI.
pFREE-xCas9-sgRNA	Kan^R, rhamnose inducible guide targeting pColE1-ISceI, aTc inducible xCas9*	ColA	[114]. The Cas9 of pFree-sgRNA were replaced by xCas9.
pFREE-eSpCas9-sgRNA	Kan^R, rhamnose inducible guide targeting pColE1-ISceI, aTc inducible eSpCas9*	ColA	[115]. The Cas9 of pFree-sgRNA was replaced by eSpCas9.

*Source refers to the endonuclease

Table 5-2 Oligonucleotides. Restriction sites are in upper case.

Primers	Sequences 5’ → 3’	Template	Final product
pColE1-ISceI Fwd	ttttagatctATGAAAaacatcaaaaaaaaccaggtaatgaacctgg	pEndoSceWT	pColE1-ISceI
pColE1-ISceI Rev	ttttGAATTCttatttcaggaaagtttcggaggagatagtgttc		
Guide fragment 1 Fwd	tGGTCTCtggccacaattcagcaaattgtgaacatcatcacgttcatctttccctggttgccaatggcccattttcctgtcagtaacg	-	Guide fragment 1
Guide fragment 1 Rev	tGGTCTCtttcaaaacagcatagctctaaaacacgaccagtctaaaagcgcctgaattcgcgaccttctcgttactgacaggaaaatgg	-	
Guide fragment 2 Fwd	aGGTCTCttgaatggtcccaaaactgcggcgagcggtatcagctcactcaaagggttttagagctatgctgttttgaatggtcc	-	Guide fragment 2
Guide fragment 2 Rev	aGGTCTCctcgaggaccagactttaattaaaaaaaaaaccccgccctgtcaggggcggggttttttttttgttttgggaccattcaaaacagcat	-	
Guide targeting pColE1-ISceI Fwd	TTAT ctgcag tggccacaattcagcaa	Ligated guide fragment 1 and 2	Guide fragment 1 and 2
Guide targeting pColE1-ISceI Rev	TAGT ctcgaggaccagactttaattaaaaa		
pFREE-xCas9-sgRNA Backbone Fwd	tggtACTAGTgatcccatgttaccggtatccaag	pFREE-sgRNA	pFREE-sgRNA backbone
pFREE-xCas9-sgRNA Backbone Rev	tggtGTCGACctatcactgatagtgctcagtatttcttatc		
pFREE-xCas9-sgRNA Insert Fwd	GTCGACagatactgagcacagaaggagatatacatatggataagaaatactcaataggctt	pxCas9CR4	xCas9 insert

Table 5-2 continued

pFREE-xCas9-sgRNA Insert Rev	**ACTAGTttagtcacctcctagctgactca**		
pFREE-eSpCas9-sgRNA Backbone Fwd	tggtACTAGTgatcccatgttaccggtatccaag	pFREE-sgRNA	pFREE-sgRNA backbone
pFREE-eSpCas9-sgRNA Backbone Rev	tggtGTCGACctatcactgatagtgctcagtatttcttatc		
pFREE-eSpCas9-sgRNA Insert Fwd	GTCGACagatactgagcacagaaggagatatacatatggataagaaatactcaataggctt	pJSC114	eSpCas9 insert
pFREE-eSpCas9-sgRNA Insert Rev	ACTAGTttagtcacctcctagctgactca		
nth TetA Fwd	ctgctttccgctcaggcgaccgatgtcagtgttaataaggcgacggcgaatacggccccaaggtc	pTKS/CS	tetA gene cassette
nth TetA Rev	cggaaaatgtgcgtgtcgacagcaatagtcggccagccgaatgcagtgttcctaggtctagggcggc		

The I-SceI lethal plasmid, pColEI-ISceI, was constructed by amplifying I-SceI from pEndoSceWT [121] (a gift from Prof. Frederick Gimble, Purdue Biochemistry) and cloned into pBAD-mTagBFP2 (a gift from Prof. Mathew Tantama, Purdue Chemistry) at the BglII and EcoRI restriction sites. An inactive I-SceI catalytic mutant, pColEI-ISceI-D44S, was constructed with the same primers and restriction sites but used pEndoSce-D44S as a template[121]. pFREE was purchased from Addgene (Addgene plasmid # 92050)[125]. The guide targeting pColEI-ISceI was cloned via golden gate from two fragments containing the rhamnose promoter with an additional PstI site and the guide that was annealed beforehand. The ligated product was then cloned to pGEMT-Easy and sequenced for identity checking. The backbone of pFREE is amplified with primers with additional PstI site and then cloned with the fragment containing rhamnose promoter and guide, generating pFREE-sgRNA. The pFREE-xCas9-sgRNA and pFREE-eSpCas9-sgRNA plasmids were constructed by amplifying the pFREE-sgRNA backbone and Cas9 mutants codon-optimized for *E. coli* from pxCas9CR4 and pJSC114[116], respectively. SalI and SpeI (BcuI) restriction sites were added to the ends of backbone and insert for ligation. pJSC114 (Addgene plasmid # 101215) and pxCas9CR4 (Addgene plasmid # 111656) were purchased from Addgene. Unlike currently published variants of the pFREE system, the constructs here lack a guide for self-curing. All plasmid constructs were verified via Sanger sequencing. Unless otherwise noted, PCR amplifications were performed with Phusion High-Fidelity DNA Polymerase (NEB, Ipswich, MA) and oligonucleotides from Sigma-Aldrich (St. Louis, MO; Table 5-2). Restriction enzymes were obtained from Thermo Fisher Scientific (Waltham, MA), and T4 ligase was sourced from NEB.

5.3.3 I-SceI lethality assay

KS165 (*E. coli* MG1655 (DE3) *nth::tetA*) transformed with pColEI-ISceI or pEndoSceWT was inoculated into 3mL of LB with and without ampicillin and grown for 4 hours at 37°C at 250 RPM. After 4 hours, the cultures were diluted 100x in fresh media (with or without ampicillin as in the parent culture) and induced with L-arabinose (inducer for I-SceI). Negative uninduced control cultures were also created in parallel. The cultures were allowed to grow for 4 hours at 37°C, 250 RPM before the optical density at 600 nm was measured. Cultures were grown in triplicate.

5.3.4 Nuclease activity assay

KS165 was transformed with pColEI-ISceI and a Cas9 variant on one of pFREE, pFREE-xCas9, or pFREE-eSpCas9 and grown on LB with ampicillin, kanamycin, and tetracycline. Individual colonies from each transformation were grown in 3mL of LB with ampicillin, kanamycin, and tetracycline for 3 hours at 37°C at 250 RPM. Cultures were diluted 100x into LB with rhamnose and aTc (inducers for guide and Cas enzyme) and allowed to grow for at least 4 h, as indicated, at 37°C at 250 RPM. Cultures were then diluted 100x again into LB with arabinose (inducer for I-SceI) and allowed to grow for 4 hours at 37°C with agitation. OD_{600} readings were taken before and after this final 4-hour growth period. For each trial with inducer, a trial without inducer was run as a negative control.

5.3.5 Statistical analysis

Each experiment was repeated in triplicate or quadruplicate. Comparisons of recovered growth between induced and uninduced endonucleases were done using one-sided and two-sided unpaired t-tests. Data shown are the mean ± standard error.

5.3.6 Enrichment assay

To validate the ability of the assay to enrich or select for more active variants, we pooled equal volumes (100 ul) of cells diluted to ~OD 0.5 expressing Cas9, xCas9, or eSpCas9. Pooled cells were then diluted 100x into LB with rhamnose and aTc (inducers for guide and Cas enzyme), and allowed to grow for 4 h at 37°C at 250 RPM. Cultures were then diluted 100x again into LB with arabinose (inducer for I-SceI) and allowed to grow for 4 hours at 37°C with agitation. OD_{600} readings were taken before and after this final 4-hour growth period. For each trial with inducers (IPTG and rhamnose), a trial without inducer was run as a negative control. After 4 hours, the cells were pelleted to remove inducer and resuspended with fresh LB containing kanamycin for 16 hours overnight culture. The overnight culture was then miniprepped and subjected to next-generation sequencing. The frequency of each Cas9 variant was then counted based on the abundance of single nucleotide polymorphisms within the sequencing dataset and averaged with all the frequencies of single nucleotide polymorphisms for a given variant in each condition.

5.4 Results

5.4.1 Design and construction of selection system

Our positive selection system links targeted DNA endonuclease activity to cell survival in a quantitative way via a two-plasmid system in a modified *E. coli* MG1655 (DE3) host strain (Figure 5-1a). The first plasmid, a lethal plasmid, encodes a homing endonuclease that creates a lethal double-stranded DNA break (DSB) at a target site introduced in the chromosome of our modified strain (KS165). The enzyme, I-SceI, targets a large recognition site, TAGGGATAACAGGGTAAT, which was integrated at the *nth* locus via standard recombineering approaches [126]. The size of this recognition site mitigates off-target cleavage of the host chromosome due to the absence of similar, slightly mismatched sequences. Chromosomal double-stranded DNA breaks are inefficiently repaired in *E. coli* inhibiting cell replication and growth [127]. Thus, we hypothesize that I-SceI will generate a lethal double-stranded DNA break that inhibits cell growth only when induced with arabinose. To prevent unintended cell death via basal expression of I-SceI, the tightly controlled, arabinose-inducible P_{araBAD} promoter [128,129] was chosen to regulate expression of I-SceI (Figure 5-1b).

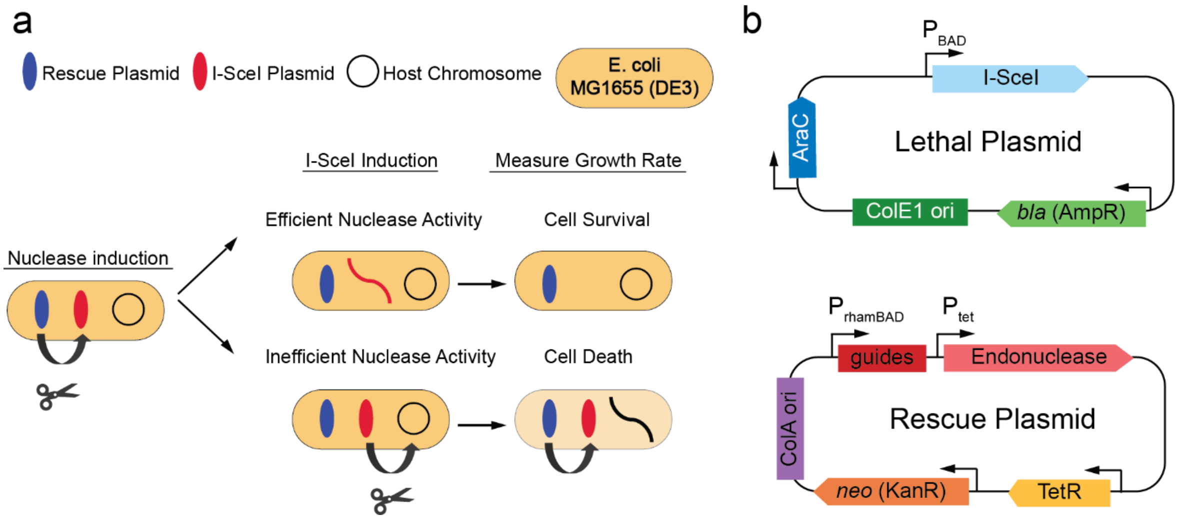


Figure 5-1. Design of *in vivo* assay system. a) Overview of the assay system. b) Plasmid designs used in the assay system. The lethal plasmid encodes for the homing endonuclease, I-SceI, which targets two sites in a modified *E. coli* MG1655 (DE3) genome. I-SceI is induced by arabinose and is on an ampicillin-resistant plasmid with a ColE1 origin of replication (~20 copies). The rescue plasmid encodes the endonuclease of choice under the control of a P_{Tet} promoter, inducible with aTc. The rescue plasmid also has corresponding endonuclease guides targeting major origins of replication, under the control of a rhamnose-inducible rhamBAD promoter. The rescue plasmid is kanamycin resistant with a ColA origin of replication (~20-40 copies).

The second plasmid serves as a rescue plasmid that encodes a programmable endonuclease targeted to the lethal plasmid, linearizing and curing it. The rescue plasmid is derived from a pFREE backbone[125], which expresses programmable endonucleases under the control of a P_{Tet} promoter that is induced by anhydrotetracycline (aTc) (Figure 5-1b). This modified backbone only contains guides targeting the pColE1-ISceI plasmid and cannot self-cure. Programmable endonuclease cleavage of the lethal plasmid is targeted via rhamnose-inducible guides for the ColE1 origin in the lethal plasmid[130]. Upon successful cleavage, the linearized lethal plasmid is rapidly degraded, rescuing growth[131]. Partial cleavage of the lethal plasmid would allow for fewer cells to escape cell death resulting in slower apparent growth. As cell growth is exponential, however, this growth-based output amplifies small differences in cleavage activity to give an exponential correlation between endonuclease activity and culture optical densities at later time points (Figure 5-2). As growth rates are finite, there will be an upper bound to the endonuclease activity that can be detected. Thus, the rescue plasmid was designed to also be self-curing by targeting its ColA origin with a separate rhamnose-inducible guide, limiting the amount of endonuclease that is expressed. This self-curing ability maintains assay sensitivity at low activity levels (poor endonuclease activity leads to longer expression times for endonuclease), while at high endonuclease activity expression is short allowing for detection of a wider range of activities.

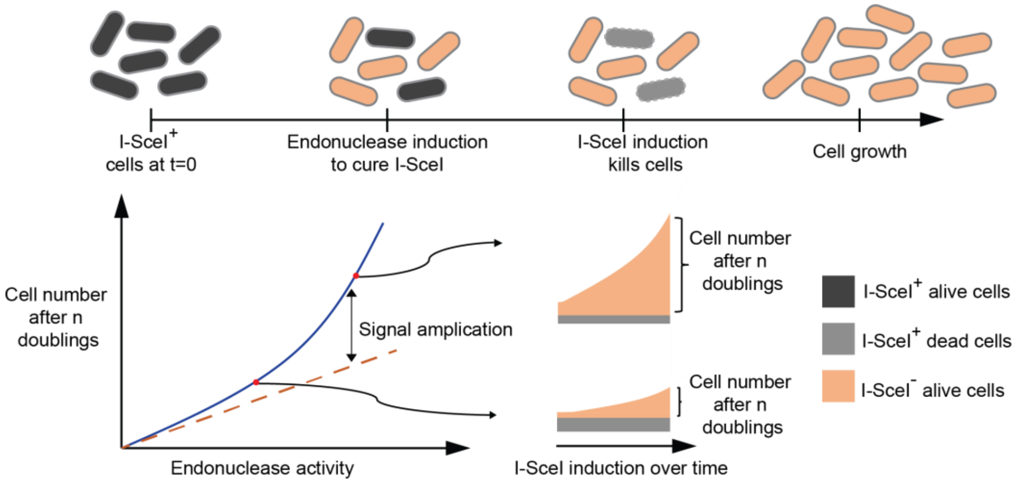

I-SceI^{+} cells at t=0
Endonuclease induction to cure I-SceI
I-SceI induction kills cells
Cell growth
Cell number after n doublings
Signal amplication
Endonuclease activity
Cell number after n doublings
Cell number after n doublings
I-SceI induction over time
I-SceI^{+} alive cells
I-SceI^{+} dead cells
I-SceI^{-} alive cells

Figure 5-2. *In vivo* assays exponentially amplify activity for increased sensitivity and detection. The exponential nature of the *in vivo* system (blue line) amplifies the signal of endonuclease activity compared to linear, *in vitro* activity assays (orange dashed line). Small differences in endonuclease activity will be magnified in an *in vivo* assay. Cells containing the I-SceI plasmid (I-SceI^{+}) will die off, leaving cells without the I-SceI plasmid (I-SceI^{-}) to replicate. Endonuclease activity is linked to cell survival because the endonuclease targets and cures the I-SceI plasmid, resulting in I-SceI^{-} cells.

5.4.2 Optimization of selection system sensitivity

We first validated the ability of I-SceI to be conditionally lethal using a pACYC vector with a low copy p15A origin of replication for the lethal plasmid. KS165 cells containing plasmid pEndoSceWT or pEndoSceD44S were plated on LB agar and induced with arabinose. As I-SceI was only induced on the plate with arabinose, cells experienced a brief period of growth before sufficient I-SceI had been expressed to arrest growth (Figure 5-3a). This resulted in a noticeable phenotypic difference in colony size between the induced and uninduced wild type I-SceI plates, indicating some I-SceI activity. Negative controls with a catalytically inactive mutant of I-SceI resulted in no phenotypic difference in cell size between the induced and uninduced cells. That is, I-SceI was indeed conditionally lethal in our cells.

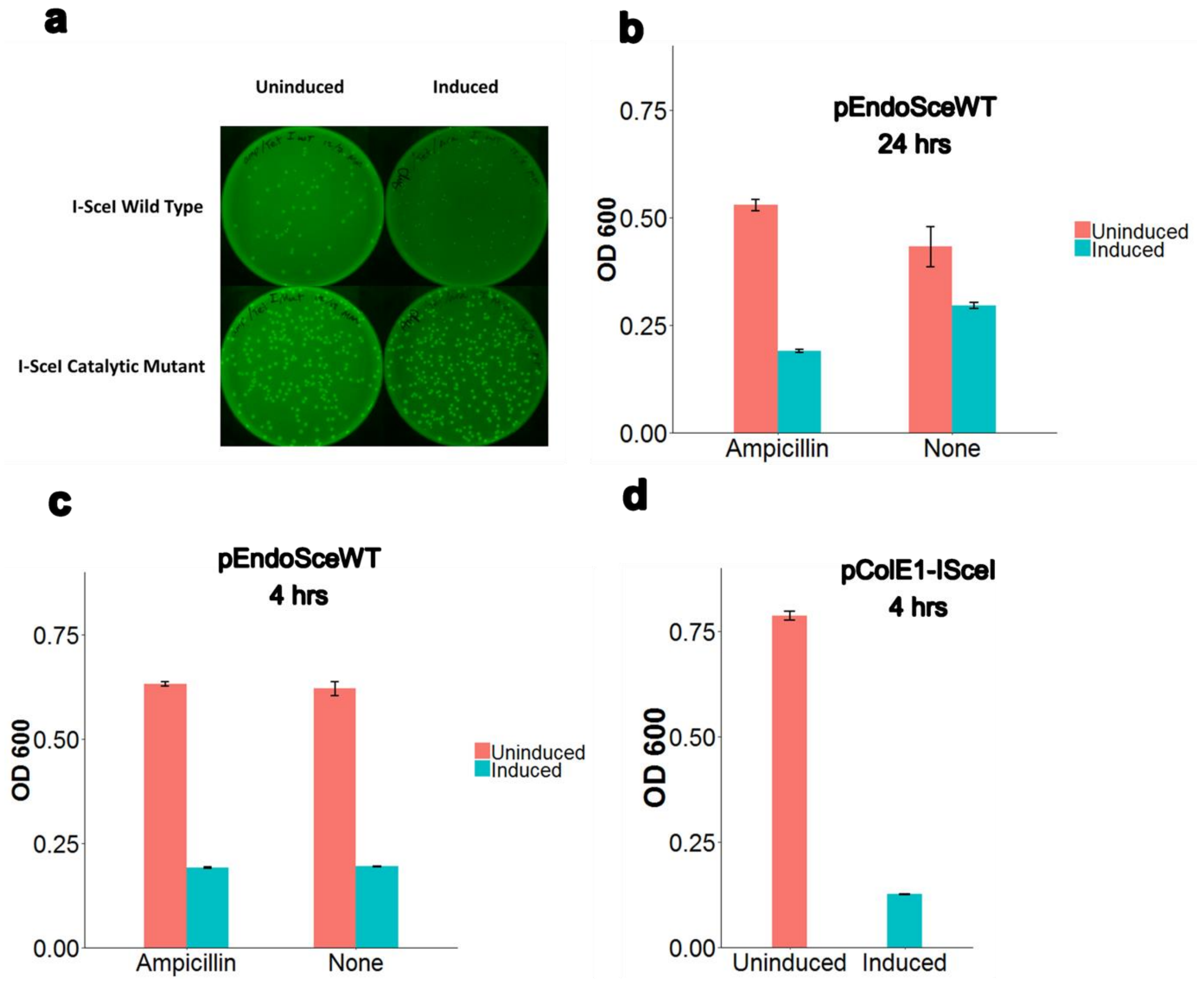


Figure 5-3. I-SceI is lethal in engineered KS 165. a) E. coli MG1655 (DE3) nth::tetA (KS165) expressing either wild type I-SceI or an inactive catalytic mutant. All cultures were able to form colonies after 16 hours at 37°C although those with the induced wild type I-SceI plasmid were smaller, suggesting I-SceI expression inhibits cell growth. b) Cells harbouring the pEndoSceWT or pColE1-ISceI plasmid were allowed to grow for 24 hours or 4 hours, followed by a 4 hour I-SceI induction period. Cultures were grown with and without ampicillin to test plasmid retention in the presence (or absence) of a selective pressure. Cells grown for 4 hours are equally sensitive with and without selective pressure while cells grown for 24 hours show less sensitivity to I-SceI without the selective pressure presumably due to spontaneous plasmid curing. Cells harbouring the higher copy number plasmid, pColE1-ISceI, increased sensitivity and exhibited more growth inhibition than pEndoSceWT when grown for 4 hours, followed by a 4 hour I-SceI induction period. NT – not tested.

To improve parallelizability and ease of detection we tested conditional lethality in liquid media by measuring the optical density of growing cultures (Figure 5-3b). We grew cultures for 24 h, to simulate potential endonuclease induction and expression from the rescue plasmid before diluting the cultures, inducing I-SceI for 4 h, and then measuring the optical density. In the presence of ampicillin, cell numbers (or OD_{600}) were reduced by 64.3% by I-SceI from pEndoSceWT, confirming the conditional lethality observed on plate-based assays. When fully implemented, the activity assay rescues cell growth by curing the I-SceI plasmid and its selection marker. That is, the cell will lose its ampicillin resistance. Thus, to prevent systematic bias in assay output, conditional lethality must also function in the absence of any antibiotic. We tested the pEndoSceWT plasmid's ability to reduce cell growth in the absence of antibiotic for 24 hours and found it only reduced growth by 31.8%. Because the pEndoSceWT plasmid has a low copy, p15A origin of replication (~10-12 copies/cell)[132], we hypothesized that the cells were spontaneously curing themselves of the plasmid over the 24 hour time period in the absence of any selection pressure. Thus, we decreased the growth time from 24 hours to 4 hours, which should allow for ample expression and curing via any tested endonucleases, and test whether the cells would retain the conditionally lethal phenotype. Counterintuitively, the ODs of these cultures were higher than those grown for 24 h before dilution. One potential explanation is that cells grown for 4 h before dilution have not yet experienced stationary phase. Thus, upon dilution they continue their rapid exponential growth, unlike the cells grown for 24 h before dilution, which may experience a longer lag phase before exponential growth. Nonetheless, the cultures with and without selective pressure (ampicillin) were able to reduce cell growth to a similar extent, 69.7% and 68.6%, respectively. This result suggests that a 4 hour growth period prior to I-SceI induction is sufficient to preserve plasmid retention within the cells.

To increase the sensitivity of the assay, we increased the copy number of the lethal plasmid. However, this copy number must be lower than that of the rescue plasmid to ensure sufficient endonuclease for activity detection. As the rescue plasmid has a ColA origin of replication which generates 20-40 plasmid copies per cell[132], we chose a pET vector with a ColE1 origin of replication (~15-20 copies/cell)[132], which is compatible with ColA[132,133]. The generated pColE1-ISceI lethal plasmid significantly reduced cell growth by 84.1%, increasing the sensitivity of the system due to the higher copy number.

5.4.3 Validation of targeted endonuclease activity

The ability of the system to function as an activity assay for endonuclease activity was validated using the popular wild type SpCas9 enzyme[134]. Cells transformed with a rescue plasmid encoding SpCas9 and the lethal plasmid were grown without any inducer to establish a baseline optical density without induced lethality. The cells were grown for 4 hours, diluted 100x, then grown for another four hours, resulting in an optical density of 2.064 ± 0.028. Expression of SpCas9 did not inhibit cell growth (OD = 2.006 ± 0.117). In the presence of uninduced rescue plasmid, I-SceI induction resulted in conditional lethality, reducing optical density to 1.133 ± 0.341. However, pre-expression of SpCas9 targeted to the I-SceI plasmid resulted in optical densities of 2.149 ± 0.167, rescuing 97.8% of wildtype growth. This difference is statistically significant (unpaired t-test, $p<0.05$, $n=3$). This result suggests that the assay successfully links endonuclease activity to cell growth allowing for an easy readout of endonuclease function as cells with induced SpCas9 and I-SceI showed an increase in cellular growth compared to the growth of cells with only I-SceI induction. A final control was added, which only included the host strain, void of any plasmid, to test the effect of metabolic burden due to plasmid maintenance and heterologous gene expression. With an OD of 2.181 ± 0.150, the effect of metabolic burden from plasmids did not meaningfully contribute to decreased growth (Figure 5-4).

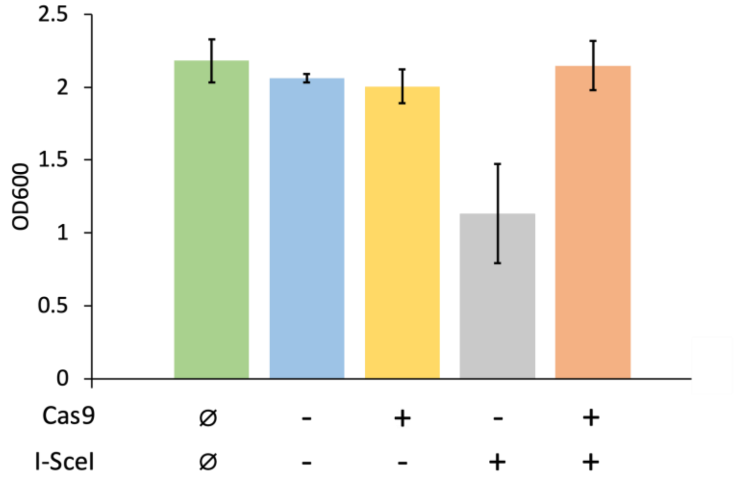


Figure 5-4. Wildtype Cas9 was tested in the endonuclease activity assay. Wildtype Cas9 was previously induced for four hours, then diluted 100x. Time after induction refers to the amount of time the cells were given to grow after I-SceI induction. + and – refer to induced or uninduced. Data shown are mean +/- standard deviation of four replicates.

We then evaluated the ability of the assay to capture the relative activities of different Cas9 variants. Cells were grown with the lethal I-SceI plasmid and a rescue plasmid containing one of three SpCas9 variants. Induction of each Cas9 mutant for 4 h prior to I-SceI was able to recover some growth relative to cells where the Cas9 mutants remain uninduced (Figures 5-5 and 5-6). However, growth recovery was different for each endonuclease tested. With this system, and under the conditions tested, wildtype Cas9 had the greatest growth recovery at 97.8%, followed by xCas9 at 27.5%, and eSpCas9 at 9.3%. (Figure 5-7). These data suggest that wildtype Cas9 has higher activity than xCas9 and eSpCas9 variants. This relative ranking of endonuclease activity is consistent with previous *in vitro* studies[119], however, the response is not linear with endonuclease activity due to the exponential signal amplification of cell growth. Nonetheless, our assay accurately ranked the activity of each endonuclease and may serve as a rapid method for the screening of relative endonuclease activity.

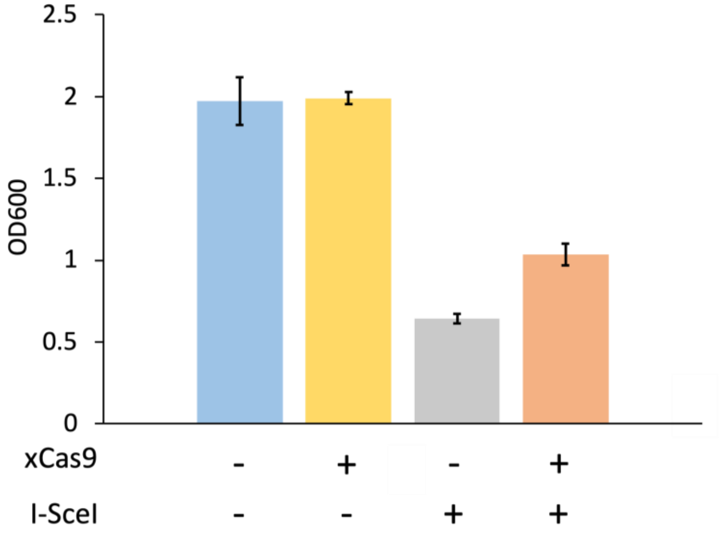


Figure 5-5. xCas9 was tested in the endonuclease activity assay. xCas9 was previously induced for four hours, then diluted 100x. Time after induction refers to the amount of time the cells were given to grow after I-SceI induction. + and – refer to induced or uninduced. Data shown are mean +/- standard deviation of four replicates.

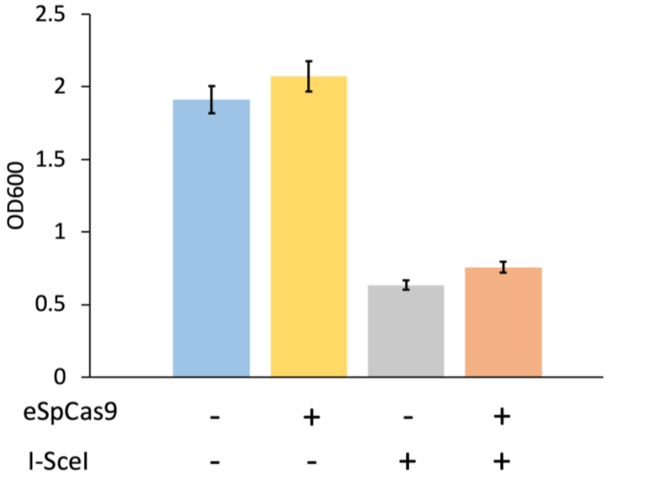


Figure 5-6. eSpCas9 was tested in the endonuclease activity assay. eSpCas9 was previously induced for four hours, then diluted 100x. Time after induction refers to the amount of time the cells were given to grow after I-SceI induction. + and – refer to induced or uninduced. Data shown are mean +/- standard deviation of four replicates.

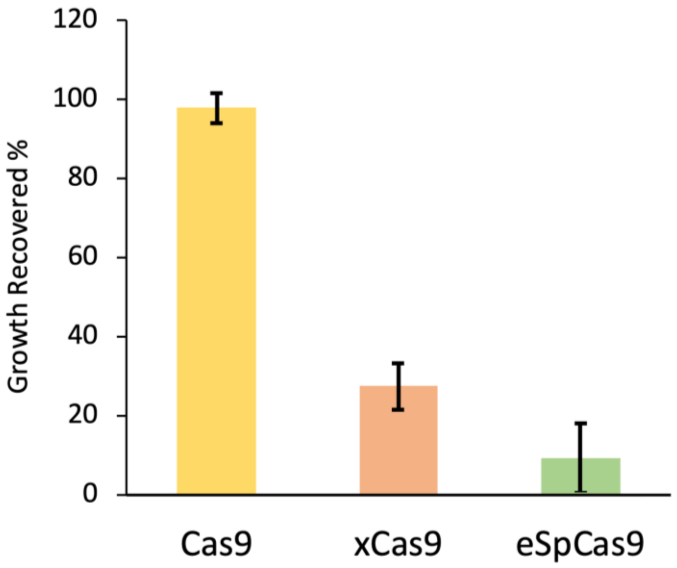


Figure 5-7. Comparison of commonly used endonucleases. Each endonuclease was given 4 h of induction and targeted toward the lethal plasmid, after which I-SceI of the lethal plasmid was induced to cause cell death. Data shown are mean +/- standard deviation of four replicates.

Given that the different rescuing capabilities of Cas9 variants, we further tested if highly active Cas9 can be enriched in a mixed culture carrying different Cas9 variants, mimicking the selection scenario for directed evolution. The cells with three Cas9 variants were mixed with equal volume of each culture and approximately equal OD before Cas9 and I-SceI induction (see method-Enrichment assay). The cells were collected, miniprepped, and subjected to next-generation sequencing to identify the variants based on their single nucleotide polymorphisms. Without Cas9 and I-SceI induction, the percentage of each Cas9 variant is approximately 33% as expected. However, after 4 h of induction, Cas9 dominated the culture at 63.4% in comparison to 22.3% and 14.3% for xCas9, and eSpCas9, respectively (Figure 5-8). Thus, our system is effective at not only measuring DNA endonuclease activity but for selecting for more active variants.

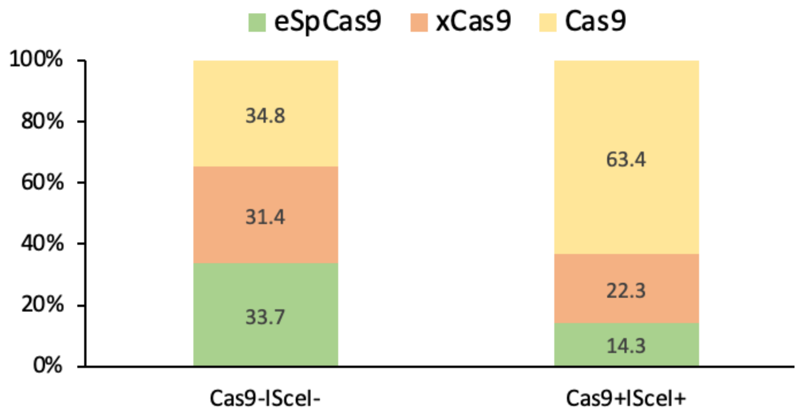


Figure 5-8. Highly active Cas9 dominates the pooled bacteria carrying Cas9 variants. Bacteria were collected and subjected to next-generation sequencing to identify the percentage of Cas9 variants with and without induction of Cas9 and I-SceI. Data shown are mean of four replicates.

5.5 Discussion

We have developed an *in vivo* endonuclease activity assay designed to identify and compare enzymes for DNA-cleavage activity. Our system makes use of the homing endonuclease, I-SceI, which recognizes and cleaves an engineered site in a modified *E. coli* genome, causing cell death. I-SceI was chosen as a lethal factor due to its low toxicity when uninduced. In contrast, previous *in vivo* endonuclease activity assays used the toxic proteins barnase and CcdB as their lethal factors. Basal expression of barnase in a low-copy plasmid with a p15A origin was reportedly lethal in *E. coli* cells, even under the tight regulation of the araBAD promoter [122]. Therefore, amber nonsense codons were introduced to cage the toxicity of barnase. While successful, this technique necessitates the use of amber suppressor tRNAs for proper functionality[135], which increases system complexity with the need for another inducer and special hosts. To combat this, a CcdB version of an endonuclease activity assay was created, removing the suppressor tRNA requirement. However, CcdB was also reported to be toxic under basal expression from the same araBAD promoter, though the copy number was much higher (100-300 copies per cell)[120]. Therefore, the authors engineered the ribosome binding site for the *ccdB* gene to thwart some of its toxicity. Unlike barnase and CcdB, our I-SceI expression system is simpler and far easier to tune with no overt toxicity issues.

Originally in a pACYC vector with a low copy p15A origin of replication, I-SceI was able to reduce cell growth by 64.3% after a 24 h growth period, in the presence of a selective pressure to retain the plasmid (Figure 5-3b). However, the assay includes a growth period to induce the tested endonuclease, at which time selective pressures cannot be incorporated. After a 24 h growth period without selective pressure, the cells were unable to fully retain the plasmid and less lethality was observed (Figure 5-3b). To mitigate this issue, we decreased the growth time to 4 hours and saw similar reductions in growth regardless of the presence of selection pressure (Figure 5-3b). We also sought to increase the sensitivity of the assay by increasing the plasmid copy number from 10-12 to 15-20 copies per cell. This new plasmid, pColE1-ISceI, was able to reduce cell growth by 84.1% after a 4 hour growth period (Figure 5-3d). In the pColE1-ISceI plasmid, I-SceI is under the control of the tightly regulated P_{BAD} promoter. This strict regulation alleviated leaky expression and prevented unwanted cell death: uninduced cells harboring the pColE1-ISceI plasmid had similar growth rates to cells not harboring the plasmid (Figure 5-4).

With the lethal plasmid properly inducing cell death, we next had to design the rescue plasmid. We modified the pFREE plasmid that was used for plasmid curing with a multiple cloning site incorporating SalI and SpeI restriction sites to allow for easy swapping of the endonuclease. The pFREE plasmid also expresses guides targeting most origins of replication, including ColE1 in the lethal plasmid and its own ColA origin to increase the range of endonuclease activities that can be detected. Critically, the copy number for ColA is higher than the copy number for ColE1, which allows for enough expression of endonuclease to sufficiently cleave each copy of lethal plasmid. By selecting a plasmid that has a higher copy number origin than the lethal plasmid, assay sensitivity is increased as more endonuclease is expressed to increase the likelihood that sufficient endonuclease activity will be generated to cure the lethal I-SceI plasmid.

To confirm our system was robust with respect to a number of endonucleases and could distinguish between differing activity levels, we tested our system with three endonucleases, Cas9, xCas9, and eSpCas9. Of the three, wild-type Cas9 displayed the highest level of endonuclease activity, followed by xCas9 and eSpCas9. We hypothesize that the mutations made to develop eSpCas9 and xCas9 affect the kinetics of the enzymes, slowing their activity levels, in agreement with published findings[136]. Meanwhile wild type spCas9 is approximately more than 5x more active at TGG PAM targets than xCas9[117]. *in vitro* assays require constant monitoring, especially at timescales of seconds, to detect differences in enzymatic activity[119]. Our *in vivo* assay, in contrast, was able to accurately capture the ranked order of endonuclease activity in a simple readout without enzyme purification. More importantly, the assay was sensitive to the relatively low levels eSpCas9 activity and amplified the small differences in xCas9 activity from wild type spCas9 because cellular growth is exponential and slight changes in endonuclease kinetics will result in large differences in total growth over time. For example, the difference between eSpCas9 and wild-type Cas9 activity levels *in vitro* may only be a percentage or two after 1 minute but would be far greater in an *in vivo* assay (Figure 5-7). Therefore, differences in activity levels between endonucleases that would otherwise be undetectable are amplified and obvious over time.

Moreover, *in vivo* activity assays require much less time than traditional *in vitro* methods. For example, *in vitro* assays require cell lysis, protein purification, protein quantification, activity assays, and finally gel electrophoresis. *In vivo* systems, on the other hand, negate cell lysis, protein purification/quantification, and gel electrophoresis. Instead, *in vivo* assays can skip from cell transformation to the activity assay directly. Typically, *in vivo* systems can be performed within 24 hours start to finish while in vitro methods take days.

Our system can also be used as a positive selection screen in directed evolution approaches to enhance enzymatic activity of endonucleases. Endonuclease activity will still be retained when the endonuclease cleaves the lethal plasmid and rescues cell growth. However, as more active mutants result in faster growth, those high performing variants will begin to dominate mixed cultures of endonuclease variants facilitating recovery, characterization and subsequent round of directed evolution (Figure 5-8). In a similar fashion, the guide sequences may also be modified to evaluate the activity as a function of target sequence composition.

5.6 Conclusion

In conclusion, our novel endonuclease activity assay can identify, rank endonuclease activity, and be used for directed evolution. By using a tightly controlled two-plasmid system, our assay positively links cell growth to endonuclease activity and serves as a ranking system among different endonucleases. The *in vivo* nature of our system drastically decreases the time required to identify endonuclease activity compared to *in vitro* methods. Moreover, the endonuclease activity signal is amplified in our *in vivo* assay compared to *in vitro* methods, facilitating the clear discrimination between endonuclease activity levels. This system may be parallelized for high-throughput screening to rapidly characterize novel programmable endonucleases including prokaryotic Argonautes[129,137–139] and can be adapted for directed evolution of endonuclease function.

ACKNOWLEDGEMENTS

We thank Frederick Gimble for his generous donation of the pEndoSceWT and pEndoSce-D44S plasmids. We are grateful to Mathew Tantama for pBAD-mTagBFP2. We also thank Kristala Prather for her generosity in gifting us the MG1655 (DE3) *E. coli* strain.

FUNDING

This work was supported by the Ralph W. and Grace M. Showalter Trust (Grant # 41000622); the Purdue Research Foundation (Grant # 60000025); and with startup funds from Purdue's Colleges of Engineering and Agriculture. Funding for open access charge was provided from startup funds.

CONFLICT OF INTEREST

None to declare

CHAPTER 6. CONCLUSIONS & FUTURE WORK

6.1 Conclusions

Gene-editing with current available CRISPR/Cas systems has changed and will continue to change many aspects of our life. As CRISPR/Cas system requires sequence-specific motifs to edit DNA, its impact is limited in organisms having GC-biased genome/genomic regions. Scientists are looking for more flexible gene-editors, and pAgos have been put forward into the spotlight as a compelling alternative. NgAgo is the first identified mesophilic pAgo claimed to have the gene-editing ability at temperatures relevant to humans or organisms relevant to biotechnology. However, the paper was retracted due to failure to reproduce key findings in other labs worldwide. Although subsequent studies showed that refolded NgAgo cleaves RNA, mutation of catalytic tetrad didn't abolish the cleavage activity, raising the possibilities of RNaseH contamination, which cleaves RNA in DNA/RNA hybrid. It should be noted that cleaving RNA doesn't exclude the possibilities of cleaving DNA as some pAgos have been shown to cut both DNA and RNA. Nonetheless, how NgAgo works remain unknown. This thesis was then designed to answer some critical questions for NgAgo. First, I would like to know if NgAgo cuts DNA as there is an unsettled debate in the literature. Second, if it cuts, what domains are essential for the cleavage. Third, can we repurpose NgAgo for gene-editing in bacteria?

To answer the first question, the protein structure of NgAgo was predicted, and the structure was used to do a structural alignment to overcome the low identity issue of conventional sequence alignment. Besides the conserved four canonical domains, NgAgo also has an intact catalytic tetrad, suggesting the presence of nucleic acid cleavage activity. Meanwhile, structure alignment also identified a new repA domain at the N-terminal of NgAgo. Phylogenetic analysis further showed that this repA domain is conserved in halophilic pAgos, suggesting functional adaptation in high-salt environments.

To get functional NgAgo proteins, the soluble fraction of NgAgo variants expressed in bacteria was purified. Alternatively, a bacterial cell-free system was used to produce NgAgo variants. Both soluble NgAgo and cell-free-system-produced NgAgo cleaved DNA. However, mutation of the catalytic tetrad of NgAgo did not abolish its function, similar to the RNA-cleaving study. Deleting the repA domain in a combination of tetrad mutation completely inactivated the DNA-cleaving ability, but the deletion of the repA domain does not abolish NgAgo activity. This true catalytically dead mutant indicates that DNA cleavage indeed depends on NgAgo and not due to other DNA endonuclease contamination. More importantly, this is the first pAgo showing that abolishing DNA-cleaving activity needs to be achieved by inactivating both repA and catalytic domains. Although the actual mechanism of the repA domain participating in the cleavage activity remains unknown, this study highlighted the unique feature of NgAgo compared to other known pAgos.

For function validation, NgAgo is able to induce cell death when targeting a nonessential pseudogene on a plasmid, indicating DNA cleavage in the host. Individual repA deletion and PIWI mutation lose the ability to induce cell death, indicating repA and PIWI domains are required for programmable DNA cleavage in a bacterial host. Similarly, NgAgo but not repA deletion or PIWI mutation enhances homologous recombination, achieving gene-editing application. These lines of evidence indicate that both repA and PIWI domains are required for the full function in the host.

Collectively, these experiments showed that NgAgo is indeed a DNA endonuclease. As the halophilic nature of NgAgo was considered, a new repA domain was identified, and a true catalytic mutant was made. These new features demonstrate the unique catalytic activity of NgAgo, providing compelling evidence and new knowledge in the literature for gene-editing tool development.

As the majority of NgAgo proteins are misfolded when expressing in *E.coli*, rational protein engineering design was used to increase solubility and activity by replacing negatively-charged residues with neutral or positively-charged residues on the protein surface. Through a series of selections, ten mutants, including NgAgo-E249Q, NgAgo-E249R, NgAgo-E252Q, NgAgo-E252R, NgAgo-D290N, NgAgo-D290K, NgAgo-E463Q, NgAgo-D533N, NgAgo-D537N, and NgAgo-E550Q, have been made to test if they have increased solubility. Three mutants, including NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q, were demonstrated to have increased solubility. Subsequent activity testing, however, showed that these three mutants lose their ability to induce DNA break, causing cell death in *E.coli*. Structural predictions of NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q revealed subtle but significant changes in the proteins, suggesting potential hypotheses regarding how these mutations break the NgAgo function.

Given that rational protein engineering design is limited by the knowledge of NgAgo, directed evolution was used for subsequent optimization. However, the toxin of current selection system, ccdb, was too toxic, as indicated by the decreased size of the colony. I then optimized the selection with another toxin, I-SceI, which induces DNA break to the host at specific sites, leading to cell death. Active endonucleases will be able to rescue the lethal effect by cleaving the I-SceI plasmid. As a prove of concept, my colleague and I showed that CRISPR/Cas9 rescued the lethal impact by I-SceI. Further experiments showed that wildtype Cas9 has higher activity compared to xCas9 and eSpCas9, similar to the result in the literature. These lines of evidence indicates that our system is reliable and able to be used as a selection system for optimizing endonuclease activity.

6.2 Future work

NgAgo is a unique pAgo among characterized pAgos. Besides having the repA domain, NgAgo is also isolated from a halophile. As I showed that halophilic pAgos also have the repA domain, this halophilic-specific feature may be an adaptation for pAgos to work in high-salt environments. Moreover, the catalytic activity of NgAgo is very different from all the known pAgos. Specifically, a mutation in the PIWI domain can still cut DNA with the help of the repA domain. However, how it works remains unclear. Future work may focus on getting the protein structure of NgAgo to understand mechanistic.

From a technological development perspective, it is needed to increase the solubility of NgAgo to improve its activity. Instead of rational design, future work may deploy directed evolution for improving its activity. As directed evolution can link the functionality of the pAgos to the survival of the host, it may be more accessible to future development as we don’t have much information about halophilic pAgos.

For guide and target nucleic acid cleavage, only experiments can figure out the preference of pAgos. Having software to reliably predict the guide and target nucleic of interested pAgos target will help researchers to identify their interested pAgos, accelerating the process for characterization.

For pAgos characterization, synthetic biology can offer many avenues for high-throughput characterization. For example, the cell-free system demonstrated in this study can be a prototype for future pAgos characterization. Given that this method does not involve time-consuming protein expression and purification, a cell-free system will accelerate the characterization of pAgos for gene-editing tool development. Moreover, the survival assay and gene-editing assays developed in this study can also serve as function validation methods in bacteria.

Like the CRISPR system, pAgos also have off-target activity via a “DNA chopping” mechanism. From a gene-editing tool development perspective, it would be valuable to develop pAgos without DNA chopping ability. Indeed, a mutational study has identified a mutation of MjAgo (D438P) at the MID domain that significantly reduced the DNA-chopping ability and guide-dependent cleavage activity[75]. Additional studies are needed to find mutants that lose DNA-chopping ability while retaining guide-dependent activity.

For applications in eukaryotic systems, pAgos may need to have accompanied by appropriate helicase as pAgos do not have a helicase domain. Indeed, *in vitro* experiment showed that adding helicase and single-stranded DNA binding domain improve the function of pAgo, overcoming the high-GC content blockage for cleaving. Future work may need to screen for appropriate helicase for pAgos to work in the eukaryotic system.

Lastly, directed evolution can be used to for activity improvement. Given that rational protein engineering design is limited the knowledge of pAgo, directed evolution is more rapid for improving activity without prior knowledge, accelerating the improvement process. Our optimization of the selection system by replacing the ccdb toxin to I-SceI can serve as a platform for directed evolution.

Despite all the challenges, new methodologies will improve the efficiency for characterizing and optimizing pAgo activity, providing revenues for understanding the pAgo activity and shed lights on the road for pAgo-based gene-editing tool development.

APPENDIX I. IMPORTANT LESSIONS FORM CHARACTERIZING NGAGO

The salt concentration is key to NgAgo activities

Unlike other pAgos, NgAgo is very sensitive to salt (NaCl) concentration, similar to the other halophilic proteins. To get functional NgAgo protein in *E.coli* bacteria, getting soluble NgAgo is the key issue that needs to be addressed. Based on my experience, *E.coli* needs to be cultured in Miller LB, instead of Lennox LB, to maximize soluble NgAgo expression. This makes sense as Miller LB (10g/L) has double the amount of NaCl than Lennox LB (5g/L). In my experience, we can get maximum soluble NgAgo with 0.05-0.1 mM working concentration of IPTG induction in Miller LB. I used 0.1mM IPTG for preparing soluble NgAgo and characterizing NgAgo in bacteria.

Guide storage

During my experiments, I also noticed that the guide function decreases over time, probably due to degradation and/or hydrolysis of the 5' phosphate group. I usually aliquot the guide after resuspension to prevent frequent freeze/thaw cycles.

Identification of guide-independent mutant of NgAgo

As DNA chopping of pAgos will impact the specificity, I tried to look for a mutant that abolishes DNA chopping ability of NgAgo. MjAgo has two modes of structure, including one without a guide and one with a guide. The one without guide chops DNA nonspecifically and gets guides to form the second mode, executing specific DNA cleavage. Mutational analysis identified the key residue (D438P) that locks the protein in the structure with guides, which greatly reduces the DNA chopping ability but also compromises its guide-dependent activity[75]. To find the corresponding residue in NgAgo, I overlayed the predicted NgAgo structure with the MjAgo structure and identified potential key residues including E588, E592, E598, D601, and E602. They were mutated to alanine to remove the negatively-charged residue or mutated to proline to remove negatively-charged residue and disrupt alpha helix structure. As this testing requires screening many mutants, I devised a more rapid way to test their DNA chopping ability. I expressed NgAgo variants, extracted the plasmid, linearized the plasmid, and ran a gel to see their DNA integrity. With protein control, BFP, the plasmid is intact. However, NgAgo expression degrades the plasmid, as evidenced by the streaking and the significantly reduced amount of intact DNA. The pattern is similar with D663A.D738A, N-del, and repA while N-del/D663A.D738A has significantly reduced the activity. This result is similar to the cell-free-based *in vitro* data, indicating the reliability of the assay.

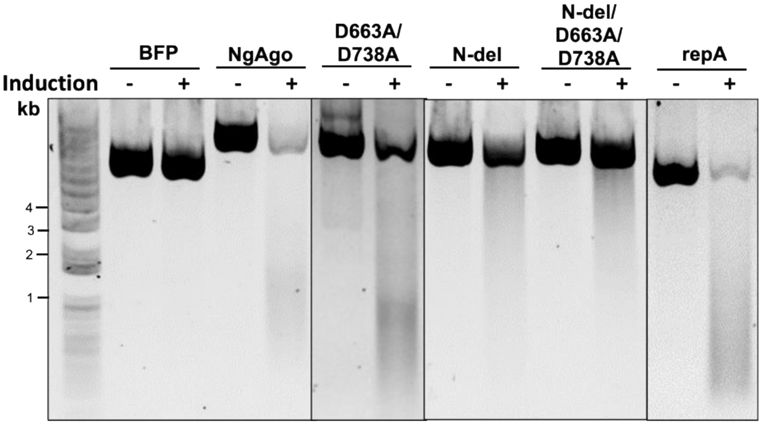


Figure 6-1. DNA chopping ability of NgAgo variants. BFP protein and NgAgo variants including wildtype, D663A/D738A, N-del, N-del/D663A/D738A, and repA were induced in BL21 (DE3) for 4 hours, miniprepped, linearized with XhoI, and checked for DNA integrity. +/- represents induced/uninduced.

Then, I repeated the experiment with guide-independent mutant candidates. Compared to the N-del, all the mutants have more intact DNA. Among them, N-del/E598A and N-del/E598P are very promising candidates that abolished guide-independent activity. Future experiment should test their guide-dependent cleavage activity.

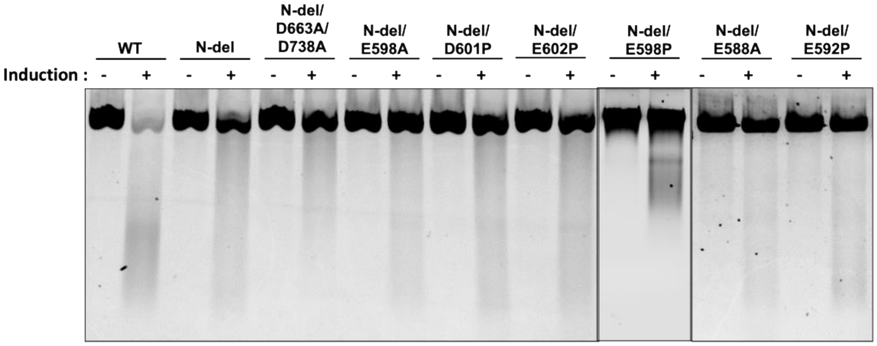


Figure 6-2 DNA chopping ability of NgAgo guide-independent mutants. BFP protein and NgAgo variants including wildtype, N-del, N-del/D663A/D738A, N-del/E598A, N-del/D601P, N-del/E602P, N-del/E598P, N-del/E588A, and N-del/E592P were induced in BL21 (DE3) for 4 hours, miniprepped, linearized with XhoI, and checked for DNA integrity.

Modeling of NgAgo with repA domain

Given that the repA domain is new to pAgos family, the structure prediction with repA yields low confidence less than 0.7, which is unreliable due to not modifying the structure fitting. Here, I presented the structure prediction of NgAgo variants with intact repA. Since the confidences of all predicted structures were less than 0.7, I predicted the structure without repA domain with more reliable confidences (Figure 4-5, 4.6, and 4-7).

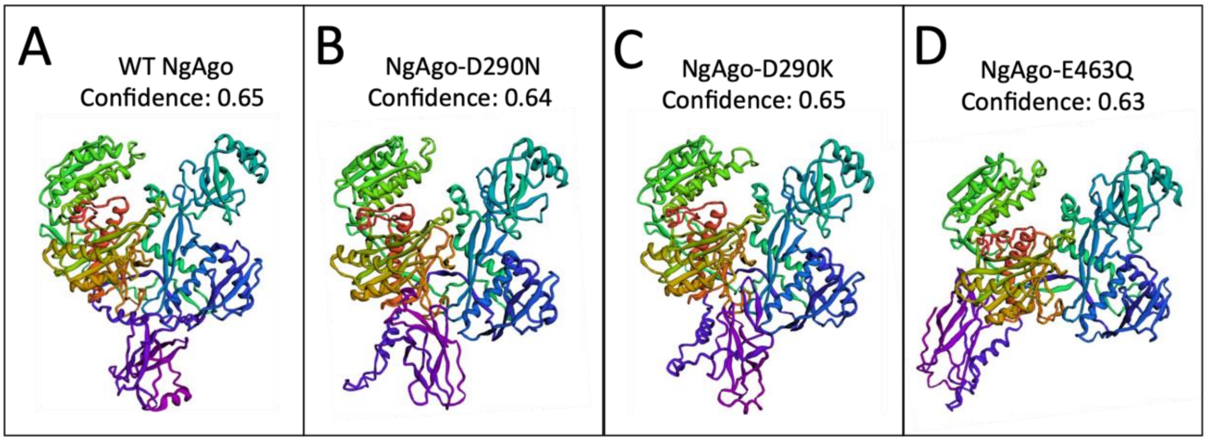


Figure 6-3. Structure modeling of NgAgo variants with intact repA domain. Wildtype, NgAgo-D290N, NgAgo-D290K, and NgAgo-E463Q were predicted with confidence, 0.65, 0.64, 0.65, and 0.63, respectively.

APPENDIX II. COPYRIGHT PERMISSIONS

Figure 2-1 was adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Structural & Molecular Biology (2014)[28].

Figure 2-2 was adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Microbiology (2018)[24].

Figure 2-3 was adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Reviews Microbiology[24].

Figure 2-4 was reprinted from Chemistry & Biology, Volume 22/Issue 12, Ortega et al., Halophilic Protein Adaptation Results from Synergistic Residue-Ion Interactions in the Folded and Unfolded States, Pages 1597-1607, Copyright (2015), with permission from Elsevier.

Figure 2-5 was reprinted from Current Opinion in Microbiology, Volume 25, Shiladitya DasSarma and Priya DasSarma, Halophiles and their enzymes: negativity put to good use, Pages 120-126, Copyright (2015), with permission from Elsevier.

Figure 2-7 was reprinted from Biophysical Journal, Volume 83, Issue 5, Gan et al., Analysis of Protein Sequence/Structure Similarity Relationships, Pages 2781-2791, Copyright (2002), with permission from Elsevier.

Figure 2-8 was adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, Nature Protocols (2015)[54].

REFERENCES

1. Wu, Y. *et al.* Correction of a genetic disease by CRISPR-Cas9-mediated gene editing in mouse spermatogonial stem cells. *Cell Res.* **25**, 67 (2015).

2. Shi, J. *et al.* ARGOS 8 variants generated by CRISPR-Cas9 improve maize grain yield under field drought stress conditions. *Plant Biotechnol. J.* **15**, 207–216 (2017).

3. Lee, J. S., Grav, L. M., Lewis, N. E. & Faustrup Kildegaard, H. CRISPR/Cas9-mediated genome engineering of CHO cell factories: application and perspectives. *Biotechnol. J.* **10**, 979–994 (2015).

4. Amoasii, L. *et al.* Gene editing restores dystrophin expression in a canine model of Duchenne muscular dystrophy. *Science* **362**, 86–91 (2018).

5. Li, M. *et al.* Reassessment of the Four Yield-related Genes Gn1a, DEP1, GS3, and IPA1 in Rice Using a CRISPR/Cas9 System. *Front. Plant Sci.* **7**, 377 (2016).

6. Jia, H., Orbovic, V., Jones, J. B. & Wang, N. Modification of the PthA4 effector binding elements in Type I CsLOB1 promoter using Cas9/sgRNA to produce transgenic Duncan grapefruit alleviating XccΔpthA4:dCsLOB1.3 infection. *Plant Biotechnol. J.* **14**, 1291–1301 (2016).

7. Ronda, C., Pedersen, L. E., Sommer, M. O. A. & Nielsen, A. T. CRMAGE: CRISPR Optimized MAGE Recombineering. *Sci. Rep.* **6**, 19452 (2016).

8. Feng, Z. *et al.* Efficient genome editing in plants using a CRISPR/Cas system. *Cell Res.* **23**, 1229 (2013).

9. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

10. Jinek, M. *et al.* A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).

11. Bassett, A. R., Tibbit, C., Ponting, C. P. & Liu, J.-L. Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System. *Cell Rep.* **4**, 220–228 (2013).

12. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).

13. Mali, P. *et al.* RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).

14. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62 (2014).

15. Li, Y. *et al.* De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.* **32**, 1045–1052 (2014).

16. Garst, A. D. *et al.* Genome-wide mapping of mutations at single-nucleotide resolution for protein, metabolic and genome engineering. *Nat. Biotechnol.* **35**, 48 (2017).

17. Zetsche, B. *et al.* Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. *Cell* **163**, 759–771 (2015).

18. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57 (2018).

19. Kim, N. *et al.* Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).

20. Wu, X., Kriz, A. J. & Sharp, P. A. Target specificity of the CRISPR-Cas9 system. *Quant. Biol.* **2**, 59–70 (2014).

21. Zhang, X.-H., Tee, L. Y., Wang, X.-G., Huang, Q.-S. & Yang, S.-H. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. *Mol. Ther. - Nucleic Acids* **4**, e264 (2015).
22. Willkomm, S. *et al.* Structural and mechanistic insights into an archaeal DNA-guided Argonaute protein. *Nat. Microbiol.* **2**, 17035 (2017).
23. Swarts, D. C. *et al.* DNA-guided DNA interference by a prokaryotic Argonaute. *Nature* **507**, 258–261 (2014).
24. Hegge, J. W., Swarts, D. C. & van der Oost, J. Prokaryotic Argonaute proteins: novel genome-editing tools? *Nat. Rev. Microbiol.* **16**, 5 (2018).
25. Cyranoski, D. Authors retract controversial NgAgo gene-editing study. *Nat. News* (2017) doi:10.1038/nature.2017.22412.
26. Sunghyeok, Y. *et al.* DNA-dependent RNA cleavage by the Natronobacterium gregoryi Argonaute. *BioRxiv* 101923 (2017).
27. Fu, L. *et al.* The prokaryotic Argonaute proteins enhance homology sequence-directed recombination in bacteria. *Nucleic Acids Res.* **47**, 3568–3579 (2019).
28. Swarts, D. C. *et al.* The evolutionary journey of Argonaute proteins. *Nat. Struct. Mol. Biol.* **21**, 743–753 (2014).
29. Swarts, D. C. *et al.* Argonaute of the archaeon Pyrococcus furiosus is a DNA-guided nuclease that targets cognate DNA. *Nucleic Acids Res.* **43**, 5120–5129 (2015).
30. Koonin, E. V. Evolution of RNA-and DNA-guided antivirus defense systems in prokaryotes and eukaryotes: common ancestry vs convergence. *Biol. Direct* **12**, 5 (2017).
31. Swarts, D. C. *et al.* Autonomous Generation and Loading of DNA Guides by Bacterial Argonaute. *Mol. Cell* **65**, 985–998 (2017).

32. Hauptmann, J. *et al.* Turning catalytically inactive human Argonaute proteins into active slicer enzymes. *Nat. Struct. Mol. Biol.* **20**, 814 (2013).

33. Faehnle, C. R., Elkayam, E., Haase, A. D., Hannon, G. J. & Joshua-Tor, L. The making of a slicer: activation of human Argonaute-1. *Cell Rep.* **3**, 1901–1909 (2013).

34. Kwak, P. B. & Tomari, Y. The N domain of Argonaute drives duplex unwinding during RISC assembly. *Nat. Struct. Mol. Biol.* **19**, 145 (2012).

35. Ma, J.-B. *et al.* Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature* **434**, 666 (2005).

36. Künne, T., Swarts, D. C. & Brouns, S. J. J. Planting the seed: target recognition of short guide RNAs. *Trends Microbiol.* **22**, 74–83 (2014).

37. Lingel, A., Simon, B., Izaurralde, E. & Sattler, M. Nucleic acid 3′-end recognition by the Argonaute2 PAZ domain. *Nat. Struct. Mol. Biol.* **11**, 576 (2004).

38. Ma, J.-B., Ye, K. & Patel, D. J. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* **429**, 318 (2004).

39. Sheng, G. *et al.* Structure-based cleavage mechanism of Thermus thermophilus Argonaute DNA guide strand-mediated DNA target cleavage. *Proc. Natl. Acad. Sci.* **111**, 652–657 (2014).

40. Wang, Y. *et al.* Structure of an argonaute silencing complex with a seed-containing guide DNA and target RNA duplex. *Nature* **456**, 921 (2008).

41. Hunt, E. A., Evans Jr, T. C. & Tanner, N. A. Single-stranded binding proteins and helicase enhance the activity of prokaryotic argonautes in vitro. *PloS One* **13**, e0203073 (2018).

42. Kaya, E. *et al.* A bacterial Argonaute with noncanonical guide RNA specificity. *Proc. Natl. Acad. Sci.* **113**, 4057–4062 (2016).

43. Wu, Z. *et al.* NgAgo-gDNA system efficiently suppresses hepatitis B virus replication through accelerating decay of pregenomic RNA. *Antiviral Res.* **145**, 20–23 (2017).

44. Qin, Y. Y., Wang, Y. M. & Liu, D. NgAgo-based fabp11a gene knockdown causes eye developmental defects in zebrafish. *Cell Res.* **26**, 1349–1352 (2016).

45. Khin, N. C., Lowe, J. L., Jensen, L. M. & Burgio, G. No evidence for genome editing in mouse zygotes and HEK293T human cell line using the DNA-guided Natronobacterium gregoryi Argonaute (NgAgo). *PloS One* **12**, e0178768 (2017).

46. Javidi-Parsijani, P. *et al.* No evidence of genome editing activity from Natronobacterium gregoryi Argonaute (NgAgo) in human cells. *Plos One* **12**, 14 (2017).

47. Burgess, S. *et al.* Questions about NgAgo. *Protein Cell* **7**, 913–915 (2016).

48. Tadeo, X. *et al.* Structural basis for the amino acid composition of proteins from halophilic archea. *PLoS Biol.* **7**, e1000257 (2009).

49. Elcock, A. H. & McCammon, J. A. Electrostatic contributions to the stability of halophilic proteins. *J. Mol. Biol.* **280**, 731–748 (1998).

50. Yamaguchi, H. & Miyazaki, M. Refolding Techniques for Recovering Biologically Active Recombinant Proteins from Inclusion Bodies. *Biomolecules* **4**, 235–251 (2014).

51. Zander, A., Holzmeister, P., Klose, D., Tinnefeld, P. & Grohmann, D. Single-molecule FRET supports the two-state model of Argonaute action. *RNA Biol.* **11**, 45–56 (2014).

52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

53. Pearson, W. R. An Introduction to Sequence Similarity (“Homology”) Searching. *Curr. Protoc. Bioinforma. Ed. Board Andreas Baxevanis Al* **0 3**, (2013).

54. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

55. Lobley, A., Sadowski, M. I. & Jones, D. T. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinforma. Oxf. Engl.* **25**, 1761–1767 (2009).

56. Yoon, B.-J. Hidden Markov Models and their Applications in Biological Sequence Analysis. *Curr. Genomics* **10**, 402–415 (2009).

57. Carpentier, M. & Chomilier, J. Protein multiple alignments: sequence-based versus structure-based programs. *Bioinformatics* **35**, 3970–3980 (2019).

58. Liu, Y. *et al.* A programmable omnipotent Argonaute nuclease from mesophilic bacteria Kurthia massiliensis. *Nucleic Acids Res.* **49**, 1597–1608 (2021).

59. Hegge, J. W. *et al.* DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. *Nucleic Acids Res.* **47**, 5809–5821 (2019).

60. Lee, K. Z. *et al.* NgAgo DNA endonuclease activity enhances homologous recombination in E. coli. *bioRxiv* 597237 (2020) doi:10.1101/597237.

61. Jolly, S. M. *et al.* Thermus thermophilus Argonaute Functions in the Completion of DNA Replication. *Cell* **182**, 1545-1559.e18 (2020).

62. Cui, L. & Bikard, D. Consequences of Cas9 cleavage in the chromosome of Escherichia coli. *Nucleic Acids Res.* gkw223 (2016).

63. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233 (2013).

64. Jiang, Y. *et al.* Multigene editing in the Escherichia coli genome via the CRISPR-Cas9 system. *Appl. Environ. Microbiol.* **81**, 2506–2514 (2015).

65. Callaway, E. ‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures. *Nature* **588**, 203–204 (2020).

66. Yang, J., Shen, C. & Huang, N. Predicting or Pretending: Artificial Intelligence for Protein-Ligand Interactions Lack of Sufficiently Large and Unbiased Datasets. *Front. Pharmacol.* **11**, (2020).

67. Arakawa, T., Yamaguchi, R., Tokunaga, H. & Tokunaga, M. Unique features of halophilic proteins. *Curr. Protein Pept. Sci.* **18**, 65–71 (2017).

68. Carlson, E. D., Gan, R., Hodgman, C. E. & Jewett, M. C. Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.* **30**, 1185–1194 (2012).

69. Marshall, R. *et al.* Rapid and Scalable Characterization of CRISPR Technologies Using an E. coli Cell-Free Transcription-Translation System. *Mol. Cell* **69**, 146-157.e3 (2018).

70. Nafisi, P. M., Aksel, T. & Douglas, S. M. Construction of a novel phagemid to produce custom DNA origami scaffolds. *Synth. Biol.* **3**, (2018).

71. Shepherd, T. R., Du, R. R., Huang, H., Wamhoff, E.-C. & Bathe, M. Bioproduction of pure, kilobase-scale single-stranded DNA. *Sci. Rep.* **9**, 6121 (2019).

72. Ryazansky, S., Kulbachinskiy, A. & Aravin, A. The expanded universe of prokaryotic Argonaute proteins. *bioRxiv* 366930 (2018).

73. Enghiad, B. & Zhao, H. Programmable DNA-guided artificial restriction enzymes. *ACS Synth. Biol.* **6**, 752–757 (2017).

74. Hur, J. K., Zinchenko, M. K., Djuranovic, S. & Green, R. Regulation of Argonaute slicer activity by guide RNA 3’end interactions with the N-terminal lobe. *J. Biol. Chem.* jbc-M112 (2013).

75. Zander, A. *et al.* Guide-independent DNA cleavage by archaeal Argonaute from Methanocaldococcus jannaschii. *Nat. Microbiol.* **2**, 17034 (2017).

76. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular cloning: a laboratory manual*. (Cold Spring Harbor Laboratory Press, 1989).

77. Tas, H., Nguyen, C. T., Patel, R., Kim, N. H. & Kuhlman, T. E. An integrated system for precise genome modification in Escherichia coli. *PloS One* **10**, e0136963 (2015).

78. Wood, W. B. Host specificity of DNA produced by Escherichia coli: bacterial mutations affecting the restriction and modification of DNA. *J. Mol. Biol.* **16**, 118-IN3 (1966).

79. Tseng, H.-C., Martin, C. H., Nielsen, D. R. & Prather, K. L. J. Metabolic engineering of Escherichia coli for enhanced production of (R)-and (S)-3-hydroxybutyrate. *Appl. Environ. Microbiol.* **75**, 3137–3145 (2009).

80. Niu, Y., Tenney, K., Li, H. & Gimble, F. S. Engineering variants of the I-SceI homing endonuclease with strand-specific and site-specific DNA-nicking activity. *J. Mol. Biol.* **382**, 188–202 (2008).

81. Reisch, C. R. & Prather, K. L. J. The no-SCAR (Scarless Cas9 Assisted Recombineering) system for genome editing in Escherichia coli. *Sci. Rep.* **5**, 15096 (2015).

82. Rhodius, V. A. *et al.* Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol. Syst. Biol.* **9**, 702 (2013).

83. Marshall, R., Maxwell, C. S., Collins, S. P., Beisel, C. L. & Noireaux, V. Short DNA containing χ sites enhances DNA stability and gene expression in E. coli cell-free transcription-translation systems. *Biotechnol. Bioeng.* **114**, 2137–2141 (2017).

84. Zimmermann, L. *et al.* A completely Reimplemented MPI bioinformatics toolkit with a new HHpred server at its Core. *J. Mol. Biol.* **430**, 2237–2243 (2018).

85. Söding, J., Biegert, A. & Lupas, A. N. The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Res.* **33**, W244–W248 (2005).

86. Flynn, R. L. & Zou, L. Oligonucleotide/oligosaccharide-binding fold proteins: a growing family of genome guardians. *Crit. Rev. Biochem. Mol. Biol.* **45**, 266–275 (2010).

87. Müller-Santos, M. *et al.* First evidence for the salt-dependent folding and activity of an esterase from the halophilic archaea Haloarcula marismortui. *Biochim. Biophys. Acta BBA-Mol. Cell Biol. Lipids* **1791**, 719–729 (2009).

88. Bell, J. C., Liu, B. & Kowalczykowski, S. C. Imaging and energetics of single SSB-ssDNA molecules reveal intramolecular condensation and insight into RecOR function. *eLife* **4**, e08646 (2015).

89. Goodall, E. C. A. *et al.* The Essential Genome of Escherichia coli K-12. *mBio* **9**, (2018).

90. Simmons, L. A. *et al.* Comparison of Responses to Double-Strand Breaks between Escherichia coli and Bacillus subtilis Reveals Different Requirements for SOS Induction. *J. Bacteriol.* **191**, 1152–1161 (2009).

91. Wu, Z. *et al.* NgAgo-gDNA system efficiently suppresses hepatitis B virus replication through accelerating decay of pregenomic RNA. *Antiviral Res.* **145**, 20–23 (2017).

92. Microbial life at high salt concentrations: phylogenetic and metabolic diversity | Aquatic Biosystems | Full Text. https://aquaticbiosystems.biomedcentral.com/articles/10.1186/1746-1448-4-2.

93. Paul, S., Bag, S. K., Das, S., Harvill, E. T. & Dutta, C. Molecular signature of hypersaline adaptation: insights from genome and proteome composition of halophilic prokaryotes. *Genome Biol.* **9**, R70 (2008).

94. Siglioccolo, A., Paiardini, A., Piscitelli, M. & Pascarella, S. Structural adaptation of extreme halophilic proteins through decrease of conserved hydrophobic contact surface. *BMC Struct. Biol.* **11**, 50 (2011).

95. Warden, A. C. *et al.* Rational engineering of a mesohalophilic carbonic anhydrase to an extreme halotolerant biocatalyst. *Nat. Commun.* **6**, 10278 (2015).

96. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. E. SuSPect: Enhanced Prediction of Single Amino Acid Variant (SAV) Phenotype Using Network Features. *J. Mol. Biol.* **426**, 2692–2701 (2014).

97. Lee, Y.-H. *et al.* Bacterial Production of Barley Stripe Mosaic Virus Biotemplates for Palladium Nanoparticle Growth. *ACS Appl. Nano Mater.* **3**, 12080–12086 (2020).

98. Wu, S.-L. *et al.* Double homeobox gene, Duxbl, promotes myoblast proliferation and abolishes myoblast differentiation by blocking MyoD transactivation. *Cell Tissue Res.* **358**, 551–566 (2014).

99. Lin, C. Y. *et al.* Extracellular Pgk1 enhances neurite outgrowth of motoneurons through Nogo66/NgR-independent targeting of NogoA. *eLife* **8**, e49175 (2019).

100. Yang, J. *et al.* Improved protein structure prediction using predicted interresidue orientations. *Proc. Natl. Acad. Sci.* **117**, 1496–1503 (2020).

101. Esvelt, K. M., Carlson, J. C. & Liu, D. R. A system for the continuous directed evolution of biomolecules. *Nature* **472**, 499–503 (2011).

102. Ravikumar, A., Arzumanyan, G. A., Obadi, M. K. A., Javanpour, A. A. & Liu, C. C. Scalable, continuous evolution of genes at mutation rates above genomic error thresholds. *Cell* **175**, 1946–1957 (2018).

103. English, J. G. *et al.* VEGAS as a Platform for Facile Directed Evolution in Mammalian Cells. *Cell* **178**, 748-761.e17 (2019).

104. Mechikoff, M. A. DEVELOPMENT OF AN ASSAY TO IDENTIFY AND QUANTIFY ENDONUCLEASE ACTIVITY. (Purdue University Graduate School, 2019). doi:10.25394/PGS.11328623.v1.

105. Gaj, T., Gersbach, C. A. & Barbas III, Carlos. F. ZFN, TALEN and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol.* **31**, 397–405 (2013).

106. Cong, L. *et al.* Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science* **339**, 819–823 (2013).

107. Mali, P. *et al.* RNA-Guided Human Genome Engineering via Cas9. *Science* **339**, 823–826 (2013).

108. Hwang, W. Y. *et al.* Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).

109. Bassett, A. R., Tibbit, C., Ponting, C. P. & Liu, J.-L. Highly Efficient Targeted Mutagenesis of Drosophila with the CRISPR/Cas9 System. *Cell Rep.* **4**, 220–228 (2013).

110. Jiang, W., Bikard, D., Cox, D., Zhang, F. & Marraffini, L. A. RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233–239 (2013).

111. Feng, Z. *et al.* Efficient genome editing in plants using a CRISPR/Cas system. *Cell Res.* **23**, 1229–1232 (2013).

112. Leenay, R. T. *et al.* Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. *Mol. Cell* **62**, 137–147 (2016).

113. Makarova, K. S. *et al.* Evolutionary classification of CRISPR–Cas systems: a burst of class 2 and derived variants. *Nat. Rev. Microbiol.* **18**, 67–83 (2020).

114. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).

115. Slaymaker, I. M. *et al.* Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).

116. Chen, J. S. *et al.* Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

117. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanding targeting space. *Science* **361**, 1259–1262 (2018).

118. Lee, J. K. *et al.* Directed evolution of CRISPR-Cas9 to increase its specificity. *Nat. Commun.* **9**, 3048 (2018).

119. Doyon, J. B., Pattanayak, V., Meyer, C. B. & Liu, D. R. Directed evolution and substrate specificity profile of homing endonuclease I-SceI. *J. Am. Chem. Soc.* **128**, 2477–2484 (2006).

120. Chen, Z. & Zhao, H. A highly sensitive selection method for directed evolution of homing endonucleases. *Nucleic Acids Res.* **33**, e154–e154 (2005).

121. Doyon, J. B., Pattanayak, V., Meyer, C. B. & Liu, D. R. Directed evolution and substrate specificity profile of homing endonuclease I-Scel. *J. Am. Chem. Soc.* **128**, 2477–2484 (2006).

122. Gruen, M., Chang, K., Serbanescu, I. & Liu, D. R. An in vivo selection system for homing endonuclease activity. *Nucleic Acids Res.* **30**, e29–e29 (2002).

123. Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual*. (Cold Spring Harbor Laboratory Press, 2012).

124. Kuhlman, T. E. & Cox, E. C. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* **38**, e92 (2010).

125. Lauritsen, I., Porse, A., Sommer, M. O. A. & Nørholm, M. H. H. A versatile one-step CRISPR-Cas9 based approach to plasmid-curing. *Microb. Cell Factories* **16**, 1–10 (2017).

126. Kuhlman, T. E. & Cox, E. C. Site-specific chromosomal integration of large synthetic constructs. *Nucleic Acids Res.* **38**, (2010).

127. Chayot, R., Montagne, B., Mazel, D. & Ricchetti, M. An end-joining repair mechanism in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 2141–2146 (2010).

128. Greenfield, L., Boone, T. & Wilcox, G. DNA sequence of the araBAD promoter in Escherichia coli B/r. *Proc. Natl. Acad. Sci. U. S. A.* **75**, 4724–4728 (1978).

129. Hegge, J. W. *et al.* DNA-guided DNA cleavage at moderate temperatures by Clostridium butyricum Argonaute. *Nucleic Acids Res.* **47**, 5809–5821 (2019).

130. Egan, S. M. & Schleif, R. F. A regulatory cascade in the induction of rhaBAD. *Journal of Molecular Biology* vol. 234 87–98 (1993).

131. Kuzminov, A. & Stahl, F. W. Stability of Linear DNA in recA Mutant Escherichia coli Cells Reflects Ongoing Chromosomal DNA Degradation. *J. Bacteriol.* **179**, 880–888 (1997).

132. Tolia, N. H. & Joshua-Tor, L. Strategies for protein coexpression in Escherichia coli. *Nat. Methods* **3**, 55–64 (2006).

133. Morlon, J., Sherratt, D. & Lazdunski, C. Identification of functional regions of the colicinogenic plasmid ColA. *MGG Mol. Gen. Genet.* **211**, 223–230 (1988).

134. Jinek, M. *et al.* A Programmable Dual-RNA – Guided DNA Endonuclease in Adaptive Bacterial Immunity. *Science* **337**, 816–822 (2012).

135. Liu, D. R. & Schultz, P. G. Progress toward the evolution of an organism with an expanded genetic code. *Proc. Natl. Acad. Sci.* **96**, 4780–4785 (1999).

136. Guo, M. *et al.* Structural insights into a high fidelity variant of SpCas9. *Cell Res.* **29**, 183–192 (2019).

137. Lee, K. Z. *et al.* NgAgo-enhanced homologous recombination in E. Coli is mediated by DNA endonuclease activity. *bioRxiv* (2019) doi:10.1101/597237.

138. Kuzmenko, A., Yudin, D., Ryazansky, S., Kulbachinskiy, A. & Aravin, A. A. Programmable DNA cleavage by Ago nucleases from mesophilic bacteria Clostridium butyricum and Limnothrix rosea. *Nucleic Acids Res.* **47**, 5822–5836 (2019).

139. Cao, Y. *et al.* Argonaute proteins from human gastrointestinal bacteria catalyze DNA-guided cleavage of single- and double-stranded DNA at 37 °C. *Cell Discov.* **5**, 38 (2019).

PUBLICATIONS

This paper is published (Lee et al, Bacterial Production of Barley Stripe Mosaic Virus Biotemplates for Palladium Nanoparticle Growth, ACS Applied Nano Materials, https://doi.org/10.1021/acsanm.0c02570, 2020). The numbering of the figures, tables, and references have been modified to fit the format of this thesis.

Bacterial Production of Barley Stripe Mosaic Virus Biotemplates for Palladium Nanoparticle Growth

Yu-Hsuan Lee,[#] Kok Zhi Lee,[#] Rachel G. Susler, Corren A. Scott, Longfei Wang, L. Sue Loesch-Fries, Michael T. Harris, and Kevin V. Solomon[*]

Abstract

Barley stripe mosaic virus (BSMV) has recently been proposed as an attractive biotemplate for direct metallic nanomaterial synthesis as it interacts with metal precursors through multiple mechanisms. These interactions more than double the coating capacity for metal and accelerate nanomaterial synthesis, reducing costs, while potentially offering economical synthesis pathways for a wider range of nanomaterials. However, these studies were only able to generate BSMV via plant production, which is not well suited to widescale industrial production and limits engineering of BSMV-templated material properties to protein mutations that maintain or enhance infectivity in plants. Here, BSMV virus-like particles (VLPs) are produced from bacteria for the first time by fusing the origin of assembly from tobacco mosaic virus (TMV) to the transcript encoding BSMV capsid protein. Purification of BSMV-VLPs produced from Escherichia coli results in nanorods that average 82 nm in length and 21 nm in diameter. We also demonstrate that these rod- shaped BSMV-VLPs can be more rapidly coated with Pd metal than in planta-produced BSMV in the absence of an external reducing agent. This study creates an alternate platform for BSMV-VLP production and enables future engineering opportunities to tune nanomaterial properties through biotemplate design.

KEYWORDS: biotemplate, barley stripe mosaic virus (BSMV), VLP, synthetic biology, E. coli, protein expression

Introduction

Bottom-up nanofabrication via templating on naturally occurring biomolecules (biotemplating) is a promising strategy for nanomaterial synthesis, as the products are frequently more uniform and less polydisperse.[1−8] For biotemplates such as plant virus capsid proteins, biotemplate structure is encoded within the protein primary sequence, which enables sponta- neous self-assembly of hierarchical complex nanomaterials for diverse applications.[9−11] These scaffolds present diverse biochemical functionalities on their surface for interaction and coating with various metals. Plant tobacco mosaic virus (TMV) capsid proteins have been utilized as biotemplates and coated with metals such as Ag,[6] Pd,[12−15] Pt,[16] and Au,[12] and Au/Pd alloys[17] for incorporation in devices as battery electrodes,[18],[19] memory devices,[20] catalysts,[3] and chemical sensors.[21]

Given the versatility of TMV, alternative biotemplates based on related viruses from the Virgaviridae family have been proposed. These alternatives present novel surface functionalities that provide new opportunities for the fabrication of nanoscaled materials. In particular, we have established that in planta produced barley stripe mosaic virus (BSMV) from the Virgaviridae family is an exciting alternate template for the synthesis of palladium nanorods.[22] The synthesized nanorods have similar properties to TMV such as morphology and a high adsorption capacity for palladium species. However, unlike TMV, BSMV-mediated reduction and deposition of metal precursor ions proceeds via a multistep Langmuir isotherm that incorporates both electrostatic and covalently driven affinity of metal precursor molecules for amino acid residues on the BSMV surface.[22] These electrostatic interactions have the potential to diversify the metals that may be deposited on template, which can be coated more quickly than TMV, leading to more economical processing for a given metal density.

Current approaches to produce BSMV are limited to in planta production, which makes its development challenging. First, modification to the genomes of in planta-synthesized viruses that may enhance biotemplating functionality may not be stable, as these modifications may not confer a selective advantage for viral propogation.[23] Second, as these viruses are plant pathogens, virus-producing plants must be grown in specialized facilities to prevent infection of plants in the wild.[24] Finally, the viral replication cycle in plants requires 2−3 weeks, which results in slow, long, and complicated processes to extract relatively small amounts of viral capsids. On the other hand, alternative production processes that use bacterial expression platforms are faster and simpler.[25−27] Bacteria grow rapidly within 24 h in fermenters, which have been developed and optimized for decades for use in the food, beverage, biopharma, and biotech sectors. Bacterial hosts can also produce virus-like particles (VLPs) of assembled viral capsid proteins at higher yields and at shorter times when compared to other hosts, including yeast, baculovirus-insect cells, mammalian cells, and plants. Moreover, the use of a heterologous host reduces the evolutionary pressures on virus replication, enabling more opportunities for favorable genetical engineering of VLP structure. For example, TMV capsid protein (CP) has been modified to present on its surface a thiol group from cysteine, an amine group from lysine, or carboxylate groups from glutamate and aspartate, enabling chemical modification of the template at precise locations via covalent coupling,[28−30] conjugation,[31−34] and click chemistry.[35] Thiol introduction on TMV surfaces also enables deposition of diverse metals, including Au,[12,18] Ni,[18] Co,[18] and Pd.[4,12−15,36,37] These engineered modifications can be introduced easily and frequently do not compromise VLP structures while enhancing functionality.[38,39]

As members of Virgaviridae, both TMV and BSMV CPs self- assemble into long nanorod-shaped structures through protein−protein and protein−nucleic acid interactions (Figures 1 and 2). CPs initially self-assemble into disk-like structures via several protein−protein interactions. Nanorod assembly is then initiated via interactions of these CP disks with an RNA stem-loop structure called an origin of assembly sequence (OAS).[40,41] Upon binding of the OAS with the disk interior, the initial nucleating CP disk shifts conformation to a right-hand helix. This conformational shift also restructures the encapsulated RNA transcript to allow for rapid CP polymerization in a helical pattern around the RNA transcript to form a nanorod. Interactions of adjacent CPs in this nanorod are stabilized via a cluster of charged residues called the Casper carboxylate center.[42]

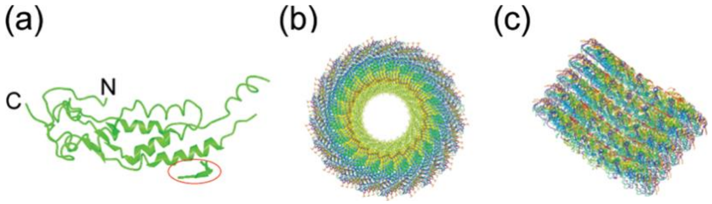


Figure 1. BSMV capsid protein and its assembly as a nanorod (a) BSMV capsid protein. N- and C-termini are labeled as N and C, respectively. The amino acids that interact with BSMV RNA are circled. Current crystal structures are only partial but capture ~90% of all residues. (b) Top-down view of assembled BSMV capsid proteins. (c) Side view of an assembled BSMV virion. The highest resolution crystal structure (PDB:5A7A[42]) was obtained from NCBI MMDB database[43] and rendered using Cn3D.[44]

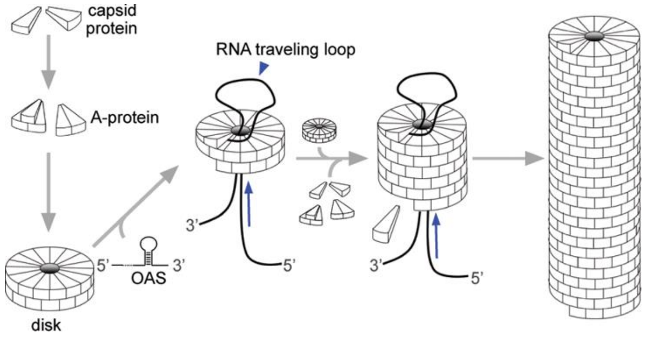


Figure 2. Assembly process of TMV involves multiple molecular interactions. Interactions between capsid proteins form “A-protein” and subsequently a disk structure before the origin of assembly sequence (OAS) on an RNA transcript initiates virus assembly. Subsequent disks and capsid proteins are assembled into a virus rod via protein−protein interaction and protein−RNA interaction.

BSMV-VLPs have not been produced in a bacterial host to date, as a native OAS to initiate its assembly remains unidentified. However, chimeric BSMV viral particles have been shown to form due to interactions between BSMV CPs and the TMV genome[41],[42] in plants. Thus, we evaluated whether the TMV OAS may be sufficient to initiate BSMV- VLP assembly in bacteria. The introduced OAS successfully enabled assembly of BSMV CPs into rod-shaped VLPs in E. coli. After optimization of the expression, purification, and processing conditions, we produced BSMV-VLPs that could be used as biotemplates for synthesis of palladium-coated nanorods via hydrothermal processing. This bacterial synthesis platform for BSMV-VLP readily accelerates the production of BSMV-templated nanomaterials and opens new opportunities to tune nanomaterial properties via template engineering.

Results and discussion

The self-assembly of BSMV-VLP relies on interactions between virus RNA and CPs, and interactions between adjacent capsid protein subunits (Figure 2). As the TMV OAS is recognized by BSMV-CP in plants,[41,42] we hypothesized that BSMV-CP transcripts with a TMV OAS at the 3′ end can initiate VLP assembly via RNA−CP interactions and lead to the formation of BSMV-VLPs (Figure 3). BSMV CP expression plasmids, which were codon-optimized for bacterial expression, were designed with and without a TMV OAS. BSMV-CP was initially expressed at 37 °C for 4 h in Escherichia coli before lysing. Crude protein lysate was then centrifuged through several rounds to purify any synthesized VLPs, which were then characterized using transmission electron microscopy (TEM). TEM images did not show any BSMV rod-shaped VLPs or disk structures (data not shown) suggesting that BSMV CPs were not produced, did not self- assemble or that the isolation procedure was insufficient to capture produced VLPs. TMV constructs were expressed and purified as a positive control. Subsequent electron microscopy displayed the presence of TMV-VLPs, excluding the possibility of inefficient VLP isolation. Moreover, disk-shaped structures will form if wild-type CPs are successfully expressed in the host but fail to self-assemble into rod VLPs.[45] Thus, the absence of BSMV disk structures suggested poor soluble CP expression. To examine expression, crude protein lysates from E. coli with induced CP plasmids were analyzed via SDS-PAGE (Figure 4). SDS-PAGE analysis revealed a relatively heavy band of BSMV protein in the bacterial insoluble pellet suggesting that the majority of the CPs was misfolded (Figure 4a).

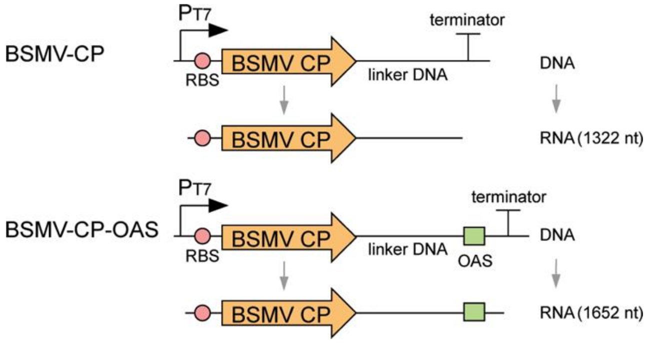


Figure 3. BSMV capsid protein expression constructs. Transcripts made from BSMV-CP (top) contain the BSMV-CP ORF and a linker RNA for a total length of 1322 nt. Transcripts made from BSMV-CP- OAS are similar, but they contain an OAS following the linker for a total length of 1652 nt (bottom). Both gene cassettes are driven by T7 promoter (P_{T7}), induced by IPTG, and translation is initiated via the native ribosomal binding site (RBS) of the parent expression vector.

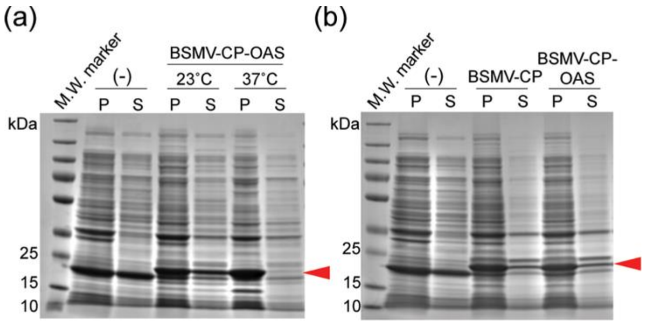


Figure 4. Expression of BSMV-CP-OAS at reduced temperatures leads to VLPs. (a) SDS-PAGE analysis of BSMV capsid protein (BSMV-CP-OAS) expression at room temperature (23 °C) and 37 °C, respectively. P represents the pellet and S represents the supernatant after centrifugation at 92,000 × g for 20 min. (b) SDS- PAGE analysis of cultures expressing BSMV-CP, BSMV-CP-OAS or a negative control lacking expression plasmid at room temperature. Arrows indicate the BSMV capsid protein at ∼22.5 kDa.

Heterologous proteins, such as BSMV CP, may misfold at high temperatures due to too rapid protein synthesis and increased hydrophobic interactions that lead to the formation of insoluble protein aggregates in inclusion bodies.[46,47] To address this, we lowered the expression temperature to room temperature and extended the expression time from 4 to 16 h in order to slow down the protein expression rate and facilitate proper protein folding.[47] SDS-PAGE analysis revealed that reducing the expression temperature significantly increased soluble CP expression (Figure 4a). TEM analysis showed that the BSMV-CP-OAS construct did indeed form rod-shaped

BSMV-VLPs while only disk structures formed from the OAS- free construct (Figure 5). SDS-PAGE confirmed soluble capsid protein expression under these conditions for both BSMV-CP and BSMV-CP-OAS constructs (Figure 4). Thus, the TMV OAS is sufficient for BSMV-VLP assembly in bacteria, which is consistent with previous studies in planta[42] (Figure 5 and Figure S1A). In the absence of an OAS, only disks formed (Figure 5a).

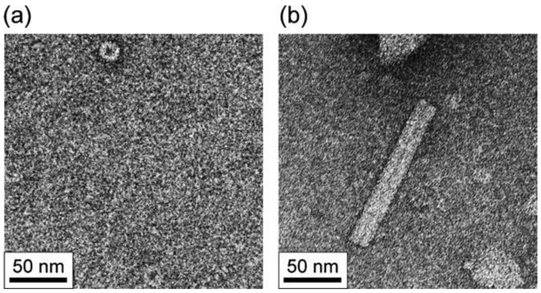


Figure 5. TEM analysis of VLPs assembled in vivo from BSMV-CP translated from an RNA (a) without an OAS and (b) with an OAS. Scale bar: 50 nm.

To increase the yield of soluble BSMV CP, we further optimized protein expression. As the inducer concentration affects the rate of protein expression and subsequent folding, we hypothesized that reduced IPTG concentrations would induce expression of more soluble protein. We tested IPTG concentrations of 0.10, 0.075, 0.050, and 0.010 mM and assessed the expression of BSMV CPs in the soluble fraction by SDS-PAGE. As shown in Figure 6, 0.075 and 0.10 mM IPTG induced higher yields of soluble CP, which is consistent with microbial expression of other VLPs such as TMV1Cys expression in E. coli.[45]

The buffer used to solubilize the final VLP product had a significant impact on the yield of VLPs. Sodium phosphate buffer (Na_2HPO_4, 10 mM, pH 7) is often used to solubilize VLPs, but its use here led to the formation of a white precipitate of aggregated VLPs over time (data not shown). That is, BSMV-VLPs were not stable in solution with Na_2HPO_4 and may have coagulated due to counterion adsorption neutralizing the electrostatic interactions of the BSMV surface.[48,49] Therefore, we investigated water and 10 mM tris(hydroxymethyl)aminomethane (Tris−HCl), pH 7, as resuspension buffers, which form ions with weaker net charge and/or different geometries that may interact more weakly. Although, BSMV-VLPs appeared to be equally soluble in water, or Tris or phosphate buffers (Figure 6b), there was no precipitation in Tris or water over time. However, due to the pH buffering, Tris buffer was used in all subsequent preparations.

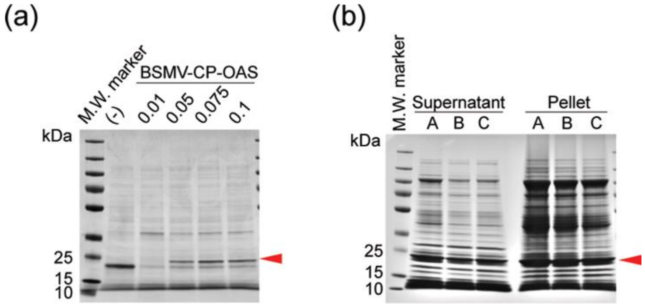


Figure 6. Analysis of BSMV-CP expression and solubilization. (a) SDS-PAGE analysis of BSMV-CP-OAS in 5 µg of bacterial preparations induced by different concentrations of IPTG (0.01, 0.05, 0.075, and 0.10 mM) compared with a negative control strain lacking the BSMV-CP-OAS expression plasmid. (b) SDS-PAGE analysis of the solubility of BSMV-CP-OAS in selected buffers: A. water, B Tris buffer, C sodium phosphate buffer. Arrows indicate BSMV capsid protein at ~22.5 kDa.

To isolate VLPs for characterization, the preparations were centrifuged at 64,000 × g for 1 h over a saturated sucrose cushion.[21] However, BSMV-VLPs of various lengths were distributed throughout the sucrose cushion as shown in Figure 7 a and b. The BSMV-VLPs were often found in end-to-end (linear) and side-by-side (raft) aggregates (Figure 7). End-to-end aggregates have been observed in some preparations of BSMV replicated in planta, which were linear aggregates of up to 40 viral particles. Similarly, raft aggregates are common drying artifacts of TEM preparations;[51] in both cases, aggregate can arise due to the interfacial forces that arise during sample drying and staining for TEM visualization,[52] and thus sample preparation needs optimization. Because the VLPs did not form homogeneous bands in sucrose (Figure 7a,b), cushions were not used for purification (Figure 7c,d).

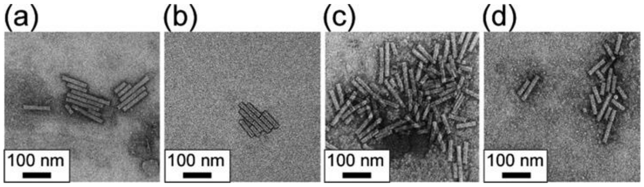


Figure 7. Analysis of BSMV-CP-OAS VLPs following centrifugation at 64,000 × g for 1 h over a saturated sucrose cushion. (a) VLPs isolated above the cushion; (b) VLPs at the bottom of the cushion. (c, d) VLPs without centrifugation through sucrose cushion. Scale bar: 100 nm.

The purified BSMV-VLPs analyzed by TEM were nanorods that ranged in size from 20 to 160 nm in length with an average of 82 nm (Figure 8). These nanorods had a measured diameter of 21 nm. BSMV viral particle length is typically a function of the size of the encapsidated RNA, which initiates self-assembly.[53] In related TMV, the length of the encapsidated OAS-containing transcript directly scales with VLP size, serving as a molecular ruler.[54] Assuming a similar scaling to wild-type BSMV, one would expect our VLP constructs to attain a size of ∼1 nm/27 nt or ∼ 61 nm. The observed lengths were more than 30% higher than expected and could be a result of the nonnative TMV OAS creating novel secondary interactions that expanded the structure of the encapsidated RNA and the resulting viral particle or uncontrolled assembly beyond the molecular ruler. To determine the heterogeneity or purity of the VLP preparations, the size distribution was measured by dynamic light scattering (DLS) (Table 1). DLS measures the hydrodynamic radius, or size of an equivalent particle that diffuses with the same rate, and is a rough estimate of particle size that is distinct from its physical dimensions for nonspherical particles. The hydrodynamic radius of rod-like VLPs is at least half the radius of gyration,[55,56] a geometrical property calculated from the measured physical dimenstions.[56] Using our TEM-measured dimensions, we estimate that the VLPs have a radius of gyration of 9.4−47 nm or a hydrodynamic radius of at least 4.7−23 nm. Analysis of the VLP preparation yielded a bimodal distribution of hydrodynamic radii with peaks at 38.0 nm (8%) and 476.5 nm (92%) (Table 1). The majority of the DLS-measured particles (476.5 nm) are two-orders of magnitude larger than expected VLP particles suggesting that the purified VLP exists primarily in aggregates as observed by TEM or complexed with other cellular impurities. The smaller peak is consistent with linear aggregates of BSMV, which are commonly observed in the literature. While future is needed to optimize the purification of individual VLP particles, E. coli produced BSMV-VLPs display similar architecture to wild-type viral particles.

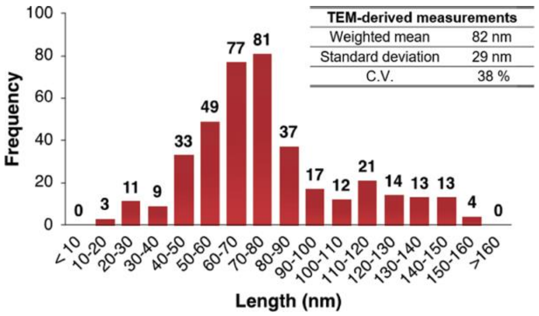


Figure 8. Length distribution histogram of BSMV-VLPs.

Table 1. Length of VLPs as Measured by Dynamic Light Scattering (DLS)

Size (hydrodynamic diameter) (nm)	Relative abundance	Standard deviation (nm)	Coefficient of variation (%)
476.5	92.3	157.7	33
38.0	7.7	6.4	17

The BSMV-VLPs were coated with palladium via a hydrothermal process to evaluate their capability as a biotemplate for nanomaterial synthesis. The BSMV-VLPs were coated in the absence of an external reducing agent by incubation in a stirred reaction vessel with 0.75 mM Na_2PdCl_4 precursor solution. As shown in Figure 9 and Figure S1B, BSMV-VLPs were successfully coated with a layer of palladium nanoparticles in a single round of processing. The Pd layer is dense and fully coated, increasing the nanomaterial diameter to 33 nm (~12 nm thick coating). Multiple coating cycles are typically required to achieve similar properties in in planta-produced TMV and BSMV. Similar single layer coatings with in planta-produced TMV[12,36] are nonuniform and only up to ~7 nm thick. Plant-generated BSMV,[22] on the other hand, was observed to require two rounds of processing to achieve similar coating thicknesses with Pd; however, this study conducted its metal deposition at 55 °C rather than the 57 °C tested here. The results indicate BSMV-VLPs produced from E. coli can serve as effective biotemplates for nanomaterial synthesis, achieving similar or better coating thicknesses in fewer processing steps than current production platforms.

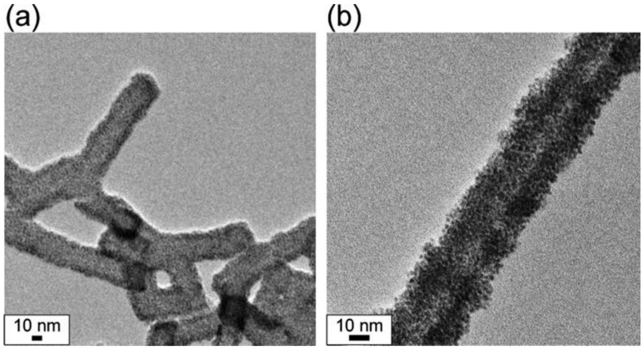


Figure 9. TEM image of a palladium-coated BMSV-VLP nanorod. Scale bar: 10 nm.

Conclusions

In summary, we have developed techniques for the production of novel BSMV-VLPs from a bacterial expression system. The expression of BSMV-VLPs was achieved by fusing an OAS sequence from TMV downstream of mRNA encoding BSMV CP. E. coli production platforms offer unique opportunities for genetic engineering and faster protein expression; therefore, the development of our system enables rapid design-build-test cycles for the engineering and production of BSMV-VLPs with desired properties. VLP expression was optimized by controlling expression temperature and isolation buffers used during purification, resulting in effective cell lysis and higher yields of soluble BSMV CP. The BSMV-VLPs were also shown to be effective biotemplates for the synthesis of inorganic nanomaterials of high quality. This work enables the generation of BSMV-derived biotemplates that expand the toolbox for bottom-up nanomaterial synthesis.

Methods

Cloning of BSMV Capsid Protein Expression Plasmid. E. coli strains and plasmids used in this study are listed in Table 2. All molecular biology manipulations were carried out according to standard practices.[57] A codon-optimized DNA sequence containing BSMV capsid protein cDNA, linker DNA[54] and TMV OAS cDNA (BSMV-CP-linker-OAS) was synthesized by IDT (Coralville, IA) (Supporting Information). This sequence was first ligated to the pGEM-T-Easy intermediate vector (Promega, Madison, WI; Cat. No.: A1360) before cloning into the vector pET21-1cys-tmv-cp (provided by Professor Culver, University of Maryland, College Park) between the NdeI and XhoI restriction sites, generating pET21-BSMV-CP- linker-OAS. pET21-BSMV-CP-linker-OAS was subsequently digested with SalI and XhoI to remove the OAS, blunt-ended with Klenow fragment (NEB, Ipswich, MA; Cat. No.: M0210S), and religated to itself, generating pET21-BSMV-CP-linker. All constructs were sequence verified via Sanger sequencing at Genewiz (South Plainfield, NJ).

BSMV Capsid Protein Expression Conditions. The BSMV CP expression plasmids were transformed into E. coli BL21-CodonPlus (DE3)-RIPL (Agilent Technologies, Santa Clara, CA; Cat. No.: #230280). The bacteria were streaked onto plates containing LB media plus 100 μg/mL ampicillin and 25 μg/ml chloramphenicol and incubated for 16−20 h at 37 °C. Single colonies were selected, inoculated into LB broth, and incubated at 37 °C for 16−20 h at 250 rpm. The liquid cultures were then diluted 100-fold in LB broth and incubated at 37 °C until an OD_{600} of 0.5. The cultures were induced with the addition of 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) for expression of the BSMV CP followed by incubation for 16−20 h at room temperature (∼23 °C) to express CP. All BL21- CodonPlus (DE3)-RIPL liquid cultures or plates contained ampicillin (100 μg/ml) and chloramphenicol (25 μg/ml). Bacteria were collected by centrifugation at room temperature for 5 min at 6000 rpm. The pellet containing the bacteria was used directly for isolation of BSMV-VLPs or stored at −80 °C.

BSMV-VLP Purification. BSMV-VLPs were isolated from E. coli by resuspension in Bugbuster protein isolation solution (Millipor- eSigma, Burlington, MA) according to the manufacturer’s instructions, supplemented with 1.2 mM dithiothreitol. Lysonase Bio- processing Reagent (MilliporeSigma, Burlington, MA) was added according per the manufacturer’s instructions and the suspension was incubated for 10 min at room temperature to lyse the cells followed by centrifugation at 19,000 × g for 10 min to remove insoluble debris. The VLPs in the preparation were isolated by centrifugation at 64,000 × g at 4 °C for 1 h followed by resuspension in 10 mM Tris−HCl, pH 7.

In an attempt to remove large aggregates from the VLP suspension, preparations were layered over a saturated sucrose cushion and spun at 19,000 × g for 10 min at room temperature in an Optima TL Ultracentrifuge (Beckman Coulter, Brea, CA). A top light-scattering band in the cushion was collected and centrifuged at 64,000 × g at 4 °C for 1 h and resuspended in 0.01 M Tris−HCl buffer at pH 7. For samples purified without a sucrose cushion, the lysate supernatant was centrifuged at 19,000 × g for 10 min at room temperature. The supernatant was collected and centrifuged at 64,000 × g at 4 °C for 1 h. The resulting pellet containing the VLPs was resuspended in 0.01 M Tris−HCl buffer at pH 7.

Verification of Coat Protein Expression. To validate coat protein expression, cell lysates were analyzed on 4−20% poly- acrylamide gels (Thermo Fisher Scientific, Waltham, MA; Cat. No.: XP04200BOX). Fourteen microliters of protein lysate was mixed with an equal volume of 2X Tris-glycine SDS Sample Buffer (Thermo Fisher Scientific, Waltham, MA; Cat. No.: LC2676) and supplemented with 2 μL of 1 M DTT (Thermo Fisher Scientific, Waltham, MA; Cat. No.: AC426380100). Samples were then incubated at 85 °C for 5 min to denature the proteins. Samples were then placed on ice for 5 min before being loaded on to the gel. PageRuler Plus Prestained Protein Ladder (Thermo Fisher Scientific, Waltham, MA; Cat. No.: 26620) was used as a molecular weight standard. The gels were run at 120 V for an hour before staining with Coomassie blue (Fisher Scientific, Pittsburgh, PA; Cat. No.: BP101−25) for 10 min. Gels were then destained with destaining buffer (10% glacial acetic acid and 10% methanol) overnight before visualization under visible light with an Azure c400 imager (Azure Biosystems, Dublin, CA).

TEM Imaging. Samples were prepared for imaging in a 200 kV Tecnai T20 TEM by placing 1.5 μl of the VLP suspension onto formvar/carbon coated copper grids followed by an equal amount of ACS-grade phosphotungstic acid (PTA, stock concentration: 1%) for negative staining. After 15 s, the excess liquid was wicked from the grid with 3MM paper and the grid was allowed to dry. At least 50 images were taken per sample using a calibrated Gatan Ultrascan 1000 CCD camera (Gatan Inc., Pleasanton, CA). More than 20 images with good contrast and focus were analyzed with the ImageJ software to measure the dimensions of ∼400 nanorods.

Size Measurement by Dynamic Light Scattering. The refractive index of purified BSMV-VLPs was measured in 0.01 M Tris buffer (pH 7) by an ABBE-3L refractometer (Thermo Fisher Scientific, Waltham, MA). The obtained VLP refractive index (1.3551) with the refractive index of the Tris resuspension buffer (1.3500) and viscosity (1.00037 cP)[58] were used for subsequent

dynamic light-scattering detection with a scattering angle of 17° by a Malvern Zetasizer Nano ZS (Malvern Panalytical Ltd., UK).

Metal Coating Process. Metal coating of the VLPs was performed in a 100 mL CSTR reactor vessel at 57 °C. The VLPs and 0.75 mM ACS-grade sodium tetrachloropalladate (II) (Na_2PdCl_4) (Sigma Aldrich, St Louis, MO) were added to the reaction vessel for 20 min and stirred continuously. The coated biotemplate aggregates spontaneously precipitated and were washed repeatedly with distilled, deionized water (18 MΩ) to remove residual salt and precursor solution.

Associated content

*Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsanm.0c02570.

DNA sequence of synthesized BSMV-CP-linker-OAS construct
SEM images of uncoated and coated BSMV-VLPs (PDF)

Author information

Corresponding Author

Kevin V. Solomon − 225 South University Street, Agricultural & Biological Engineering, 1203 West State Street, Bindley Bioscience Center, and 500 Central Drive, Laboratory of Renewable Resources Engineering (LORRE), Purdue University, West Lafayette, Indiana 47907-2093, United States; orcid.org/0000-0003-2904-9118; Email: kvs@ purdue.edu

Authors

Yu-Hsuan Lee − School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907, United States

Kok Zhi Lee − 225 South University Street, Agricultural & Biological Engineering and 1203 West State Street, Bindley Bioscience Center, Purdue University, West Lafayette, Indiana 47907-2093, United States

Rachel G. Susler − School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907, United States

Corren A. Scott − School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907, United States

Longfei Wang − 915 West State Street, Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana 47907, United States

L. Sue Loesch-Fries − 915 West State Street, Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana 47907, United States

Michael T. Harris − School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907, United States;

orcid.org/0000-0002-0797-8701

Complete contact information is available at:

https://pubs.acs.org/10.1021/acsanm.0c02570

Author Contributions

[#]Y.-H. L. and K.Z.L. contributed equally.

Notes

The authors declare no competing financial interest.

Acknowledgments

We would like to gratefully thank Dr. James Culver at the University of Maryland Biotechnology Institute, Center for Biosystems Research, for his insights on the production of VLPs. This material is based upon work supported by the National Science Foundation under Grant No. CBET-2028618. The research was also supported by the Robert B. and Virginia V. Covalt Professorship of Chemical Engineering, the Purdue Department of Botany & Plant Pathology, the Purdue Research Foundation (Fellowships #60000025 and #60000029), the USDA National Institute of Food and Agriculture (Hatch Multistate Project S1075) and startup funds from the Purdue Colleges of Agriculture and Engineering. Electron microscopy was performed with the help of Dr. Christopher J. Gilpin, Laurie Mueller, and Robert Seiler in the Purdue Life Science Microscopy Facility.

References

(1) Flynn, C. E.; Lee, S.-W.; Peelle, B. R.; Belcher, A. M. Viruses as vehicles for growth organization and assembly of materials. Acta Mater. 2003, 51, 5867−5880.

(2) Knez, M.; Sumser, M.; Bittner, A. M.; Wege, C.; Jeske, H.; Martin, T. P.; Kern, K. Spatially selective nucleation of metal clusters on the tobacco mosaic virus. Adv. Funct. Mater. 2004, 14, 116−124.

(3) Yang, C.; Manocchi, A. K.; Lee, B.; Yi, H. Viral-templated palladium nanocatalysts for Suzuki coupling reaction. J. Mater. Chem. 2011, 21, 187−194.

(4) Manocchi, A. K.; Horelik, N. E.; Lee, B.; Yi, H. Simple, readily controllable palladium nanoparticle formation on surface-assembled viral nanotemplates. Langmuir 2010, 26, 3670−3677.

(5) Dujardin, E.; Peet, C.; Stubbs, G.; Culver, J. N.; Mann, S. Organization of metallic nanoparticles using tobacco mosaic virus templates. Nano Lett. 2003, 3, 413−417.

(6) Lee, S.-Y.; Royston, E.; Culver, J. N.; Harris, M. T. Improved Metal Cluster Deposition on a Genetically Engineered Tobacco Mosaic Virus Template. Nanotechnology 2005, 16, S435−S441.

(7) Zhang, Y.; Dong, Y.; Zhou, J.; Li, X.; Wang, F. Application of plant viruses as a biotemplate for nanomaterial fabrication. Molecules 2018, 23, 2311.

(8) Kim, I.; Kang, K.; Oh, M. H.; Yang, M. Y.; Park, I.; Nam, Y. S. Virus-Templated Self-Mineralization of Ligand-Free Colloidal Palla- dium Nanostructures for High Surface Activity and Stability. Adv. Funct. Mater. 2017, 27, 1703262.

(9) Jeevanandam, J.; Pal, K.; Danquah, M. K. Virus-like nano- particles as a novel delivery tool in gene therapy. Biochimie 2019, 157, 38−47.

(10) Chu, S.; Brown, A. D.; Culver, J. N.; Ghodssi, R. Tobacco Mosaic Virus as a Versatile Platform for Molecular Assembly and Device Fabrication. Biotechnol. J. 2018, 13, 1800147.

(11) Larkin, E. J.; Brown, A. D.; Culver, J. N., Fabrication of tobacco mosaic virus-like nanorods for peptide display. In Virus-Derived Nanoparticles for Advanced Technologies, Springer: 2018; pp. 51−60.

(12) Lim, J.-S.; Kim, S.-M.; Lee, S.-Y.; Stach, E. A.; Culver, J. N.; Harris, M. T. Quantitative Study of Au(III) and Pd(II) Ion Biosorption on Genetically Engineered Tobacco Mosaic Virus. J. Colloid Interface Sci. 2010, 342, 455−461.

(13) Freer, A. S.; Guarnaccio, L.; Wafford, K.; Smith, J.; Steilberg, J.; Culver, J. N.; Harris, M. T. SAXS characterization of genetically engineered tobacco mosaic virus nanorods coated with palladium in the absence of external reducing agents. J. Colloid Interface Sci. 2013, 392, 213−218.

(14) Adigun, O. O.; Freer, A. S.; Miller, J. T.; Loesch-Fries, L. S.; Kim, B. S.; Harris, M. T. Mechanistic Study of the Hydrothermal Reduction of Palladium on the Tobacco Mosaic Virus. J. Colloid Interface Sci. 2015, 450, 1−6.

(15) Adigun, O. O.; Novikova, G.; Retzlaff-Roberts, E. L.; Kim, B.; Miller, J. T.; Loesch-Fries, L. S.; Harris, M. T. Decoupling and elucidation of surface-driven processes during inorganic mineraliza- tion on virus templates. J. Colloid Interface Sci. 2016, 483, 165−176.

(16) Lee, S.-Y.; Choi, J.; Royston, E.; Janes, D. B.; Culver, J. N.; Harris, M. T. Deposition of Platinum Clusters on Surface-Modified Tobacco Mosaic Virus. J. Nanosci. Nanotechnol. 2006, 6, 974−981.

(17) Lim, J.-S.; Kim, S.-M.; Lee, S.-Y.; Stach, E. A.; Culver, J. N.; Harris, M. T. Formation of au/pd alloy nanoparticles on TMV. J. Nanomaterials 2010, 2010, 1−6.

(18) Royston, E.; Ghosh, A.; Kofinas, P.; Harris, M. T.; Culver, J. N. Self-assembly of virus-structured high surface area nanomaterials and their application as battery electrodes. Langmuir 2008, 24, 906−912.

(19) Chen, X.; Gerasopoulos, K.; Guo, J.; Brown, A.; Wang, C.; Ghodssi, R.; Culver, J. N. Virus-Enabled Silicon Anode for Lithium- Ion Batteries. ACS Nano 2010, 4, 5366−5372.

(20) Tseng, R. J.; Tsai, C.; Ma, L.; Ouyang, J.; Ozkan, C. S.; Yang, Y. Digital memory device based on tobacco mosaic virus conjugated with nanoparticles. Nat. Nanotechnol. 2006, 1, 72.

(21) Bruckman, M. A.; Liu, J.; Koley, G.; Li, Y.; Benicewicz, B.; Niu, Z.; Wang, Q. Tobacco mosaic virus based thin film sensor for detection of volatile organic compounds. J. Mater. Chem. 2010, 20, 5715−5719.

(22) Adigun, O. O.; Retzlaff-Roberts, E. L.; Novikova, G.; Wang, L.; Kim, B.-S.; Ilavsky, J.; Miller, J. T.; Loesch-Fries, L. S.; Harris, M. T. BSMV as a biotemplate for palladium nanomaterial synthesis. Langmuir 2017, 33, 1716−1724.

(23) Yusibov, V.; Shivprasad, S.; Turpen, T.; Dawson, W.; Koprowski, H., Plant viral vectors based on tobamoviruses. In Plant Biotechnology, Springer: 2000; pp. 81−94.

(24) Brewer, H. C.; Hird, D. L.; Bailey, A. M.; Seal, S. E.; Foster, G. D. A guide to the contained use of plant virus infectious clones. Plant biotechnol. J. 2018, 16, 832−843.

(25) Jeong, H.; Seong, B. L. Exploiting virus-like particles as innovative vaccines against emerging viral infections. J. Microbiology 2017, 55, 220−230.

(26) Zeltins, A. Construction and characterization of virus-like particles: a review. Mol. Biotechnol. 2013, 53, 92−107.

(27) Schneemann, A.; Young, M. J. Viral assembly using heterologous expression systems and cell extracts. Adv. Protein Chem. 2003, 64, 1−36.

(28) Meunier, S.; Strable, E.; Finn, M. Crosslinking of and coupling to viral capsid proteins by tyrosine oxidation. Chem. Biol. 2004, 11, 319−326.

(29) Pokorski, J. K.; Steinmetz, N. F. The art of engineering viral nanoparticles. Mol. Pharmaceutics 2010, 8, 29−43.

(30) Aljabali, A. A.; Barclay, J. E.; Butt, J. N.; Lomonossoff, G. P.; Evans, D. J. Redox-active ferrocene-modified Cowpea mosaic virus nanoparticles. Dalton Trans. 2010, 39, 7569−7574.

(31) Molino, N. M.; Wang, S.-W. Caged protein nanoparticles for drug delivery. Curr. Opin. Biotechnol. 2014, 28, 75−82.

(32) Wen, A. M.; Shukla, S.; Saxena, P.; Aljabali, A. A.; Yildiz, I.; Dey, S.; Mealy, J. E.; Yang, A. C.; Evans, D. J.; Lomonossoff, G. P. Interior engineering of a viral nanoparticle and its tumor homing properties. Biomacromolecules 2012, 13, 3990−4001.

(33) Zhou, K.; Li, F.; Dai, G.; Meng, C.; Wang, Q. Disulfide bond: dramatically enhanced assembly capability and structural stability of tobacco mosaic virus nanorods. Biomacromolecules 2013, 14, 2593− 2600.

(34) Zhou, K.; Zhang, J.; Wang, Q. Site-Selective Nucleation and Controlled Growth of Gold Nanostructures in Tobacco Mosaic Virus Nanotubulars. Small 2015, 11, 2505−2509.

(35) Bruckman, M. A.; Kaur, G.; Lee, L. A.; Xie, F.; Sepulveda, J.; Breitenkamp, R.; Zhang, X.; Joralemon, M.; Russell, T. P.; Emrick, T. Surface modification of tobacco mosaic virus with "click" chemistry. ChemBioChem 2008, 9, 519−523.

(36) Lim, J.-S.; Kim, S.-M.; Lee, S.-Y.; Stach, E. A.; Culver, J. N.; Harris, M. T. Biotemplated aqueous-phase palladium crystallization in the absence of external reducing agents. Nano Lett. 2010, 10, 3863− 3867.

(37) Lee, S. Y.; Lim, J. S.; Harris, M. T. Synthesis and application of virus-based hybrid nanomaterials. Biotechnol. Bioeng. 2012, 109, 16− 30.

(38) Yi, H.; Nisar, S.; Lee, S.-Y.; Powers, M. A.; Bentley, W. E.; Payne, G. F.; Ghodssi, R.; Rubloff, G. W.; Harris, M. T.; Culver, J. N. Patterned assembly of genetically modified viral nanotemplates via nucleic acid hybridization. Nano Lett. 2005, 5, 1931−1936.

(39) Geiger, F. C.; Eber, F. J.; Eiben, S.; Mueller, A.; Jeske, H.; Spatz, J. P.; Wege, C. TMV nanorods with programmed longitudinal domains of differently addressable coat proteins. Nanoscale 2013, 5, 3808−3816.

(40) Butler, P. J. G. The current picture of the structure and assembly of tobacco mosaic virus. J. Gen. Virol. 1984, 65, 253−279.

(41) Butler, P. Self−assembly of tobacco mosaic virus: the role of an intermediate aggregate in generating both specificity and speed. Philos. Trans. R. Soc. London, Ser. B 1999, 354, 537−550.

(42) Clare, D. K.; Pechnikova, E. V.; Skurat, E. V.; Makarov, V. V.; Sokolova, O. S.; Solovyev, A. G.; Orlova, E. V. Novel inter-subunit contacts in barley stripe mosaic virus revealed by cryo-electron microscopy. Structure 2015, 23, 1815−1826.

(43) Madej, T.; Lanczycki, C. J.; Zhang, D.; Thiessen, P. A.; Geer, R. C.; Marchler-Bauer, A.; Bryant, S. H. MMDB and VAST+: tracking structural similarities between macromolecular complexes. Nucleic Acids Res. 2013, 42, D297−D303.

(44) Wang, Y.; Geer, L. Y.; Chappey, C.; Kans, J. A.; Bryant, S. H. Cn3D: sequence and structure views for Entrez. Trends Biochem. Sci. 2000, 25, 300−302.

(45) Brown, A. D.; Naves, L.; Wang, X.; Ghodssi, R.; Culver, J. N. Carboxylate-directed in vivo assembly of virus-like nanorods and tubes for the display of functional peptides and residues. Biomacromolecules 2013, 14, 3123−3129.

(46) Villaverde, A.; Carrió, M. M. Protein aggregation in recombinant bacteria: biological role of inclusion bodies. Biotechnol. Lett. 2003, 25, 1385−1395.

(47) Sørensen, H. P.; Mortensen, K. K. Soluble expression of recombinant proteins in the cytoplasm of Escherichia coli. Microb. cell fac. 2005, 4, 1.

(48) Donovan, A. R.; Adams, C. D.; Ma, Y.; Stephan, C.; Eichholz, T.; Shi, H. Fate of nanoparticles during alum and ferric coagulation monitored using single particle ICP-MS. Chemosphere 2018, 195, 531−541.

(49) Xu, C.-Y.; Xu, R.-K.; Li, J.-Y.; Deng, K.-Y. Phosphate-induced aggregation kinetics of hematite and goethite nanoparticles. J. Soils and Sed. 2017, 17, 352−363.

(50) Kassanis, B.; Slykhuis, J. Some properties of barley stripe mosaic virus. Ann. of Appl. Biol. 1959, 47, 254−263.

(51) Michen, B.; Geers, C.; Vanhecke, D.; Endes, C.; Rothen- Rutishauser, B.; Balog, S.; Petri-Fink, A. Avoiding drying-artifacts in transmission electron microscopy: Characterizing the size and colloidal state of nanoparticles. Sci. Rep. 2015, 5, 9793.

(52) Israelachvili, J. N., Intermolecular and Surface Forces. Academic Press: 2011; 704.

(53) Chiko, A. W. Evidence of multiple virion components in leaf- dip preparations of barley stripe mosaic virus. Virology 1975, 63, 115− 122.

(54) Saunders, K.; Lomonossoff, G. P. In planta synthesis of designer-length tobacco mosaic virus-based nano-rods that can be used to fabricate nano-wires. Front. in plant sci. 2017, 8, 1335.

(55) Burchard, W., Static and dynamic light scattering approaches to structure determination of biopolymers. In Laser light scattering in biochemistry, Harding, S. E.; Sattelle, D. B.; Bloomfield, V. A., Eds. Royal Society of Chemis: Cambridge, 1992; 3−22.

(56) Santos, N. C.; Castanho, M. A. Teaching light scattering spectroscopy: the dimension and shape of tobacco mosaic virus. Biophys. J. 1996, 71, 1641−1650.

(57) Green, M. R., Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory 2012.

(58) Chairatana, P.; Chu, H.; Castillo, P. A.; Shen, B.; Bevins, C. L.; Nolan, E. M. Proteolysis triggers self-assembly and unmasks innate immune function of a human α-defensin peptide. Chem. Sci. 2016, 7, 1738−1752.

This paper is published (Lee et al, Engineering Tobacco Mosaic Virus and Its Virus-Like-Particles for Synthesis of Biotemplated Nanomaterials, Biotechnology Journal, https://doi.org/10.1002/biot.202000311, 2020). The numbering of the figures, tables, and references have been modified to fit the format of this thesis.

Engineering Tobacco Mosaic Virus and Its Virus-Like-Particles for Synthesis of Biotemplated Nanomaterials

Kok Zhi Lee, Vindula Basnayake Pussepitiyalage, Yu-Hsuan Lee, L. Sue Loesch-Fries, Michael T. Harris, Shohreh Hemmati, and Kevin V. Solomon*

Abstract

Biomolecules are increasingly attractive templates for the synthesis of functional nanomaterials. Chief among them is the plant tobacco mosaic virus (TMV) due to its high aspect ratio, narrow size distribution, diverse biochemical functionalities presented on the surface, and compatibility with a number of chemical conjugations. These properties are also easily manipulated by genetic modification to enable the synthesis of a range of metallic and non-metallic nanomaterials for diverse applications. This article reviews the characteristics of TMV and related viruses, and their virus-like particle (VLP) derivatives, and how these may be manipulated to extend their use and function. A focus of recent efforts has been on greater understanding and control of the self-assembly processes that drive biotemplate formation. How these features have been exploited in engineering applications such as, sensing, catalysis, and energy storage are briefly outlined. While control of VLP surface features is well-established, fewer tools exist to control VLP self-assembly, which limits efforts to control template uniformity and synthesis of certain templated nanomaterials. However, emerging advances in synthetic biology, machine learning, and other fields promise to accelerate efforts to control template uniformity and nanomaterial synthesis enabling more widescale industrial use of VLP-based biotemplates.

1. Introduction

Nanoscale materials offer precise tuning of material properties through atomistic control of matter and energy interactions at increasingly small length scales. This precision in material properties enables vast new opportunities in sensing, data storage, energy storage, and catalysis, among other areas.[1–5] Key to this process is the synthesis of nanoscale materials with well-defined and uniform architectures.[4,6] Traditional chemical and physical synthesis technologies rely on purely materialand energy-intensive processes that are difficult to control and scale, rely on toxic or otherwise non-green chemicals, have non-uniform outputs, and limited control of atomistic features.[7–10] In contrast, biology synthesizes uniform nanoscale biomolecules via well-defined design rules, which may be engineered and serve as biotemplates for the synthesis of metallic nanomaterials.[11,12] Biomolecules, such as nucleic acids, microtubules, amyloid fibers, and viruses, have been used as scaffolds for the construction of hierarchical complex nanomaterials.[13–17] Their surfaces present diverse biochemical functionalities that are used to organize nanoparticle synthesis and may be modified via conjugation with organic or inorganic materials to create novel devices and control metal mineralization.[18,19] Finally, biomolecules possess well-defined nanoscale architectures, are structurally stable across a wide range of conditions, and can be easily manipulated via genetic engineering. All these features make biomolecules attractive biotemplates for bottom-up nanomaterial assembly. Viruses possess many advantages over other types of biomolecules for nanoparticle synthesis as they occur in a wide range of shapes and sizes, and present diverse chemical functionalities for nanoparticle synthesis. Plant viruses are widely used because they are harmless to human beings.[11] For instance, cowpea chlorotic mottle virus, cowpea mosaic virus, and brome mosaic virus form icosahedral structures that range in size from 18 to 30 nm while tobacco mosaic virus (TMV) and barley stripe mosaic virus (BSMV) assume rod-shaped structures up to 300 nm in length.[11] This diversity enables biotemplating of diverse nanomaterials for incorporation as catalysts, sensors, battery anodes, and semiconductor digital memory devices.[1–5] Viral particles consist of self-assembled capsid proteins (CPs) and nucleic acids that genetically encode the CPs. The CPs present diverse biochemical functionalities via amino acid residues on the particle surface that interact with metals in solution and drive nanoparticle synthesis. These residues may be conjugated to other compounds to enable synthesis of different nanomaterials and create novel functional properties.[18,19] Similarly, the presented protein functionalities and dimensions can be directly modified via engineering the encoding nucleic acid sequence without dramatically altering the viral structure to enable synthesis of new materials.[1,18,20] Non-infectious virus-like particles (VLPs) may be generated via heterologous expression of CPs in non-native species without using the complete viral genome.[21] Expressed CPs spontaneously self-assemble into VLPs that possess the same rich chemical diversity on their surfaces to drive nanoparticle synthesis. Plant VLPs also offer several compelling features over real viruses for VLP engineering and industrial-scale production. First, VLPs are more tolerant of mutations than live viruses enabling more engineering opportunities to enhance function. For example, genetic modifications that enhance particle structural stability to improve nanomaterial synthesis yields enable the formation of nucleic acid-free VLPs that are unable to infect host cells.[22] VLP production does not rely on infection for production and may be stably produced with this enhancement in a heterologous host. Second, heterologous microbial hosts replicate and produce CPs much more rapidly than plants, which need several weeks to grow and mature before infection with the virus for production.[23] Moreover, live viruses are infectious agents and must be grown in a biosafety level 2 greenhouse by plant virologists, to contain potential environmental contamination, adding to their costs.[24] Bacterial VLP production also leverages a wealth of bioprocessing infrastructure that has been developed for large scale production of food, pharmaceuticals, and chemicals.[25] However, bacterially-produced TMV VLPs display aberrant assembly properties. Unlike plant-produced TMV, bacterial TMV VLPs are unable to self-assemble into nanorods in the absence of an appropriate nucleic acid except under acidic conditions ($pH < 5.5$; see Section 2 for a description of viral assembly).[26] This may be attributed to the limited capacity for bacteria to post-translationally acetylate the N-terminal residue as is found in the plant-produced virus.[27,28] Nonetheless, in the presence of an appropriate nucleic acid, bacterially-derived VLPs assemble into nanorods that are suitable for metal coating.[12,18,29] Thus, VLPs are more compelling platforms for the development of viral biotemplates. TMV is widely used for biotemplating due to its architecture and physicochemical properties (Table 1). The dimensions of TMV are well suited to biotemplating applications such as the production of batteries and sensors.[1,6,30] TMV is self-assembled around a single nucleic acid from over a thousand copies of a single CP into a 300 nm long nanotube whose inner and outer diameters are 4 and 18 nm, respectively.[31] This aspect ratio maximizes the available surface area in compact volumes enabling more efficient battery electrodes with higher charge densities and increased sensitivity to chemical analytes as sensors.

Moreover, the biochemical/physicochemical properties of TMV enable reduction of metal ions and nanoparticle synthesis on the template under ambient conditions.[3,32] Finally, TMV and its VLPs consist of a single CP that is amenable to genetic and chemical modifications that expand the types of nanomaterials that may be synthesized, enhance morphological uniformity, and increase particle density.[29] While TMV is the most commonly used in bionanotechnology, the evolutionarily-related BSMV provides a promising alternative biotemplate.[16] BSMV has a similar architecture to TMV (1) but presents distinct surface functional groups that accelerate nanoparticle synthesis and increase nanoparticle density for increased electrical and thermal conductivities and analyte sensitivity as sensors.[16] BSMV particles encapsidate one of three nucleic acids to produce a distribution of particles at one of three lengths (108, 125, 148 nm).[33] The evolutionary similarities between TMV and BSMV may allow successful engineering strategies from TMV to be applied in BSMV to expand and enhance the properties of BSMV-derived biotemplates. Thus, BSMV is emerging as an attractive virus biotemplate for nanomaterial synthesis.

The convergence of advances from materials science, structural biology, molecular biology, chemistry, machine learning, and synthetic biology, now enable the rapid engineering and development of TMV and its VLPs for biotemplating. Their physicochemical properties make them well-suited for the synthesis of diverse nanomaterials for applications such as catalysis, energy storage, and sensing. In this review, we provide an overview of the properties and use of TMV, and its VLPs for nanoparticle synthesis, and focus on emerging technologies, approaches, and opportunities to engineer VLPs to enhance their function and broaden the nanomaterials that may be synthesized.

2. Viral Particle Self-Assembly and Metal Nanoparticle Synthesis

Viral biotemplates such as TMV self-assemble in the presence of nucleic acid molecular rulers from identical CP subunits due to several covalent, hydrogen, and electrostatic interactions encoded within the CP primary sequence and its associated nucleic acids (Figure 1).[37] These interactions must be strong enough to withstand the pHs, temperatures, and ionic strengths required for successful nanoparticle synthesis. CP is first translated and folded before ultimately forming flat disks via hydrophobic interactions between residues on the CP surface.[37] An RNA sequence naturally found in the TMV viral genome, known as the origin of assembly sequence (OAS), assumes a hairpin secondary structure that then serves as a nucleus for nanotube formation. CP disks are threaded by OAS-containing RNA, allowing for electrostatic interactions between the OAS and CP.[38] This interaction forces the CP disks to assume a helical conformation, reorienting themselves by induced proton adsorption and subsequent hydrogen-bonding, which then rapidly polymerizes with other CP molecules into a nanotube.[37] Strengthening the final assembly are electrostatic interactions between adjacent CPs, which are mediated by a handful of negatively charged residues in a motif known as the Caspar carboxylate cluster.[39] Calcium ions typically neutralize the repulsive negative charges of adjacent CPs and drive viral polymerization. Control of these interactions via genetic engineering of the CP and/or the OAS offers tremendous potential to modify nanotube architecture and its stability in subsequent biotemplating.[22,40] Metal nanoparticles are frequently synthesized spontaneously on viral biotemplates using aqueous metal solutions. The metal precursor ions adsorb and are chemically reduced on the viral particle surfaces at many adsorption and nucleation sites to form a metallic nanomaterial.[41,42] The chemical interactions that drive metal precursor adsorption and reduction to a metal deposit are not well understood. However, the adsorption process is frequently described by a single-step Langmuir isotherm that is solely driven by covalent interactions, for example, palladium on TMV.[41] As metal ion precursor adsorption and reduction are the fundamental processes that drive metal coating formation, the oxidation potential of surface accessible residues must be sufficient to drive metal reduction (Table 2). Amino acid residues that are easily oxidized, such as, cysteine, tyrosine, and lysine, more readily interact with metals driving deposition.[43–45] Metals with higher positive reduction potentials, including gold, silver, and platinum, can be reduced by the various functional groups present on the CP of TMV (Table 2).[42] This deposition is frequently enhanced by engineering the amino acid residues that are presented on the virus/VLP.[29,42] Other metals such as nickel, iron, and cobalt cannot be reduced this way as they have negative reduction potentials. Instead, a different metal that is more readily reduced such as palladium is mineralized first onto the CP, which then catalyzes to reduce target metal ions to metal atoms.[46,47] For example, nickel and cobalt are deposited in the inner channel of TMV after mineralization of TMV with Pd and Pt.[47] Fundamentally, an appropriate pairing of amino acid side chains that can chemically reduce and interact with metals intrinsically sets the metal adsorption capacity and controls the rate of reaction as nanoparticle synthesis proceeds spontaneously under ambient conditions. Plant-produced BSMV has been demonstrated to be a viable biotemplate for mineralization of palladium nanowires, however, biomineralization with BSMV differs from that of TMV.[16] The surface of BSMV allows metal ion precursor deposition to proceed via a multi-step Langmuir isotherm that incorporates both electrostatic and covalently adsorbent-adsorbate interactions. This difference may arise in part due to the larger amount of BSMV surface-exposed residues, compared to TMV, in an unstructured insertion loop containing 10 amino acids, that protrudes from the particle surface.[36] These stronger interactions increase the adsorption capacity for Pd on BSMV twofold compared to that on TMV.[16] Similarly, the rate of adsorption is increased compared to TMV, suggesting that BSMV can be fully coated in fewer processing cycles saving both time and expensive precursor material. Furthermore, BSMV biotemplates produce more uniformly sized nanoparticles relative to TMV. The additional opportunities to engineer metal deposition via the insertion loops and superior adsorption and metal nanoparticle synthesis characteristics make BSMV an attractive alternative to TMV that may generate more uniform metal nanostructures more economically.

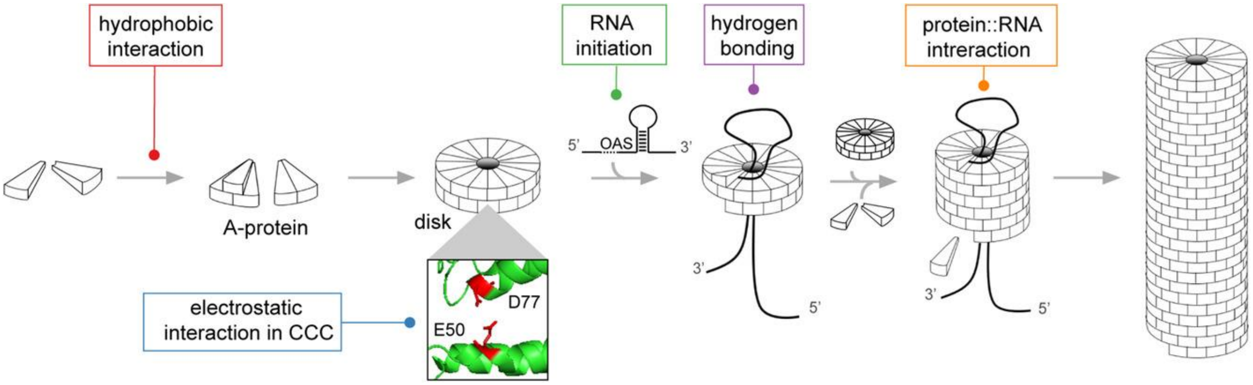
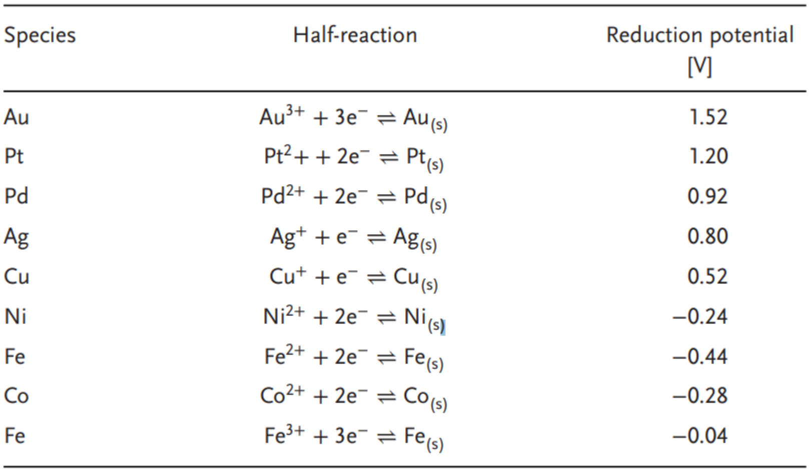


Figure 1. Driving forces for TMV assembly. TMV CP experiences various interactions, including hydrophobic interactions, RNA initiation, electrostatic interaction in Caspar carboxylate cluster (CCC), hydrogen bonding, and RNA: Protein interaction at different stages of assembly.

Table 2. Reduction potential of some metals at 25 C.[48].



Species	Half-reaction	Reduction potential [V]
Au	$Au^{3+} + 3e^{-} \rightleftharpoons Au_{(s)}$	1.52
Pt	$Pt^{2}+ + 2e^{-} \rightleftharpoons Pt_{(s)}$	1.20
Pd	$Pd^{2+} + 2e^{-} \rightleftharpoons Pd_{(s)}$	0.92
Ag	$Ag^{+} + e^{-} \rightleftharpoons Ag_{(s)}$	0.80
Cu	$Cu^{+} + e^{-} \rightleftharpoons Cu_{(s)}$	0.52
Ni	$Ni^{2+} + 2e^{-} \rightleftharpoons Ni_{(s)}$	−0.24
Fe	$Fe^{2+} + 2e^{-} \rightleftharpoons Fe_{(s)}$	−0.44
Co	$Co^{2+} + 2e^{-} \rightleftharpoons Co_{(s)}$	−0.28
Fe	$Fe^{3+} + 3e^{-} \rightleftharpoons Fe_{(s)}$	−0.04

3. Nanostructure Synthesis with Viral Biotemplates for Diverse Applications

Nanostructures are increasingly sought to create advanced materials that improve electrical conductivity and capacitance, current generation, catalytic activity, and detection sensitivity.[3,5,19,49,50] These applications benefit from high aspect ratio nanomaterials that can be biotemplated on TMV or its VLPs. However, each application relies on distinct metals whose unique electronic structure is key to enabling these applications, for example, metal with high conductivity such as copper is used in conductive nanowires while gold is relatively inert making it an ideal substrate for a sensor.[19,51] Moreover, the precise function of the nanomaterial in these roles relies on their incorporation into devices with unique architectures (Figure 2). To enable the synthesis of these nanomaterials, the viral biotemplates are modified and processed in innovative ways. We describe only a handful below.

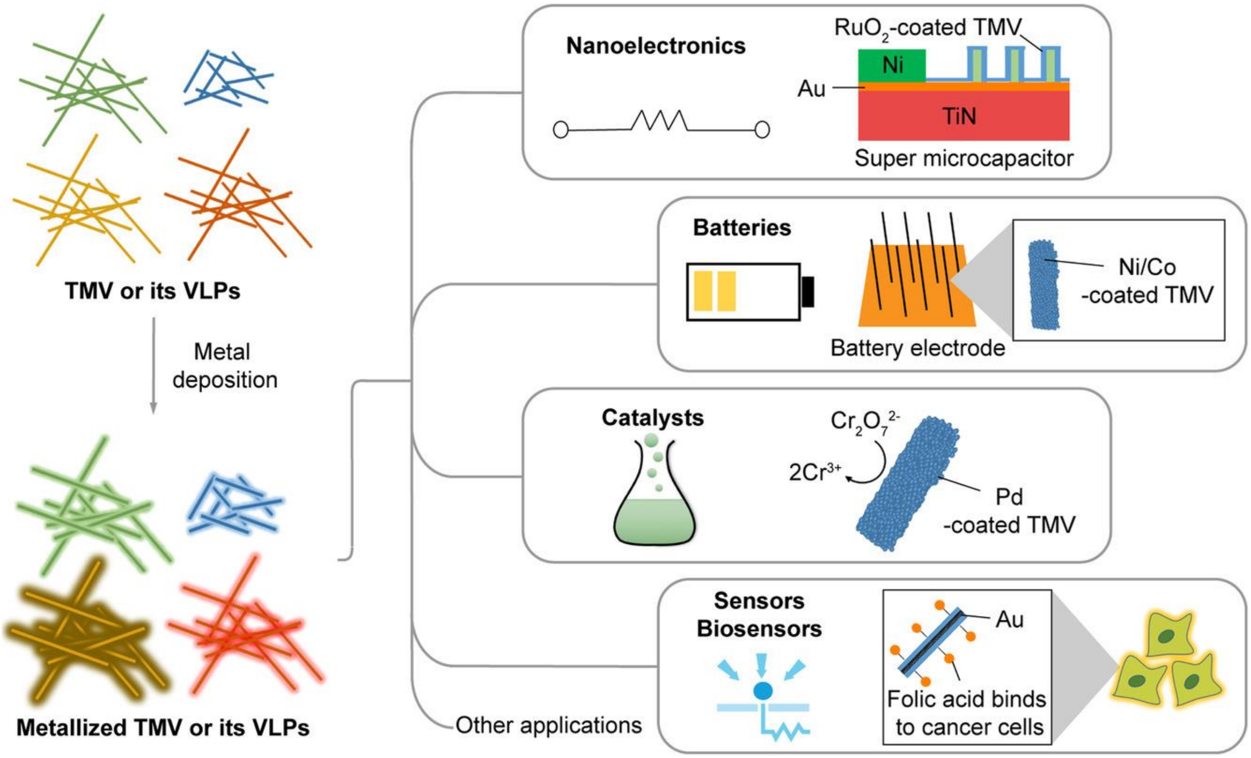


Figure 2. Potential applications of VLPs. Nanorod VLPs such as those made from TMV can be coated and functionalized with metals and other small molecules. The specific functional properties of this coating and the orientation of the templated nanorods on functionalized surfaces confer unique properties for diverse applications such as those above.

3.1. Nanoelectronics

Nanoelectronics, such as those found in modern sensors and lasers, use metal nanowires due to their compact size and improved electrical conductivity.[52] Industrial-scale production of metal nanowires is most commonly achieved via polyol synthesis. However, it is not environmentally-friendly or sustainable as it uses various toxic polyhydric reducing agents such as ethylene glycol and propylene glycol in an energy-intensive reaction process that occurs at high temperatures.[8,53–55] In contrast, highly conductive and uniform nanowires may be synthesized without toxic reagents or high temperatures via templating on TMV or its VLP.[56] To enable the creation of palladium, platinum, or gold nanowires, additional lysine, cysteine, and/or tyrosine side chains are added to the surface of the viral template to provide sufficient reducing power for nanoparticle synthesis and potential nucleation sites for metal binding. The introduction of this reduction of power can decrease or completely remove the need for additional reducing agents in nanowire synthesis, thereby lowering costs.[32,57] For thinner nanowires, the metal may be deposited in the narrow viral channel of TMV, rather than on the surface.[58,59] Buffering agents are used to alter the ionic state of amino acid side chains, changing their electrostatic interactions to decrease metal deposition on the surface. The inner channel, whose side chains are distinct from the surface, remains charged and electrostatically interacts with metal solutions to reduce and adsorb synthesized metal nanoparticles. These techniques may be directly applied to platinum or palladium nanowires; however, nanowires composed of other metals such as copper, cobalt, and/or nickel require a core of Pd and Pt, which then serves as catalytic sites for reduction of other metals forming thin alloy nanowires such as CoFe, CoNi, FeNi, and CoFeNi.[47,58] Bimetallic ferromagnetic alloys with Pt such as CoPt and FePt3 may also be synthesized directly by mixing precursor metal salt solutions with platinum or palladium salts in the same deposition bath thus simplifying the two-step process.[46]

3.2. Batteries and Energy Storage

Batteries produce current via chemical reactions on the surface of electrodes, which is limited by the available surface area.[3] To generate the currents needed to power modern electronic devices, metal coated TMV virions can be vertically oriented on top of a gold substrate to form a carpet-like nanoforest structure (2). The vertical attachment is achieved by introducing a cysteine residue at the amino terminus of CP to produce TMV1cys virions.[3,6] Most cysteines in TMV are partially recessed and thus are incapable of metal-binding; however, the N-terminal residue is fully exposed allowing for near-vertical assembly of the biotemplate on a gold substrate via covalent interactions between the gold and the thiol group of the cysteine. Subsequent coating with nickel and/or cobalt increases the active electrode surface area more than tenfold with a doubling in electrode discharge capacity (or current generation).[3] Similarly, Pd/Ni/Sicoated TMV anodes in lithium ion batteries can increase the discharge capacity by nearly tenfold compared to then available graphite anodes.[6] Finally, incorporating TMV in sodium-ion batteries as a carbon/tin/nickel-coated TMV anode can increase battery cycling lifespan with little degradation in charge capacity over 150 deep charging cycles. This capability made it the longest-cycling nano-Sn anode material for Na-ion batteries at the time.[30] Coated TMV patterned with a similar nanoforest structure has also been shown to increase the performance of microsupercapacitors (2), which also rely on the surface area to store charges.[60]

3.3. Catalysis

Catalyst effectiveness is proportional to its exposed surface area, as more sites are available for reaction. To minimize the costs of large volumes of catalytic metal, an economical strategy for catalyst synthesis is to deposit catalytic metal on an inexpensive inert substrate with a high surface area to volume ratios such as carbon or a biotemplate. However, biotemplated catalysts are more efficient; Pd catalyst synthesized from TMV1cys immobilized on a gold substrate in a nanostar configuration increased the reaction rate 68% relative to commercial Pd/C catalyst of comparable size.[61] TMV-templated catalysts are stable and can be recycled with negligible degradation in performance over several cycles.[4,5,61] The superior performance of TMV-templated catalyst is attributed primarily to two factors. One, palladium particles mineralized on TMV tend to be more uniform as compared to commercially available palladium catalyst supported on carbon material leading to more efficient catalysis.[61,62] And second, fabrication of palladium on TMV does not use surfactants and capping agents like the synthesis of traditional catalysts that can block active sites.[63] Removal of these surfactants to rescue catalysis is incomplete and adds to the cost of catalyst manufacture. Thus, biotemplated catalysts are both more catalytically- and economically-advantaged in many scenarios.

3.4. Sensors

Sensors are used in a wide range of areas including environmental and disease monitoring to detect and quantify selected analytes from complex dirty solutions. Effective sensors must have high selectivity, sensitivity, and structural stability over the range of environmental conditions they are likely to experience. Selectivity for a specific analyte in mixed samples is achieved by functionalizing the viral surface with metals, enzymes, and/or other chemicals. For example, palladium-coated TMV rapidly, and reversibly detects hydrogen at room temperatures.[50] Similarly, gold-coated nanowires conjugated with folic acid via coupling agents can interact with the folate receptors on tumor cells to detect cancers with high sensitivity.[19] Additionally, TMV may be decorated with biotin to enable binding of enzymes to detect penicillin.[64] Other desirable sensor properties such as sensitivity and stability arise from the nature of viral biotemplates themselves. TMV’s high surface area to volume ratio allows for more contact with analytes, amplifying signal generation, and improving device sensitivity.[65] Similarly, TMV structure is very stable in high temperatures and over a wide range of pHs (Table 1) that may be encountered in the analysis of complex mixtures, making biosensors derived from TMV more durable and rugged.

4. Engineering Viruses and Virus-like Particles for Enhanced Biotemplate Properties

Improved understanding of TMV structural biology and molecular biology has informed several strategies to control and enhance biotemplate properties. These include improved structural stability to enable synthesis of more diverse nanomaterials, control of VLP architecture, and enhanced uniform coating to control physical properties and more versatile conjugation capabilities to tune function. These modifications are powered by the flexibility of the CP-encoding nucleic acids, which can be readily engineered via genetic tools.

4.1. Structural Stability in High Temperatures and Extreme pH

Despite the successes in coating TMV and its VLPs with metals such as gold, palladium, and platinum, many industriallydesirable metals are not readily deposited due to unfavorable electrochemistry (2).[29,66] This issue may be mitigated somewhat through the use of buffering and reducing agents, activation by primary deposition with another metal such as Pt, and higher temperatures to create more favorable processing conditions. These strategies ultimately increase metal-template interactions via altering residue ionization state, changing the reduction potentials of the metal ions and amino acid side chains, and/or create nucleation sites to drive metal mineralization. However, these conditions may destabilize the CP interactions that drive the self-assembly of the template leading to biotemplate loss and low yields of metal mineralization. Biotemplates are only stable within fixed pH and temperature ranges (1). To address this challenge, TMV and VLPs have been engineered to increase their structural stability to resist disassembly. The stability of TMV biotemplates has been increased by control of the intermolecular forces that drive self-assembly (1), enabling more rapid biotemplating of a wider range of materials.[67–69] For example, single point mutations within the Caspar carboxylate cluster can enhance viral assembly. Neutralization of a negatively-charged residue (E50Q) or replacing a negatively-charged residue with a positively-charged residue in the Caspar carboxylate center (E50R or D77R) has been shown to produce longer virions that spontaneously self-assemble without an RNA that contain an OAS or other nucleic acids in transgenic plants.[70] Similarly, neutralizing the negative-charged residues with site-specific mutations, E50Q/D77N, was sufficient to rescue TMV-VLP assembly in E.coli even in the absence of RNA with an OAS, which typically results in viral disks alone.[18] In combination with cysteine engineering, E50Q/D77N mutants can attach to a gold-coated plate.[18] These mutations within the Caspar carboxylate cluster are sufficient to overcome the poor stability of nucleic acid-free VLPs and recover resistance to pH, attaining similar stability to wildtype virions. In so doing, they create nucleic acid-free VLPs that have a free internal channel for the synthesis of thin nanowires.[67] Engineering of the hydrophobic interactions between CPs (1) may further improve the stability of VLPs; however, modifications of this nature have yet to be evaluated. OAS-containing nucleic acids that initiate the assembly of wildtype virus and VLPs also act as a molecular ruler that sets the length of produced VLPs.[20] While VLPs that have been engineered appropriately (e.g., at the Caspar carboxylate cluster) can self-assemble without this molecular ruler, the resulting VLPs are more heterogeneous in size than OAS-containing viral particles with sizes as small as 20 nm.[22] Thus, future work should include efforts to control the precise dimensions of VLPs or the development of efficient separation technologies to isolate VLPs with specific dimensions.

4.2. Coating Uniformity

Nanomaterial properties are controlled in part by the coating uniformity or smoothness of the surface of metal deposited on biotemplate. This uniformity relies on the molecular interactions between the deposited metal and residues on the surface or in the channel of the biotemplate. Coating agents such as citric acid are added to the deposition reaction to alter the charge of surface-exposed residues and increase metal-biotemplate interactions during the metal deposition process.[42,71,72] However, these pH-sensitive electrostatic interactions can be disrupted by other reagents of the deposition process leading to poor nanoparticle synthesis and nonuniform coating. To address this challenge, covalent bonds that are resistant to modest pH changes can be introduced to drive nanoparticle synthesis. Site-specific mutagenesis of the coat protein to introduce cysteine and lysine residues have been demonstrated to improve nanoparticle synthesis.[29,73] These residues may be introduced either on the surface or inner channel of TMV VLPs (e.g., S3C, T103C, T158K) or either terminus of the CP.[73–75] These modifications allow for metal deposition under a wider range of pHs and promote high-density nucleation sites for nanoparticle synthesis, unlike wildtype whose nanoparticle synthesis is driven by weaker surface drying effects. For example, two cysteine residues have been genetically added to the amino-terminus of TMV CP, resulting in 4260 available thiol groups per VLP for metal reduction and deposition[29]. The introduced thiol groups on the TMV surface serve as a reaction site for nanomaterials such as gold and palladium to reduce metal ions or form covalent bonds with non-metal materials via thiol coupling. Cysteine enhances the deposition of several metals including gold, silver, and palladium, and also enables novel attachments for device manufacture (e.g., nanoforests for battery anode synthesis).[29] Lysine substitution has also served as a functional group for chemical conjugation of silica coatings.[73]

4.3. Diversification of Surface Characteristics

While TMV/BSMV virions and VLPs serve as exceptional biotemplates for metal mineralization, they are not directly compatible with the synthesis of non-metallic nanomaterials. However, the introduction of surface-accessible amino acids, including aspartate, glutamate, and tyrosine provides conjugation sites for the synthesis of non-metallic nanomaterials via covalent bonding. For example, the carboxylic acid moiety of aspartate and glutamate can react with amines, which have been functionalized with biotin, chromophores, and crown ethers while tyrosine can conjugate with PEG.[76] These conjugations enable medical applications of TMV such as vectors for drug delivery. In addition to canonical amino acids, synthetic noncanonical amino acids with unique functional groups have also been incorporated in other VLPs such as the hepatitis B virus or bacteriophage Qβ to expand coating capabilities further.[77] Noncanonical amino acids provide unique functional groups, such as alkynes, and aminophenylalanine, that facilitate bioorthogonal coating.[78] Bioorthogonal reactions enable specific conjugation via the noncanonical amino acids without cross-reaction with canonical amino acids of the virions, reducing the risk of disrupting the physicochemical properties that drive biotemplate self-assembly. Alkyne-containing noncanonical amino acids have been incorporated into virions to attach PEG, oligonucleotides, antibody fragments, and other peptides via click chemistry,[77] while, aminophenylalanine incorporated VLPs, have been used to conjugate cell type-specific targeting peptides, enabling drug delivery to specific cell types.[79] The number of novel viral-derived biomaterials are certain to expand further as new noncanonical amino acids are engineered and incorporated into CPs.[80]

5. Future Directions

Current advances in many fields now converge to enhance the function and potential of viral biotemplates. These activities are greatly accelerated by the innovations of synthetic biology that enable more rapid building and testing of engineered VLPs. These constructs are thoroughly characterized by spectroscopic techniques with the rich datasets fed to machine learning algorithms to guide iterative design and improvements (Figure 3). Several emerging opportunities are as follows.

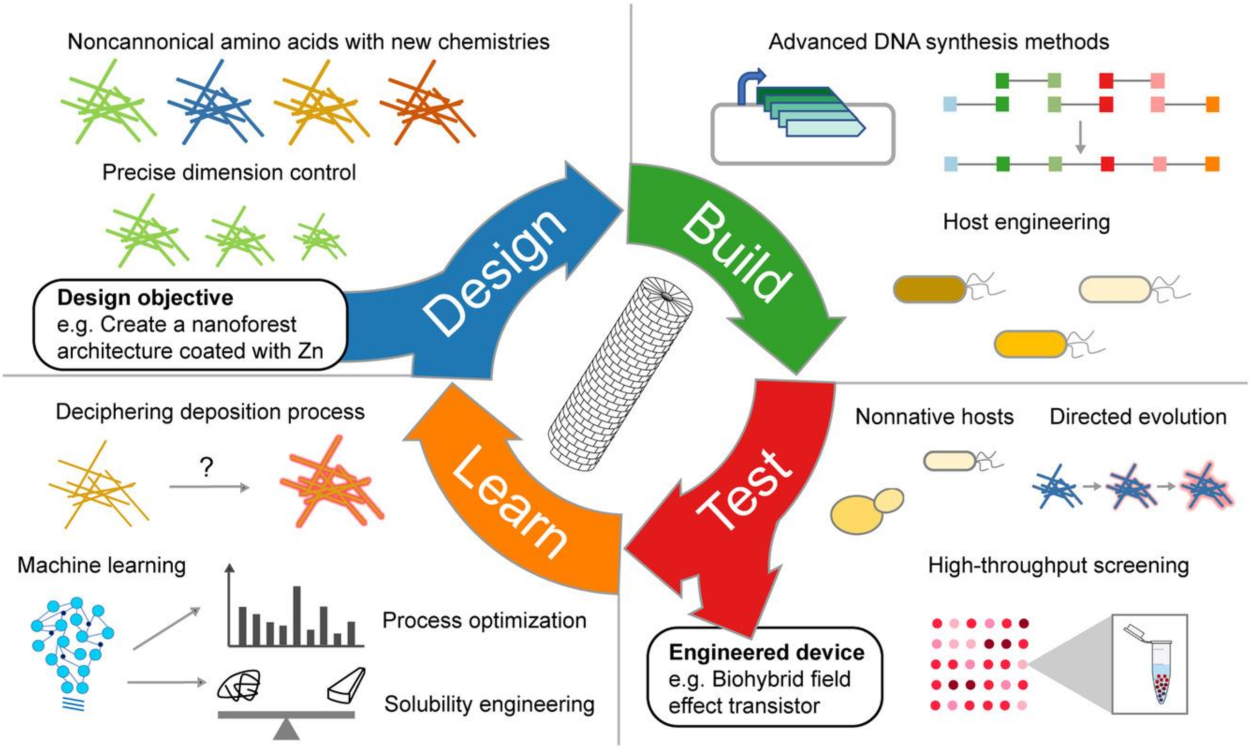

Noncannonical amino acids with new chemistries
Precise dimension control
Design objective
e.g. Create a nanoforest architecture coated with Zn
Design
Build
Advanced DNA synthesis methods
Host engineering
Test
Nonnative hosts
Directed evolution
High-throughput screening
Engineered device
e.g. Biohybrid field effect transistor
Learn
Deciphering deposition process
?
Machine learning
Process optimization
Solubility engineering

Figure 3. Emerging opportunities to accelerate the development of TMV,anditsVLPs. In the "design" phase, VLPs with specific properties (e.g., conjugation to an enzyme via noncanonical amino acids) for an application guide the design of DNA sequences that are then rapidly assembled in the "build" phase. Advanced DNA synthesis methods rapidly assemble DNA to produce thousands of variants while host engineering boosts VLP production. In the "test" phase, these constructs are expressed and screened. Directed evolution develops VLPs with desirable features while cell-free systems enable high-throughput screening. In the "learn" phase, characterization of the resulting products and analysis via empirical and machine learning approaches refine the initial design ultimately leading to an optimal biotemplate and bioproduction platform.

5.1. Rapid Prototyping and Screening of Novel Biotemplates

Despite the exciting future for VLP development, the production and characterization of newly engineered VLPs are timeconsuming processes, spanning several days. Thus, new screening tools are necessary to expedite the development of newly engineered VLPs. Cell-free systems that contain the transcriptional and translational machinery needed for VLP production could become integral to rapid VLP prototyping and characterization without laborious purification. Cell-free systems can be used to produce engineered VLP candidates in a highly parallel or high-throughput manner with minimal inputs due to their microliter scale reaction volumes.[81] VLPs derived from non-plant viruses are already being developed via cell-free technologies. For example, a cell-free system was used to incorporate noncanonical amino acids in bacteriophage MS2 and bacteriophage Qβ for click-chemistry functionalization.[82] Rapid screening is also needed to identify engineered mutants with desirable properties. Advanced DNA synthesis methods such as Gibson and Golden-gate assemblies can produce thousands of DNA-encoding VLPs variants in a single reaction.[83–85] However, screening of these variants for improved function requires the characterization of individual mutants, which makes the screening process slow and inefficient. Instead of screening every variant, directed evolution can be applied to evolve mutants with desirable traits.[86] Directed evolution is a process where mutants are propagated and those with desirable properties are selected for. The ability of a given mutant to propagate or replicate is linked to the property that is desired, allowing surviving mutants to "select" for enhanced properties. Thus, thousands of variants can be simultaneously evaluated in hours without screening every single variant. The selected variants can then be mutated to introduce additional genetic diversity and subject to another round of selection. The iteration of these processes can generate protein mutants with optimized or even new properties such as producing VLPs with enhanced structural stability.[87] Although directed evolution has not been widely used for VLP engineering, the recent development of new directed evolution methods, including assisted machine learning,[88] which identifies more promising mutations to be constructed and screened, and in vivo continuous directed evolution,[89] which streamlines the genetic engineering process, will accelerate the engineering of VLPs.

5.2. Improved Biotemplate Production Yields

VLPs frequently form insoluble aggregates that are unsuitable for biotemplating processes when expressed in non-native hosts.[90] To increase the yield of usable VLP biotemplates, the solubility of heterologous CP must be enhanced. There are many ways to increase protein solubility;[91] however, increasing CP solubility for VLPs production is challenging as one of the drivers of protein insolubility, strong hydrophobic interactions, is also the driving force for virion and VLP assembly. However, machine learning algorithms have begun to identify the design rules for soluble VLP CP design without disrupting VLP selfassembly based on hydrophobicity scales.[92] An early report with these tools counterintuitively suggests that CP solubility may be highly correlated with arginine content in some VLPs, despite its negative correlation with a solubility of short-chain variable fragments. Arginine reduces many hydrophobic protein-protein interactions while interacting favorably with other amino acids to reduce random aggregation. However, these tools have yet to be applied to forward engineer novel VLP CPs.

5.3. Molecular Understanding and Control of the Mineralization Process

While the basic processes of metal mineralization are qualitatively understood, mechanistic details regarding the reaction mechanisms of metal reduction and adsorption remain elusive. Current studies on metal mineralization using TMV, BSMV, and their corresponding VLPs as biotemplates focused on the characterization of synthesized nanostructures using different techniques such as transmission electron microscopy,[24] Fourier transform infrared spectroscopy (FTIR), and X-ray scattering analysis.[19,62] These methods reveal the structure of the final synthesized structure and inform hypotheses of how metal mineralization has occurred. However, a more effective approach to understanding how metal mineralization takes place is to perform in situ FTIR.[93] This allows direct observation of the reaction progress enabling determination of the mechanism of the mineralization. It is also possible to observe how changes in reaction conditions such as pH, temperature, and concentrations of precursors, reducing agents and biotemplate would affect metal mineralization including particle size, particle size homogeneity, and the type of metal nanostructures mineralized on the surface. This would require an extensive design of experiments to systematically evaluate the effect of each parameter and their interactions. Machine learning algorithms such as artificial neural networks have already been applied to these rich datasets to create predictive models for nanoparticle synthesis as a function of processing parameters.[94]

Similarly, neural networks first used to predict the binding of metallic ion cofactors to enzymes could be extended to VLPs to predict and model metal-biotemplate interactions as a function of engineered CP protein sequence.[95] Such computational tools would greatly accelerate biotemplate engineering efforts and optimize deposition processes for metallic nanomaterial synthesis. In addition to empirical studies of the deposition process, much progress may be made with improved theoretical models of metal deposition. While theoretical studies of TMV mineralization have yet to be conducted, the metal selectivity of other viral templates can be partially explained via theory. For example, first-principle density functional theory calculations of the binding energies between E3-M13 virus surface residues and metal precursor complexes successfully predict the deposition of Pd and Ag nanoparticles and explain the inability to template Au or Pt.[96] These results suggest that coulombic interactions between the metallic precursor and the template play an important role in the templating of metals on viruses. More generally, molecular dynamics (MD) simulations have successfully predicted the affinity of peptide sequences for various metals.[97–101] These simulations confirm the importance of electrostatic interactions for metal adsorption and highlight the impact of amino acid side chain geometry on these interactions. Moreover, these calculations suggest that metal adsorption is entropically driven. Application of these approaches to TMV biotemplating of nanoparticles would provide an in-depth molecular-level understanding of the physio-chemical and biological reactions governing biomineralization and guide template engineering efforts for truly bottom-up nanofabrication.

6. Concluding Remarks

Integrated concepts from material science, biology, chemistry, and engineering have developed TMV, its VLPs for the synthesis of high-quality nanomaterials for diverse applications. TMV, and emerging BSMV, are particularly attractive as biotemplates due to their ease of manipulation, large scale production in bacterial culture, and surface moieties that facilitate interactions with a number of materials. New insights afforded by structural biology and spectroscopic characterization now inform the engineering of CP molecular interactions within a viral particle/VLP, and between the viral particle/VLP and deposited nanomaterials. These have led to several innovative genetic strategies that enhance biotemplate structural stability and improve the uniformity and properties of synthesized nanoscale architecture enabling diverse applications such as sensing, catalysis, and energy storage. While further engineering is needed to scale up these systems for widespread industrial nanomaterial synthesis, emerging technologies from synthetic biology and machine learning promise to accelerate the pace of TMV- and BSMV- derived biotemplate development in the years to come.

Acknowledgements

This review is based upon work supported by the National Science Foundation under Grant No. CBET- 2028618. This article was partially supported by the USDA National Institute of Food and Agriculture (Hatch Multistate Project S1075) and the Robert B and Virginia V. Covalt Endowed Professorship of Chemical Engineering at Purdue. K.Z.L. was partially supported by Purdue Research Foundation Fellowships (#60000025 and #60000029) and a Ralph W. and Grace M. Showalter Research Trust Award (#41000622). V.B.P. was supported by startup funds from the College of Engineering, Architecture, and Technology at Oklahoma State University. Y.-H.L. was supported by the Robert B and Virginia V. Covalt Endowed Professorship of Chemical Engineering. The funding sources had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

V.B.P. and Y.-H.L. contributed equally to this work. K.Z.L.: Conceptualization (equal), visualization (lead), writing-original draft (lead), writing-review, and editing (lead); V.B.P.: Visualization (supporting), writing-original draft (supporting), writing-review, and editing (supporting); Y.-H.L.: Writing-original draft (supporting), writing-review, and editing (supporting); S.L.-F.: Supervision (equal), writing-original draft (supporting), writing-review, and editing (supporting); M.T.H.: Supervision (equal), writing-original draft (supporting), writing-review, and editing (supporting); S.H.: Supervision (equal), visualization (equal), writing-original draft (supporting), writing-review, and editing (supporting); K.V.S.: Conceptualization (lead), project administration (lead), funding acquisition (lead), supervision (lead), writing-original draft (lead), writing-review, and editing (lead).

Data Availability

Statement Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Keywords

barley stripe mosaic virus, biotemplate, nanomaterial, tobacco mosaic virus, virus-like particle

References

[1] C. Koch, K. Wabbel, F. J. Eber, P. Krolla-Sidenstein, C. Azucena, H. Gliemann, S. Eiben, F. Geiger, C. Wege, Front. Plant Sci. 2015, 6, 1137.

[2] N. G. Portney, R. J. Tseng, G. Destito, E. Strable, Y. Yang, M. Manchester, M. G. Finn, M. Ozkan, Appl. Phys. Lett. 2007, 90, 214104.

[3] E. Royston, A. Ghosh, P. Kofinas, M. T. Harris, J. N. Culver, Langmuir 2008, 24, 906.

[4] C. Yang, A. K. Manocchi, B. Lee, H. Yi, Appl. Catal., B 2010, 93, 282.

[5] C. Yang, A. K. Manocchi, B. Lee, H. Yi, J. Mater. Chem. 2010, 21, 187.

[6] X. Chen, K. Gerasopoulos, J. Guo, A. Brown, C. Wang, R. Ghodssi, J. N. Culver, ACS Nano 2010, 4, 5366. [7] R. Mueller, L. Mädler, S. E. Pratsinis, Chem. Eng. Sci. 2003, 58, 1969.

[8] Y. Sun, Y. Xia, Adv. Mater. 2002, 14, 833.

[9] S. Cattaneo, S. Althahban, S. J. Freakley, M. Sankar, T. Davies, Q. D. N. He, C. J. Kielyab, G. J. Hutchings, Nanoscale 2019, 11, 8247.

[10] M. A. K. Zak, Solid State Sci. 2012, 14, 488.

[11] Y. Zhang, Y. Dong, J. Zhou, X. Li, F. Wang, Molecules 2018, 23, 2311. [12] S.-Y. Lee, J.-S. Lim, M. T. Harris, Biotechnol. Bioeng. 2012, 109, 16.

[13] S. M. D. Watson, H. D. A. Mohamed, B. R. Horrocks, A. Houlton, Nanoscale 2013, 5, 5349.

[14] J. C. Zhou, Y. Gao, A. A. Martinez-Molares, X. Jing, D. Yan, J. Lau, T. Hamasaki, C. S. Ozkan, M. Ozkan, E. Hu, B. Dunn, Small 2008, 4, 1507.

[15] M. Malisauskas, R. Meskys, L. A. Morozova-Roche, Biotechnol. Prog. 2008, 24, 1166.

[16] O. O. Adigun, E. L. Retzlaff-Roberts, G. Novikova, L. Wang, B.-S. Kim, J. Ilavsky, J. T. Miller, L. S. Loesch-Fries, M. T. Harris, Langmuir 2017, 33, 1716.

[17] S. Chu, A. D. Brown, J. N. Culver, R. Ghodssi, Biotechnol. J. 2018, 13, 1800147.

[18] A. D. Brown, L. Naves, X. Wang, R. Ghodssi, J. N. Culver, Biomacromolecules 2013, 14, 3123. [19] Y. Qu, Y. Yang, R. Du, M. Zhao, Appl. Microbiol. Biotechnol. 2020, 104, 3947.

[20] K. Saunders, G. P. Lomonossoff, Front. Plant Sci. 2017, 8, 1335.

[21] A. Zeltins, Mol. Biotechnol. 2013, 53, 92.

[22] A. Kadri, C. Wege, H. Jeske, J. Virol. Methods 2013, 189, 328.

[23] H. Jeong, B. L. Seong, J. Microbiol. 2017, 55, 220.

[24] V. V. Makarov, E. V. Skurat, P. I. Semenyuk, D. A. Abashkin, N. O. Kalinina, A. M. Arutyunyan, A. G. Solovyev, E. N. Dobrov, PLoS One 2013, 8, e60942.

[25] K. Schügerl, J. Hubbuch, I. bioprocesses, Curr. Opin. Microbiol. 2005, 8, 294.

[26] D. J. Hwang, I. M. Roberts, T. M. Wilson, Proc. Natl. Acad. Sci. USA 1994, 91, 9067.

[27] K. Narita, Bioch. Biophys. Acta 1958, 28, 184.

[28] B. Polevoda, F. Sherman, J. Mol. Biol. 2003, 325, 595.

[29] S.-Y. Lee, E. Royston, J. N. Culver, M. T. Harris, Nanotechnology 2005, 16, S435.

[30] Y. Liu, Y. Xu, Y. Zhu, J. N. Culver, C. A. Lundgren, K. W. C. Xu, ACS Nano 2013, 7, 3627.

[31] J. M. Alonso, M. Ł. Górzny, A. M. Bittner, Trends Biotechnol. 2013, 31, 530.

[32] A. S. Freer, L. Guarnaccio, K. Wafford, J. Smith, J. Steilberg, H. N. Culver, M. T. Harris, J. Colloid Interface Sci. 2013, 392, 213.

[33] A. W. Chiko, Virology 1975, 63, 115.

[34] G. Oster, J. Biol. Chem. 1951, 190, 55.

[35] Ninth Report of the International Committee on Taxonomy of Viruses, in Virus Taxonomy (Eds: A. M. Q. King, M. J. Adams, E. B. Carstens, E. J. Lefkowitz), Elsevier, San Diego, USA 2012, pp. 1139– 1162.

[36] D. K. Clare, E. V. Pechnikova, E. V. Skurat, V. V. Makarov, O. S. Sokolova, A. G. Solovyev, E. V. Orlova, Structure 2015, 23, 1815.

[37] W. K. Kegel, P. van der Schoot, Biophys. J. 2006, 91, 1501.

[38] P. J. Butler, Philos. Trans. R. Soc., B 1999, 354, 537.

[39] H. Wang, A. Planchart, G. Stubbs, Biophys. J. 1998, 74, 633.

[40] C. Wege, C. Koch, WIREs Nanomed. Nanobiotechnol. 2020, 12, e1591.

[41] O. O. Adigun, G. Novikova, E. L. Retzlaff-Roberts, B. Kim, J. T. Millera, L. S. Loesch-Fries, M. T. Harrisa, J. Colloid Interface Sci. 2016, 483, 165.

[42] E. Dujardin, C. Peet, G. Stubbs, J. N. Culver, S. Mann, Nano Lett. 2003, 3, 413.

[43] T. J. Bechtel, E. Weerapana, Proteomics 2017, 17, 1600391.

[44] R. Aeschbach, R. Amadoò, H. Neukom, Biochim. Biophys. Acta, Protein Struct. 1976, 439, 292.

[45] M. Utrera, J.-G. Rodríguez-Carpena, D. Morcuende, M. Estévez, J. Agric. Food Chem. 2012, 60, 3917.

[46] R. Tsukamoto, M. Muraoka, M. Seki, H. Tabata, I. Yamashita, Chem. Mater. 2007, 19, 2389.

[47] M. Knez, A. M. Bittner, F. Boes, C. Wege, H. Jeske, E. Maiβ, K. Kern, Nano Lett. 2003, 3, 1079.

[48] S. G. Bratsch, J. Phys. Chem. Ref. Data 1989, 18, 1989.

[49] A. S. Freer, C. Gilpin, L. Mueller, M. Harris, Chem. Eng. Commun. 2015, 202, 1216.

[50] K. Srinivasan, S. Cular, V. R. Bhethanabotla, S. Y. Lee, M. T. Harris, Nanomaterial sensing layer based surface acoustic wave hydrogen sensors in Proceedings of IEEE Ultrasonics Symposium, 2005, 1, 645– 648. https://ieeexplore.ieee.org/document/1602935.

[51] J. C. Zhou, C. M. Soto, M.-S. Chen, M. A. Bruckman, M. H. Moore, E. Barry, B. R. Ratna, P. E. Pehrsson, B. R. Spies, T. S. Confer, J. Nanobiotechnol. 2012, 10, 18.

[52] S. M. Bergin, Y.-H. Chen, A. R. Rathmell, P. Charbonneau, Z.-Y. Lib, B. J. Wiley, Nanoscale 2012, 4, 1996.

[53] Y.-H. Choi, Y.-S. Chae, J.-H. Lee, Y. Kwon, Y. S. Kim, Electron. Mater. Lett. 2015, 11, 735.

[54] S. Hemmati, D. P. Barkey, L. Eggleston, B. Zukas, N. Gupta, M. Harris, ECS J. Solid State Sci. Technol. 2017, 6, P144.

[55] S. Hemmati, D. P. Barkey, N. Gupta, R. Banfield, ECS J. Solid State Sci. Technol. 2015, 4, P3075.

[56] M. Wnek, M. Ł. Górzny, M. B. Ward, C. Wälti, A. G. Davies, R. ˛ Brydson, S. D. Evans, P. G. Stockley, Nanotechnology 2013, 24, 025605.

[57] J.-S. Lim, S.-M. Kim, S.-Y. Lee, E. A. Stach, J. N. Culver, M. T. Harris, Nano Lett. 2010, 10, 3863.

[58] S. Balci, K. Hahn, P. Kopold, A. Kadri, C. Wege, K. Kern, A. M. Bittner, Nanotechnology 2012, 23, 045603.

[59] S. Balci, A. M. Bittner, K. Hahn, C. Scheu, M. Knez, A. Kadri, C. Wege, H. Jeske, K. Kerna, Electrochim. Acta 2006, 51, 6251.

[60] M. Gnerlich, H. Ben-Yoav, J. N. Culver, D. R. Ketchum, R. Ghodssi, J. Power Sources 2015, 293, 649.

[61] C. Yang, J. H. Meldon, B. Lee, H. Yi, Catal. Today 2014, 233, 108.

[62] A. K. Manocchi, S. Seifert, B. Lee, H. Yi, Langmuir 2011, 27, 7052.

[63] D. Li, C. Wang, D. Tripkovic, S. Sun, N. M. Markovic, V. R. Stamenkovic, ACS Catal. 2012, 2, 1358.

[64] C. Koch, A. Poghossian, M. J. Schöning, C. Wege, Nanotheranostics 2018, 2, 184.

[65] H. Zhou, J. Liu, J.-J. Xu, S. Zhang, H. Y. Chen, in Advances in Clinical Chemistry (Ed: G. S. Makowski), Elsevier, Cambridge 2019, pp. 31– 98.

[66] S.-Y. Lee, J. Choi, E. Royston, D. B. Janes, J. N. Culver, M. T. Harris, J. Nanosci. Nanotechnol. 2006, 6, 974. [67] A. Kadri, E. Maiß, N. Amsharov, A. M. Bittner, S. Balci, K. Kern, H. Jeske, C. Wege, Virus Res. 2011, 157, 35.

[68] E. Royston, S.-Y. Lee, J. N. Culver, M. T. Harris, J. Colloid Interface Sci. 2006, 298, 706.

[69] P. Atanasova, R. C. Hoffmann, N. Stitz, S. Sanctis, Z. Burghard, J. Bill, J. J. Schneider, S. Eiben, Bioinspired, Biomimetic Nanobiomater. 2019, 8, 2.

[70] M. Bendahmane, I. Chen, S. Asurmendi, A. A. Bazzini, J. Szecsi, R. N. Beachy, Virology 2007, 366, 107.

[71] A. A. Khan, E. K. Fox, M. Ł. Górzny, E. Nikulina, D. F. Brougham, C. Wege, A. M. Bittner, 2013, 29, 2094.

[72] M. Knez, M. Sumser, A. M. Bittner, C. Wege, H. Jeske, T. P. Martin, K Kern, Adv. Funct. Mater. 2004, 14, 116.

[73] K. Altintoprak, A. Seidenstücker, A. Welle, S. Eiben, P. Atanasova, N. Stitz, A. Plettl, J. Bill, H. Gliemann, H. Jeske, D. Rothenstein, F. Geiger, C. Wege, Beilstein J. Nanotechnol. 2015, 6, 1399.

[74] F. C. Geiger, F. J. Eber, S. Eiben, A. Mueller, H. Jeske, J. P. Spatza, C. Wege, Nanoscale 2013, 5, 3808.

[75] K. Zhou, F. Li, G. Dai, C. Meng, Q. Wang, Biomacromolecules 2013, 14, 2593.

[76] T. L. Schlick, Z. Ding, E. W. Kovacs, M. B. Francis, J. Am. Chem. Soc. 2005, 127, 3718.

[77] E. Strable, D. E. Prasuhn, A. K. Udit, S. Brown, A. J. Link, J. T. Ngo, G. Lander, J. Quispe, C. S. Potter, B. Carragher, D. A. Tirrell, M. G. Finn, Bioconjugate Chem. 2008, 19, 866.

[78] M. Boyce, C. R. Bertozzi, Nat. Methods 2011, 8, 638.

[79] Z. M. Carrico, D. W. Romanini, R. A. Mehl, M. B. Francis, Chem. Commun. 2008, 1205.

[80] D. L. Dunkelmann, J. C. W. Willis, A. T. Beattie, J. W. Chin, Nat. Chem. 2020, 12, 535.

[81] E. D. Carlson, R. Gan, C. E. Hodgman, M. C. Jewett, Biotechnol. Adv. 2012, 30, 1185.

[82] K. G. Patel, J. R. Swartz, Bioconjugate Chem. 2011, 22, 376.

[83] D. G. Gibson, L. Young, R.-Y. Chuang, J. C. Venter, C. A. Hutchison III, H. O. Smith, Nat. Methods 2009, 6, 343. [84] C. Engler, R. Kandzia, S. Marillonnet, PLoS One 2008, 3, e3647.

[85] C. Engler, S. Marillonnet, Methods Mol. Biol. 2013, 1073, 141. [86] M. S. Packer, D. R. Liu, Nat. Rev. Genet. 2015, 16, 379.

[87] R. Chapman, W. R. Bourn, E. Shephard, H. Stutz, N. Douglass, T. Mgwebi, A. Meyers, N. Chin’ombe, A.-L. Williamson, PLoS One 2014, 9, e103314.

[88] Z. Wu, S. B. J. Kan, R. D. Lewis, B. J. Wittmann, R. H. Arnold, Proc. Natl. Acad. Sci. U. S. A. 2019, 116, 8852.

[89] A. Ravikumar, G. A. Arzumanyan, M. K. A. Obadi, A. A. Javanpour, C. C. Liu, Cell 2018, 175, 1946.

[90] Y. Huo, X. Wan, T. Ling, J. Wu, W. Wang, S. Shen, Mol. Immunol. 2018, 93, 278.

[91] S. R. Trevino, J. M. Scholtz, C. N. Pace, J. Pharm. Sci. 2008, 97, 4155.

[92] P. Vormittag, T. Klamp, J. Hubbuch, Front. Bioeng. Biotechnol. 2020, 8, 395.

[93] D. J. Ahn, A. Berman, D. Charych, J. Phys. Chem. 1996, 100, 12455.

[94] A. Shafaei, G. R. Khayati, Measurement 2020, 151, 107199.

[95] P. K. Naik, P. Ranjan, P. Kesari, S. Jain, J. Biophys. Chem. 2011, 02, 112.

[96] I. Kim, K. Kang, M. H. Oh, M. Y. Yang, I. Park, Y. S. Nam, Adv. Funct. Mater. 2017, 27, 1703262.

[97] J. P. Palafox-Hernandez, Z. Tang, Z. E. Hughes, Y. Li, M. T. Swihart, P. N. Prasad, T. R. Walsh, M. R. Knecht, Chem. Mater. 2014, 26, 4960.

[98] Z. Tang, J. P. Palafox-Hernandez, W.-C. Law, Z. E. Hughes, M. T. Swihart, P. N. Prasad, M. R. Knecht, T. R. Walsh, ACS Nano 2013, 7, 9632.

[99] H. Chen, X. Su, K.-G. Neoh, W.-S. Choe, Langmuir 2008, 24, 6852.

[100] M. Hnilova, E. E. Oren, U. O. S. Seker, B. R. Wilson, S. Collino, J. S. Evans, C. Tamerler, M. Sarikaya, Langmuir 2008, 24, 12440.

[101] N. Kantarci, C. Tamerler, M. Sarikaya, T. Haliloglu, P. Doruker, Polymer 2005, 46, 4307.

This paper is published (Hooker et al, Leveraging Anaerobic Fungi for Biotechnology, Current Opinion in Biotechnology, https://doi.org/10.1016/j.copbio.2019.03.013, 2019). The numbering of the figures, tables, and references have been modified to fit the format of this thesis.

Leveraging Anaerobic Fungi for Biotechnology

Casey A Hooker[1,2], Kok Zhi Lee[1], and Kevin V Solomon[1,2,3]

Abstract

Early-branching anaerobic fungi are critical for hydrolyzing untreated lignocellulose in the digestive tracts of large herbivorous animals. While these fungi were discovered more than 40 years ago, they remain understudied and underexploited. Recent advances in -omics technologies, however, have enabled studies that reveal significant biosynthetic potential within anaerobic fungal genomes for diverse biotechnological applications. Applications range from enhanced second-generation bioenergy platforms to improved animal health. However, developing gut fungi for these applications will require significant advances in genome engineering technologies for these organisms. Here, we review the biotechnological abilities of anaerobic fungi and highlight challenges that must be addressed to develop them for a range of biotechnological applications.

Introduction

Anaerobic fungi (phylum Neocallimastigomycota) are central to hydrolyzing crude, ingested plant material in the digestive tracts of ruminant and hindgut fermenting animals [1,2,3**]. Initially identified as zooflagellates at the turn of the 20th century, a series of papers by Orpin in the mid-70s describing Neocallimastix frontalis, Caecomyces communis (originally classified as Sphaeromonas communis) and Piromyces communis established the foundation for their classification as a distinct class of lignocellulolytic fungi [2,3**,4–6]. While best known for colonizing the gastrointestinal tracts of herbivorous animals, anaerobic fungi have since been found in diverse niches worldwide including landfill soils, coastal marine sediments, and the deep biosphere (i.e. several kilometers below the surface of the earth in igneous bedrock) [7,8]. Anaerobic fungi are critical for carbon cycling as they decompose an array of lignocellulose-rich biomass streams. Thus, these organisms and their lignocellulolytic enzymes are vital to agricultural, nutritional, and possibly biogeochemical processes [7,9,10*].

As anaerobic fungi were first identified in herbivore digestion, initial research focused on elucidating their role in breaking down consumed forages. Anaerobic fungi are the primary colonizers of ingested plant material, and increase particle surface area and access for other lignocellulolytic microbes leading to more complete and rapid digestion of animal nutrients [11,12]. Despite accounting for very small percentages of the gut microflora in their hosts (7–9%), anaerobic fungi release more than 50% of the fermentable sugars from ingested plant matter [2,9]. Nonetheless, fungi in this phylum are understudied due to a lack of complete genomes and genetic tools to manipulate them. Here, we review the past and current development of anaerobic fungi, and highlight emerging applications of anaerobic gut fungi as microbial cell factories. We also discuss the key engineering tools that will need to be developed to enable these organisms to address a range of global grand challenges.

Original anaerobic fungal applications in animal health and nutrition

Gut microbes are essential to plant biomass digestion in large herbivores as the host animals lack the enzymes needed to break down crude fiber rich materials. This recalcitrant lignocellulose is degraded in part by anaerobic fungi that grow invasively into the material and secrete the largest and most diverse repertoire of fungal lignocellulolytic enzymes [16]. The enzymes of anaerobic fungi hydrolyze an array of feedstocks and release theirsugars with high efficiency [17]. Anaerobic fungi are thought to be the primary colonizers of the ingested plant matter, and are thus critical for efficient digestion in their hosts [16]. This was demonstrated in studies that showed that removal of anaerobic fungi from the digestive tracts of sheep decreased feed intake by 40% [18]. Similar experiments supplementing buffalo feed with anaerobic fungal cultures increased digestion of wheat straw and green fodder leading to improved weight gain [19]. Therefore, anaerobic fungi enhance digestion of recalcitrant forages, and may be introduced in the animal diet as a probiotic to promote animal growth. Developing anaerobic fungi as a probiotic or dietary supplement has significant room to positively benefit agricultural industries. For example, it has been demonstrated that diet type contributes significantly to which populations of anaerobic fungi are most prominent in the digestive tracts of their respective hosts [10* ,20]. Thus, in the future, one might envision a range of livestock probiotics based on cultures of anaerobic gut fungi that are refined based upon dietary availability. This becomes particularly attractive if changes in climate or economic conditions induce significant alterations in feedstock availability for the host animals. Moreover, analysis of recently sequenced fungal genomes demonstrate that anaerobic fungi may synthesize natural products; natural products form the basis of current animal antibiotics and growth promoters that could be produced in situ by anaerobic fungal probiotics to reduce the current demand for antibiotics in the cattle industry [14,21–23].

Anaerobic fungal centered communities and consortia

The digestive tracts of ruminant and hindgut fermenting animals harbor a plethora of bacteria, archaea, and anaerobic fungi that work together to break down plant biomass [24]. Yet the complex interactions occurring between these microorganisms is only partially understood. For example, methanogenic archaea metabolize a variety of fermentation products including fungal produced hydrogen mutually benefitting both organisms [2,12,25]. In so doing they enhance fungal enzyme activity in coculture by at least 30–60 % depending on the strain of anaerobic fungi [26]. Similarly, monocultures of Neocallimastix frontalis solubilized 16% of crystalline cellulose in 72 hours, whereas cocultures with Methanobrevibacter smithii solubilized 98% of crystalline cellulose in the same time [27]. While the complex interactions between these two kingdoms are extremely beneficial for biomass hydrolysis, few attempts have leveraged this knowledge for biotechnological applications.

Engineering or modifying the interactions between gut fungi and archaea could be exploited for many technologies. For example, methanogenesis in livestock animals is a significant contributor to global methane production, and consequently, greenhouse gas emissions [28]. By building a genome engineering toolkit for anaerobic fungi, it is possible that the fungal metabolome could be shifted to hinder methane production in these animals, or to enhance biogas production in anaerobic digesters. Similarly, metabolic engineering of anaerobic fungi could be leveraged to enhance the mutualism between archaea and fungi for processes such as biogas generation. Anaerobic fungi have been added to biogas production processes increasing biogas production by (+4–22%) depending on the substrate composition and the strain of anaerobic fungi [29], which may improve the economic viability of livestock operations, and reduce greenhouse gas emissions

[30]. As biogas production can be significantly influenced by feedstock composition, developing microbial consortia robust to compositional changes will be critical to maintain production stability and efficiency [31]. Alternatively, metabolic engineering may be used to enhance biological hydrogen production by anaerobic fungi. However, developing anaerobic fungi for biological hydrogen production or for biogas platforms will require significant advances in genome, metabolic, and organelle engineering. Synthetic cocultures of anaerobic fungi with bacteria, yeasts, or methanogens could be leveraged as a strategy to improve biomanufacturing efficiency, which has been successful for model organisms [32]. As anaerobic fungi release significant portions of the fermentable sugars in plant biomass and produce useful metabolic byproducts, an opportunity exists to convert inexpensive feedstocks (i.e. lignocellulose) into high value products. Current approaches use two stage operations, where anaerobic fungi are initially used to break down and ferment plant material. Subsequently, more genetically tractable organisms (e.g. Saccharomyces cerevisiae or Escherichia coli) are added to the fermentation liquor to make a higher value product (Figure 1) [33,34]. While these reports may not be true cocultures, these demonstrations serve as exciting proofs of concept to exploit anaerobic fungi in synthetic communities to produce value-added products directly from inexpensive lignocellulose without pretreatment.

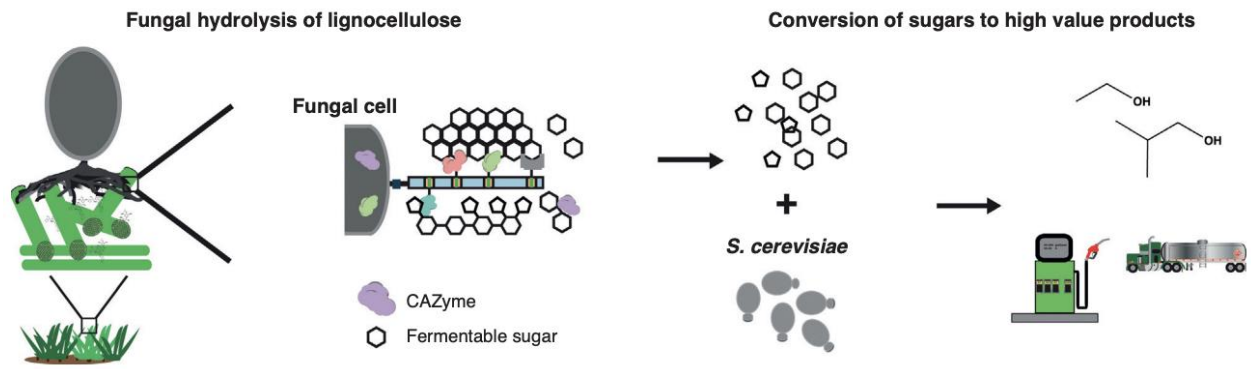
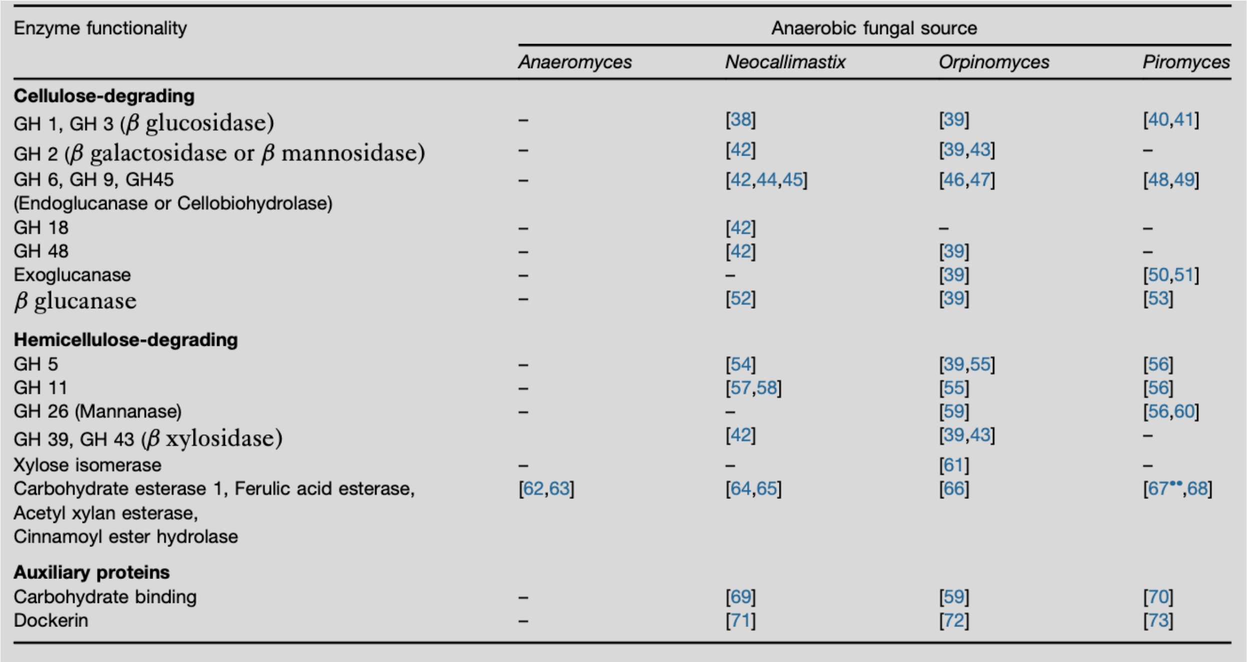

Fungal hydrolysis of lignocellulose
Conversion of sugars to high value products
Fungal cell
CAZyme
Fermentable sugar
+
S. cerevisiae
OH
OH

Figure 1. Anaerobic efficiently degrade crude lignocellulose via a cellulosome. Anaerobic fungi grow invasively into crude lignocellulose and secrete CAZymes that release the trapped sugars. Anaerobic fungi are currently the only fungal species known to form extracellular cellulosomes where CAZymes self-assemble on a scaffold to more efficiently depolymerize lignocellulose. These sugar monomers can then be fermented by more genetically tractable organisms to make high value products such as ethanol or isobutanol with applications as biofuels.

Heterologous cloning of anaerobic fungal enzymes

Lignocellulose is widely regarded as a potential feedstock for bioenergy production as it does not compete with food crops, and has enormous annual production yields estimated to be greater than 150 billion tons [35]. Yet developing robust enzyme technologies to readily release the sugars in plant biomass has been a significant bottleneck hindering second generation energy platforms from being sustainable [15]. Most industrial enzyme preparations rely on fungi such as Trichoderma reesei or Aspergillus spp. that encode significantly fewer carbohydrate active enzymes (CAZymes), which release sugars from biomass, than anaerobic fungi [13]. Thus, anaerobic fungi may be a superior platform for the production of more active enzyme cocktails. Many anaerobic fungal CAZymes (e.g. bglucosidases and hemicellulases) were acquired through horizontal gene transfer, which integrates both fungal and bacterial hydrolytic strategies [36,37**]. Further, gut fungi not only secrete free CAZymes, but also form cellulosomes that allow for concerted enzyme hydrolysis of lignocellulosic material (Figure 1) [37**]. Thus, lower enzyme concentrations are required for efficient hydrolysis, which could be extremely useful to decrease production costs [15]. While the repertoire of CAZymes expressed by anaerobic fungi have long been attractive for biomass hydrolysis, most early efforts focused on characterizing enzymes expressed by these organisms. However, given the naturally low enzyme expression in anaerobic fungi, a large body of work focused on cloning specific enzymes into more genetically tractable organisms (Table 1). While these approaches have had some success, the complex hydrolysis of crude plant material requires a cocktail of several enzymes, which may be more readily produced and optimized from its native host, anaerobic fungi.

Table 1. Heterologously expressed anaerobic fungal enzymes from various genera.



Enzyme functionality	Anaerobic fungal source			
	Anaeromyces	*Neocallimastix*	*Orpinomyces*	*Piromyces*
Cellulose-degrading				
GH 1, GH 3 (β glucosidase)	–	[38]	[39]	[40,41]
GH 2 (β galactosidase or β mannosidase)	–	[42]	[39,43]	–
GH 6, GH 9, GH45 (Endoglucanase or Cellobiohydrolase)	–	[42,44,45]	[46,47]	[48,49]
GH 18	–	[42]	–	–
GH 48	–	[42]	[39]	–
Exoglucanase	–	–	[39]	[50,51]
β glucanase	–	[52]	[39]	[53]
Hemicellulose-degrading				
GH 5	–	[54]	[39,55]	[56]
GH 11	–	[57,58]	[55]	[56]
GH 26 (Mannanase)	–	–	[59]	[56,60]
GH 39, GH 43 (β xylosidase)		[42]	[39,43]	–
Xylose isomerase	–	–	[61]	–
Carbohydrate esterase 1, Ferulic acid esterase, Acetyl xylan esterase, Cinnamoyl ester hydrolase	[62,63]	[64,65]	[66]	[67**,68]
Auxiliary proteins				
Carbohydrate binding	–	[69]	[59]	[70]
Dockerin	–	[71]	[72]	[73]

Omics studies of anaerobic fungi reveals extensive biosynthetic potential

Although anaerobic fungi were discovered more than 40 years ago, only recent innovations in gut fungal techniques have made it feasible to disrupt the chitinous cell wall and isolate quality DNA for genome acquisition [36,74]. Similarly, advances in next generation sequencing have overcome the extremely high AT bias (~80%) and homopolymer repeats in anaerobic fungal genomes to complete the first few draft genomes [36,37**]. In 2013, the first draft genome for a species of anaerobic fungi was published [36]. In the subsequent four years five more draft genomes and multiple transcriptomes have become available representing 5 of the 11 total genera (Anaeromyces, Caecomyces, Neocallimastix, Orpinomyces, and Piromyces) in the phylum Neocallimastigomycota [36,37**,75* ,76**]. These works have begun to elucidate how anaerobic fungi regulate gene expression and break down plant biomass. Initially, most -omics analyses focused on the regulation and number of biomass degrading enzymes within the genomes of anaerobic fungi. Through RNAseq, Solomon et al. identified the large numbers of CAZymes expressed by anaerobic fungi, and the corresponding regulatory changes associated with unique substrates [76**]. Accordingly, subsequent work focused on the mechanisms by which these enzymes break down plant biomass, including discovery of fungal cellulosomes (i.e. protein superstructures allowing for concerted hydrolysis and reduced enzyme loadings) in anaerobic fungi, which were hypothesized in as early as 1992 [77]. Recently, Haitjema et al. demonstrated that anaerobic fungi form cellulosomes by the interactions of CAZyme dockerin domains with corresponding cohesin domains on scaffoldin proteins that are tethered to the exterior of the fungal cell (Figure 1) [37**]. Most notably, the cellulosomes of these organisms have compositional promiscuity that may help to confer a selective advantage over the bacterial counterparts in the digestive tracts of their animal hosts [37**]. As these recent omics-based analyses helped to elucidate the expression and organization of the CAZymes in gut fungi, ensuing work has begun to analyze the complex regulation of these enzymes, and provides potential targets by which to modify and engineer gene expression. The development of anaerobic fungi as a base platform for lignocellulose degrading enzymes requires a deep knowledge of gene expression regulation and its relationship with fungal metabolism. Carbon catabolite repression (CCR), whereby the end products of substrate hydrolysis limit the expression of CAZymes is present in many organisms as a strategy to conserve energy [78]. CCR significantly affects the expression of biomass degrading enzymes in anaerobic fungi [76**]. Recently, it was shown that natural antisense transcripts are one of the key mechanisms by which the CCR is mediated, and thus provides one target by which CCR may be manipulated for genome engineering technologies [79*]. Similarly, the eukaryotic nature of anaerobic fungisuggestssome level of epigenetic regulation may be present. One example is DNA base methylation (e.g. cytosine or adenine methylation). A recent work has confirmed the presence of adenine methylation in anaerobic fungi near coding areas, and especially near transcriptional start sites to as far as +1500 bases downstream [80**]. Taken together, the rise in omics technologies has significantly improved knowledge of the extensive array of biomass degrading enzymes and their regulation in anaerobic fungi.

Emerging techniques to engineer anaerobic fungi

Attempts to engineer gut fungi for biotechnology have been limited by a lack of a robust genetics toolset. Initial attempts to develop this toolset have focused on the enolase promoter of the genus Neocallimastix [81*]. This promoter was successful in transiently expressing a reporter gene although the complex biolistic nature of this transformation method, and the absence of a plasmid autonomous replicating sequence for expression stability make this method challenging for routine use [82*]. Recently, Calkins et al. demonstrated that anaerobic fungi are naturally competent for nucleic acids, which has greatly sped the ability to introduce foreign DNA or RNA into the nucleus of gut fungi [83*]. The authors used this transformation methodology to knock down expression of a lactose dehydrogenase with RNA transcripts, which has opened the door for RNAi-based transcriptomic engineering. While significant strides have been made to engineer anaerobic fungi, there is still a critical need to expand the genetic toolbox of anaerobic fungi to include more robust permanent transformation methodologies, and identify more promoters as well as other standard biological parts for gene expression. Strain engineering efforts in more highly studied fungi that introduce permanent genetic changes frequently rely on homology-based recombination where transformed linear DNA is integrated into the host chromosome [84–87]. Homologous recombination typically requires significant amounts of homology in each arm (>350 nucleotides) to avoid non-homologous end joining pathways, which lead to random integration, pleotropic effects, or potentially lethal insertions [88]. More accurate site-specific recombination events are favored when CRISPR-based technologies are used to create targeted double stranded DNA breaks, which must be repaired for cell viability. These techniques have been successfully used in other fungi for genome editing and may be equally successful in anaerobic fungi [84,86,87,89]. However, to leverage these techniques, stable plasmid or artificial chromosome maintenance will be necessary. Thus, identifying autonomous replication sequences, centromeres, and selectable markers such as antibiotic resistant genes will be critical. Lastly, as anaerobic fungi are eukaryotic organisms, tools for cellular localization (i.e. nuclear localization signals) and reporters including anaerobic fluorescent proteins will be necessary. We anticipate robust transformation methodologies, stable artificial chromosomes or plasmids, and CRISPR technologies (Figure 2) will enable high throughput functional genomics methods akin to CREATE or TRMR to screen and fully annotate anaerobic fungal genomes [90,91].

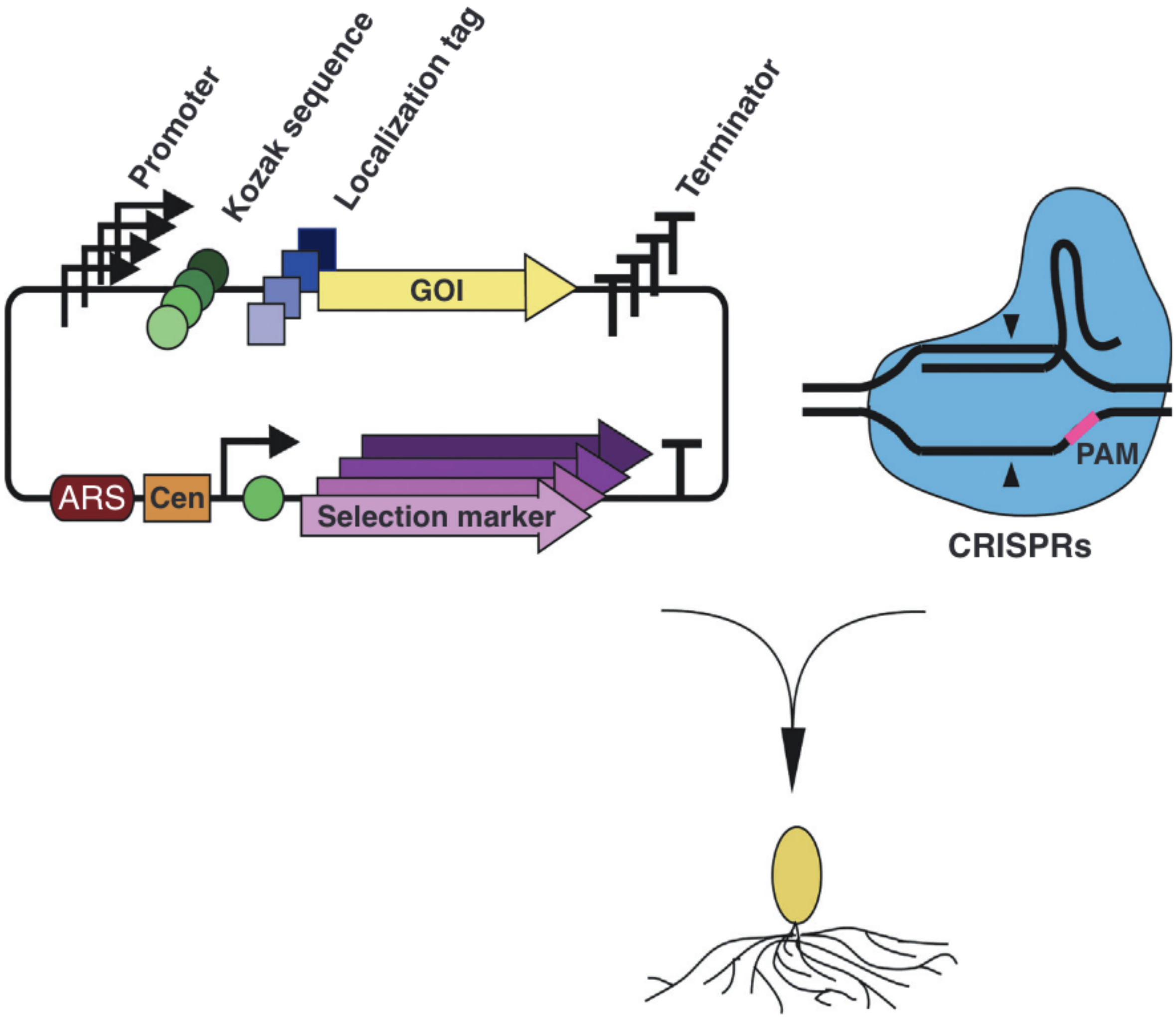


Figure 2. Genome editing technologies potentially enable anaerobic fungal engineering. CRISPR based engineering and omics technologies will allow forrapid production, screening, and characterization of mutant libraries of anaerobic fungi for biotechnology purposes.www.sciencedirect.com Current Opinion in Biotechnology 2019, 59:103–110

Perspective and outlook

Since the discovery of anaerobic fungi more than 40 years ago, anaerobic fungi have been pursued for their ability to readily break down crude plant biomass for animal nutrition and bioenergy applications. These microorganisms provide a wealth of biosynthetic potential that may be leveraged for a range of technologies; however, engineering anaerobic fungi remains as a key bottleneck. The significant advances in synthetic biology and the rise in sequencing data will allow for many of the challenges to be overcome. The extent to which these organisms may be exploited for biotechnology is only beginning to be realized, thus highlighting the critical need for future studies. These biotechnological applications range from improving animal health, lignocellulolytic enzyme production, natural product identification, and biogas technologies among many more. In summary, anaerobic fungi provide an exciting and untapped resource for a large and diverse range of biotechnologies.

Conflict of interest statement

Nothing declared.

Acknowledgements

This research was supported by the startup funds from the Colleges of Engineering and Agriculture, and the USDA National Institute of Food and Agriculture (Hatch Multistate Project S1075). CH is supported by a National Science Foundation graduate research fellowship (DGE-1333468). The funding sources had no role in the study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

** of outstanding interest

1. Nicholson MJ, Theodorou MK, Brookman JL: Molecular analysis of the anaerobic rumen fungus Orpinomyces – insights into an AT-rich genome. Microbiology 2005, 151:121-133.

2. Orpin CG: The rumen flagellate Piromonas communis: its lifehistory and invasion of plant material in the rumen. J Gen Microbiol 1977, 99:107-117.

3. **Orpin CG: Studies on the rumen flagellate Neocallimastix frontalis. J Gen Microbiol 1975, 91:249-262. This seminal article by Orpin is the first description of anaerobic fungi and their fibrolytic properties, paving the way for all future research in anaerobic fungi.

4. Gold JJ, Brent Heath I, Bauchop T: Ultrastructural description of a new chytrid genus of caecum anaerobe, Caecomyces equi gen. nov., sp. nov., assigned to the Neocallimasticaceae. Biosystems 1988, 21:403-415.

5. Orpin CG: Invasion of plant tissue in the rumen by the flagellate Neocallimastix frontalis. Microbiology 1977, 98:423-430.

6. Orpin C: The occurrence of chitin in the cell walls of the rumen organisms Neocallimastix frontalis, Piromonas communis and Sphaeromonas communis. Microbiology 1977, 99:215-218.

7. Drake H, Ivarsson M: The role of anaerobic fungi in fundamental biogeochemical cycles in the deep biosphere. Fungal Biol Revi 2018, 32:20-25.

8. Picard KT: Coastal marine habitats harbor novel earlydiverging fungal diversity. Fungal Ecol 2017, 25:1-13.

9. Theodorou MK, Mennim G, Davies DR, Zhu W-Y, Trinci APJ, Brookman JL: Anaerobic fungi in the digestive tract of mammalian herbivores and their potential for exploitation. Proc Nutr Soc 1996, 55:913-926.

10. Boots B, Lillis L, Clipson N, Petrie K, Kenny DA, Boland TM, Doyle E: Responses of anaerobic rumen fungal diversity (phylum Neocallimastigomycota) to changes in bovine diet. J Appl Microbiol 2013, 114:626-635. This study shows that alterations in dietary fiber content and soya oil significantly affect fungal diversity in the rumen. Thus, this work provides precedence for ways to mitigate green house gas emissions in livestock animals.

11. Orpin CG, Joblin KN: The rumen anaerobic fungi. In The Rumen Microbial Ecosystem. Edited by Hobson PN, Stewart CS. Netherlands: Springer; 1997:140-195.

12. Bauchop T, Mountfort DO: Cellulose fermentation by a rumen anaerobic fungus in both the absence and the presence of rumen methanogens. Appl Environ Microbiol 1981, 42:1103-1110.

13. Seppa¨ la¨ S, Wilken SE, Knop D, Solomon KV, O'Malley MA: The importance of sourcing enzymes from non-conventional fungi for metabolic engineering and biomass breakdown. Metab Eng 2017, 44:45-59.

14. Hillman ET, Readnour LR, Solomon KV: Exploiting the natural product potential of fungi with integrated -omics and synthetic biology approaches. Curr Opin Syst Biol 2017, 5:50-56.

15. Klein-Marcuschamer D, Oleskowicz-Popiel P, Simmons BA, Blanch HW: The challenge of enzyme cost in the production of lignocellulosic biofuels. Biotechnol Bioeng 2012, 109:1083-1087.

16. Gruninger RJ, Puniya AK, Callaghan TM, Edwards JE, Youssef N, Dagar SS, Fliegerova K, Griffith GW, Forster R, Tsang A et al.: Anaerobic fungi (phylum Neocallimastigomycota): advances in understanding their taxonomy, life cycle, ecology, role and biotechnological potential. FEMS Microbiol Ecol 2014, 90:1-17.

17. Hooker CA, Hillman ET, Overton JC, Ortiz-Velez A, Schacht M, Hunnicutt A, Mosier NS, Solomon KV: Hydrolysis of untreated lignocellulosic feedstock is independent of S-lignin composition in newly classified anaerobic fungal isolate, Piromyces sp. UH3-1. Biotechnol Biofuels 2018, 11:293.

18. Gordon GLR, Phillips MW: Removal of anaerobic fungi from the rumen of sheep by chemical treatment and the effect on feed consumption and in vivo fibre digestion. Lett Appl Microbiol 1993, 17:220-223.

19. Paul SS, Deb SM, Punia BS, Das KS, Singh G, Ashar MN, Kumar R: Effect of feeding isolates of anaerobic fungus Neocallimastix CF 17 on growth rate and fibre digestion in buffalo calves. Arch Anim Nutr 2011, 65:215-228.

20. Khejornsart P, Wanapat M: Diversity of rumen anaerobic fungi and methanogenic archaea in swamp buffalo influenced by various diets. J Anim Vet Adv 2010, 9:3062-3069.

21. Russell JB, Rychlik JL: Factors that alter rumen microbial ecology. Science 2001, 292:1119-1122.

22. Nagaraja T, Chengappa M: Liver abscesses in feedlot cattle: a review. J Anim Sci 1998, 76:287-298.

23. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F et al.: MycoCosm portal: gearing up for 1000 fungal genomes. Nucl Acids Res 2014, 42: D699-D704.

24. Lin KW, Patterson JA, Ladisch MR: Anaerobic fermentation: microbes from ruminants. Enzyme Microb Technol 1985, 7:98-107.

25. Yarlett N, Orpin CG, Munn EA, Yarlett NC, Greenwood CA: Hydrogenosomes in the rumen fungus Neocallimastix patriciarum. Biochem J 1986, 236:729-739.

26. Teunissen MJ, Kets EPW, Op den Camp HJM, Huis in't Veld JHJ, Vogels GD: Effect of coculture of anaerobic fungi isolated from ruminants and non-ruminants with methanogenic bacteria on cellulolytic and xylanolytic enzyme activities. Arch Microbiol 1992, 157:176-182.

27. Wood TM, Wilson CA, McCrae SI, Joblin KN: A highly active extracellular cellulase from the anaerobic rumen fungus Neocallimastix frontalis. FEMS Microbiol Lett 1986, 34:37-40.

28. Caro D, Kebreab E, Mitloehner FM: Mitigation of enteric methane emissions from global livestock systems through nutrition strategies. Clim Change 2016, 137:467-480.

29. Procha´ zka J, Mra´ zek J, trosova´ L, Fliegerova´ K, Za´ branska´ J, Doha´ nyos M: Enhanced biogas yield from energy crops with rumen anaerobic fungi. Eng Life Sci 2012, 12:343-351.

30. Masse´ DI, Talbot G, Gilbert Y: On farm biogas production: a method to reduce GHG emissions and develop more sustainable livestock operations. Anim Feed Sci Technol 2011, 166–167:436-445.

31. Amon T, Amon B, Kryvoruchko V, Zollitsch W, Mayer K, Gruber L: Biogas production from maize and dairy cattle manure— influence of biomass composition on the methane yield. Agric Ecosyst Environ 2007, 118:173-182.

32. Jones JA, Wang X: Use of bacterial co-cultures for the efficient production of chemicals. Curr Opin Biotechnol 2018, 53:33-38.

33. Henske JK, Wilken SE, Solomon KV, Smallwood CR, Shutthanandan V, Evans JE, Theodorou MK, O'Malley MA: Metabolic characterization of anaerobic fungi provides a path forward for bioprocessing of crude lignocellulose. Biotechnol Bioeng 2018, 115:874-884. 3

4. Ranganathan A, Smith OP, Youssef NH, Struchtemeyer CG, Atiyeh HK, Elshahed MS: Utilizing anaerobic fungi for two-stage sugar extraction and biofuel production from lignocellulosic biomass. Front Microbiol 2017, 8:635.

35. Pauly M, Keegstra K: Cell wall carbohydrates and their modification as a resource for biofuels. Plant J 2008, 54:559-568.

36. Youssef NH, Couger MB, Struchtemeyer CG, Liggenstoffer AS, Prade RA, Najar FZ, Atiyeh HK, Wilkins MR, Elshahed MS: The genome of the anaerobic fungus Orpinomyces sp. strain C1A reveals the unique evolutionary history of a remarkable plant biomass degrader. Appl Environ Microbiol 2013, 79:4620-4634.

37. ** Haitjema CH, Gilmore SP, Henske JK, Solomon KV, de Groot R, Kuo A, Mondo SJ, Salamov AA, LaButti K, Zhao Z et al.: A parts list for fungal cellulosomes revealed by comparative genomics. Nat Microbiol 2017, 2:17087. This recent work elucidates the components and assembly mechanisms for cellulosomes in anaerobic fungi. Notably, through their analyses Haitjema et al. show that these cellulsomes are evolutionary chimeric structures and have compositional promiscuity that appears to provide a competitive advantage to anaerobic fungi to outcompete their bacterial counterparts.

38. Chen H-L, Chen Y-C, Lu M-Y, Chang J-J, Wang H-T, Ke H-M, Wang T-Y, Ruan S-K, Wang T-Y, Hung K-Y et al.: A highly efficient beta -glucosidase from the buffalo rumen fungus Neocallimastix patriciarum W5. Biotechnol Biofuels 2012, 5 24-24.

39. Morrison JM, Elshahed MS, Youssef NH: Defined enzyme cocktail from the anaerobic fungus Orpinomyces sp. strain C1A effectively releases sugars from pretreated corn stover and switchgrass. Sci Rep 2016, 6:29217.

40. Harhangi HR, Akhmanova AS, Emmens R, Cvander Drift, de Laat WTAM, van Dijken JP, Jetten MSM, Pronk JT, Op den Camp HJM: Xylose metabolism in the anaerobic fungus Piromyces sp. strain E2 follows the bacterial pathway. Arch Microbiol 2003, 180:134-141.

41. Steenbakkers PJM, Harhangi HR, Bosscher MW, vd Hooft MMC, Keltjens JT, vd Drift C, Vogels GD, Camp HJMOD: betaGlucosidase in cellulosome of the anaerobic fungus Piromyces sp. strain E2 is a family 3 glycoside hydrolase. Biochem J 2003, 370:963-970.

42. Wang T-Y, Chen H-L, Lu M-Y, Chen Y-C, Sung H-M, Mao C-T, Cho H-Y, Ke H-M, Hwa T-Y, Ruan S-K et al.: Functional characterization of cellulases identified from the cow rumen fungus Neocallimastix patriciarum W5 by transcriptomic and secretomic analyses. Biotechnol Biofuels 2011, 4:24.

43. Morrison JM, Elshahed MS, Youssef N: A multifunctional GH39 glycoside hydrolase from the anaerobic gut fungus Orpinomyces sp. strain C1A. PeerJ 2016, 4:e2289.

44. Mingardon F, Chanal A, Lo´ pez-Contreras AM, Dray C, Bayer EA, Fierobe H-P: Incorporation of fungal cellulases in bacterial minicellulosomes yields viable, synergistically acting cellulolytic complexes. Appl Environ Microbiol 2007, 73:3822-3832.

45. Lo´ pez-Contreras AM, Smidt H, van der Oost J, Claassen PA, Mooibroek H, de Vos WM: Clostridium beijerinckii cells expressing Neocallimastix patriciarum glycoside hydrolases show enhanced lichenan utilization and solvent production. Appl Environ Microbiol 2001, 67:5127-5133.

46. Jin X, Meng N, Xia L-m: Expression of an endo-b-1,4-glucanase gene from Orpinomyces PC-2 in Pichia pastoris. Int J Mol Sci 2011, 12:3366-3380.

47. Hughes SR, Riedmuller SB, Mertens JA, Li X-L, Bischoff KM, Qureshi N, Cotta MA, Farrelly PJ: High-throughput screening of cellulase F mutants from multiplexed plasmid sets using an automated plate assay on a functional proteomic robotic workcell. Proteome Sci 2006, 4:10.

48. Tsai C-F, Qiu X, J-H Liu: A comparative analysis of two cDNA clones of the cellulase gene family from anaerobic fungus Piromyces rhizinflata. Anaerobe 2003, 9:131-140.

49. Eberhardt RY, Gilbert HJ, Hazlewood GP: Primary sequence and enzymic properties of two modular endoglucanases, Cel5A and Cel45A, from the anaerobic fungus Piromyces equi. Microbiology 2000, 146:1999-2008.

50. Liu J-H, Tsai C-F, Liu J-W, Cheng K-J, Cheng C-L: The catalytic domain of a Piromyces rhizinflata cellulase expressed in Escherichia coli was stabilized by the linker peptide of the enzyme. Enzyme Microb Technol 2001, 28:582-589.

51. O’Malley MA, Theodorou MK, Kaiser CA: Evaluating expression and catalytic activity of anaerobic fungal fibrolytic enzymes native to Piromyces sp. E2 in Saccharomyces cerevisiae. Environ Prog Sustain Energy 2011, 31:37-46.

52. Hung Y-L, Chen H-J, Liu J-C, Chen Y-C: Catalytic efficiency diversification of duplicate b-1, 3-1, 4-glucanases from Neocallimastix patriciarum J11. Appl Environ Microbiol 2012, 78:4294-4300 AEM. 07473-07411.

53. Chu C-Y, Tseng C-W, Yueh P-Y, Duan C-H, Liu J-R: Molecular cloning and characterization of a b-glucanase from Piromyces rhizinflatus. J Biosci Bioeng 2011, 111:541-546.

54. Lee JMT, Hu Y, Zhu H, Cheng KJ, Krell PJ, Forsberg CW: Cloning of a xylanase gene from the ruminal fungus Neocallimastix patriciarum 27 and its expression in Escherichia coli. Can J Microbiol 1993, 39:134-139.

55. Liab K, Azadi P, Collins R, Tolan J, Kim JS, Eriksson KEL: Relationships between activities of xylanases and xylan structures. Enzyme Microb Technol 2000, 27:89-94.

56. Fanutti C, Ponyi T, Black GW, Hazlewood GP, Gilbert HJ: The conserved noncatalytic 40-residue sequence in cellulases and hemicellulases from anaerobic fungi functions as a protein docking domain. J Biol Chem 1995, 270:29314-29322.

57. Liu J-R, Duan C-H, Zhao X, Tzen JT, Cheng K-J, Pai C-K: Cloning of a rumen fungal xylanase gene and purification of the recombinant enzyme via artificial oil bodies. Appl Microbiol Biotechnol 2008, 79:225-233.

58. Xue H, Zhou J, You C, Huang Q, Lu H: Amino acid substitutions in the N-terminus, cord and a-helix domains improved the thermostability of a family 11 xylanase XynR8. J Ind Microbiol Biotechnol 2012, 39:1279-1288.

59. Ximenes EA, Chen H, Kataeva IA, Cotta MA, Felix CR, Ljungdahl LG, Li X-L: A mannanase, ManA, of the polycentric anaerobic fungus Orpinomyces sp. strain PC-2 has carbohydrate binding and docking modules. Can J Microbiol 2005, 51:559-568.

60. Millward-Sadler SJ, Hall J, Black GW, Hazlewood GP, Gilbert HJ: Evidence that the Piromycesgene family encoding endo-l, 4- mannanases arose through gene duplication. FEMS Microbiol Lett 1996, 141:183-188.

61. Madhavan A, Tamalampudi S, Ushida K, Kanai D, Katahira S, Srivastava A, Fukuda H, Bisaria VS, Kondo A: Xylose isomerase from polycentric fungus Orpinomyces: gene sequencing, cloning, and expression in Saccharomyces cerevisiae for bioconversion of xylose to ethanol. Appl Microbiol Biotechnol 2008, 82:1067.

62. Qi M, Wang P, Selinger LB, Yanke LJ, Forster RJ, McAllister TA: Isolation and characterization of a ferulic acid esterase (Fae1A) from the rumen fungus Anaeromyces mucronatus. J Appl Microbiol 2011, 110:1341-1350.

63. Gruninger RJ, Cote C, McAllister TA, Abbott DW: Contributions of a unique b-clamp to substrate recognition illuminates the molecular basis of exolysis in ferulic acid esterases. Biochem J 2016, 473:839-849.

64. Dalrymple BP, Cybinski DH, Layton I, McSweeney CS, Xue G-P, Swadling YJ, Lowry JB: Three Neocallimastix patriciarum esterases associated with the degradation of complex polysaccharides are members of a new family of hydrolases. Microbiology 1997, 143:2605-2614.

65. Pai C-K, Wu Z-Y, Chen M-J, Zeng Y-F, Chen J-W, Duan C-H, Li ML, Liu J-R: Molecular cloning and characterization of a bifunctional xylanolytic enzyme from Neocallimastix patriciarum. Appl Microbiol Biotechnol 2010, 85:1451-1462.

66. Blum DL, Li X-L, Chen H, Ljungdahl LG: Characterization of an acetyl xylan esterase from the anaerobic fungus Orpinomyces sp. strain PC-2. Appl Environ Microbiol 1999, 65:3990-3995.

67. ** Poidevin L, Levasseur A, Pae¨ s G, Navarro D, Heiss-Blanquet S, Asther M, Record E: Heterologous production of the Piromyces equi cinnamoyl esterase in Trichoderma reesei for biotechnological applications. Lett Appl Microbiol 2009, 49:673-678. This work investigated enzyme activity of a cinnamoyl esterase from Piromyces equi in hydrolyzing various feedstocks. Notably, the activity of the P. equi esterase was higher than the aerobic enzymes from other organisms which were investigated. This paper provides exciting proof for sourcing biomass degrading enzymes from anaerobic fungi as a way to boost enzyme cocktail performance on various feedstocks.

68. Fillingham IJ, Kroon PA, Williamson G, Gilbert HJ, Hazlewood GP: A modular cinnamoyl ester hydrolase from the anaerobic fungus Piromyces equi acts synergistically with xylanase and is part of a multiprotein cellulose-binding cellulase– hemicellulase complex. Biochem J 1999, 343:215-224.

69. Gustavsson M, Lehtio¨ J, Denman S, Teeri TT, Hult K, Martinelle M: Stable linker peptides for a cellulose-binding domain–lipase fusion protein expressed in Pichia pastoris. Protein Eng 2001, 14:711-715.

70. Charnock SJ, Bolam DN, Nurizzo D, Szabo´ L, McKie VA, Gilbert HJ, Davies GJ: Promiscuity in ligand-binding: the threedimensional structure of a Piromyces carbohydrate-binding module, CBM29-2, in complex with cello- and mannohexaose. Proc Natl Acad Sci U S A 2002, 99:14077-14082.

71. Wang H-C, Chen Y-C, Hseu R-S: Purification and characterization of a cellulolytic multienzyme complex produced by Neocallimastix patriciarum J11. Biochem Biophys Res Commun 2014, 451:190-195.

72. Steenbakkers PJ, Li X-L, Ximenes EA, Arts JG, Chen H, Ljungdahl LG, den Camp HJO: Noncatalytic docking domains of cellulosomes of anaerobic fungi. J Bacteriol 2001, 183:5325-5333.

73. Raghothama S, Eberhardt RY, Simpson P, Wigelsworth D, White P, Hazlewood GP, Nagy T, Gilbert HJ, Williamson MP: Characterization of a cellulosome dockerin domain from the anaerobic fungus Piromyces equi. Nat Struct Mol Biol 2001, 8:775-778.

74. Solomon KV, Henske JK, Theodorou MK, Amp, Apos, Malley MA: Robust and effective methodologies for cryopreservation and DNA extraction from anaerobic gut fungi. Anaerobe 2016, 38:39-46.

75. Henske JK, Gilmore SP, Knop D, Cunningham FJ, Sexton JA, Smallwood CR, Shutthanandan V, Evans JE, Theodorou MK, O'Malley MA: Transcriptomic characterization of Caecomyces churrovis: a novel, non-rhizoid-forming lignocellulolytic anaerobic fungus. Biotechnol Biofuels 2017, 10:305. This paper characterizes the ability of anaerobic fungi to degrade lignocellulosic feedstocks and provides an in-depth analysis of sugar hydrolysis and fungal enzyme expression. These authors also develop a pipeline to use anaerobic fungi in two-stage cocultures which are exciting proofs of concept for developing gut fungi for bioprocess operations.

76. ** Solomon KV, Haitjema CH, Henske JK, Gilmore SP, BorgesRivera D, Lipzen A, Brewer HM, Purvine SO, Wright AT, Theodorou MK et al.: Early-branching gut fungi possess a large, comprehensive array of biomass-degrading enzymes. Science 2016, 351:1192-1195. This paper provides one of the earliest omics-based analyes of anaerobic fungi. These authors demonstrate the ability of anaerobic fungi to express enormous amounts of biomass degrading enzymes. This work also shows the ability of anaerobic fungi to match enzyme expression to feedstock complexity thus paving the way to understand how anaerobic fungal enzymes are tightly regulated.

77. Wilson CA, Wood TM: The anaerobic fungus Neocallimastix frontalis: isolation and properties of a cellulosome-type enzyme fraction with the capacity to solubilize hydrogenbond-ordered cellulose. Appl Microbiol Biotechnol 1992, 37:125-129.

78. Strauss J, Mach RL, Zeilinger S, Hartler G, Sto¨ ffler G, Wolschek M, Kubicek C: Crel, the carbon catabolite repressor protein from Trichoderma reesei. FEBS Lett 1995, 376:103-107.

79. Solomon KV, Henske JK, Gilmore SP, Lipzen A, Grigoriev IV, Thompson D, O'Malley MA: Catabolic repression in earlydiverging anaerobic fungi is partially mediated by natural antisense transcripts. Fungal Genet Biol 2018, 121:1-9. This work provides one of the only published mechanisms of CAZyme regulation in anaerobic fungi and provides insight into the complexity of these fungal genomes.

80. ** Mondo SJ, Dannebaum RO, Kuo RC, Louie KB, Bewick AJ, LaButti K, Haridas S, Kuo A, Salamov A, Ahrendt SR et al.: Widespread adenine N6-methylation of active genes in fungi. Nat Genet 2017, 49:964-968. This work analyses multiple early diverging fungi and elucidates the role of adenine methylation and genomic context and function. Notably, these authors discuss the potential roles of adenine methylation in anaerobic fungi around open reading frames.

81. Fischer M, Durand R, Fe` vre M: Characterization of the "promoter region" of the enolase-encoding gene enol from the anaerobic fungus Neocallimastix frontalis: sequence and promoter analysis. Curr Genet 1995, 28:80-86. This work provides the only published promoter sequence of anaerobic fungi.

82. Durand R, Rascle C, Fischer M, Fevre M: Transient expression of the b-glucuronidase gene after biolistic transformation of the anaerobic fungus Neocallimastix frontalis. Curr Genet 1997, 31:158-161. This paper demonstrates a potential transformation method for introducing nucleic acids into the nuclei of anaerobic fungi.

83. Calkins SS, Elledge NC, Mueller KE, Marek SM, Couger M, Elshahed MS, Youssef NH: Development of an RNA interference (RNAi) gene knockdown protocol in the anaerobic gut fungus Pecoramyces ruminantiumstrain C1A. PeerJ 2018, 6:e4276. This paper provides the only published evidence for natural competency in anaerobic fungi and leverages it to influence gene expression. This work paves the way for future synthetic biology approaches to engineering anaerobic fungi.

84. Vyas VK, Barrasa MI, Fink GR: A Candida albicans CRISPR system permits genetic engineering of essential genes and gene families. Sci Adv 2015, 1:e1500248.

85. Druzhinina IS, Kubicek CP: Genetic engineering of Trichoderma reesei cellulases and their production. Microb Biotechnol 2017, 10:1485-1499.

86. Grahl N, Demers EG, Crocker AW, Hogan DA: Use of RNAprotein complexes for genome editing in non-albicans Candida species. mSphere 2017, 2:e00218-00217.

87. Norton EL, Sherwood RK, Bennett RJ: Development of a CRISPR-Cas9 system for efficient genome editing of Candida lusitaniae. mSphere 2017, 2:e00217-00217.

88. Derntl C, Kiesenhofer DP, Mach RL, Mach-Aigner AR: Novel strategies for genomic manipulation of Trichoderma reesei with the purpose of strain engineering. Appl Environ Microbiol 2015, 81:6314-6323 AEM. 01545-01515.

89. Wang Q, Cobine PA, Coleman JJ: Efficient genome editing in Fusarium oxysporum based on CRISPR/Cas9 ribonucleoprotein complexes. Fungal Genet Biol 2018, 117:21-29.

90. Liang L, Liu R, Garst AD, Lee T, Nogue´ VSI, Beckham GT, Gill RT: CRISPR EnAbled trackable genome engineering for isopropanol production in Escherichia coli. Metab Eng 2017, 41:1-10.

91. Warner JR, Reeder PJ, Karimpour-Fard A, Woodruff LB, Gill RT: Rapid profiling of a microbial genome using mixtures of barcoded oligonucleotides. Nat Biotechnol 2010, 28:856-862.