

UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

PROTEIN ENGINEERING OF CAS9 FOR SAFER GENOME TOOLS, AND  
PHYLOGENETIC ANALYSIS OF CORONAVIRUS SPIKE PROTEIN FOR EFFICIENT  
ANTIVIRAL TARGETS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

SYDNEY NEWSOM

Norman, OK

2022

PROTEIN ENGINEERING OF CAS9 FOR SAFER GENOME TOOLS, AND  
PHYLOGENETIC ANALYSIS OF CORONAVIRUS SPIKE PROTEIN FOR EFFICIENT  
ANTIVIRAL TARGETS

A DISSERTATION APPROVED FOR THE DEPARTMENT OF CHEMISTRY AND  
BIOCHEMISTRY

BY THE COMMITTEE CONSISTING OF

Dr. Rakhi Rajan, Chair

Dr. Charles Rice

Dr. Elena Zgurskaya

Dr. Laura-Isobel McCall

Dr. Krithi Sankaranarayanan





## Acknowledgements

I would like to acknowledge Dr. Hari Priya Parameshwaran, graduate of the Rajan lab at the University of Oklahoma Chemistry and Biochemistry Department, for co-authored work in the Frontiers article and designing and cloning the SpyCas9<sup>H982A</sup> variant. Also, graduate student Richard Van from the Shao lab at the University of Oklahoma Chemistry and Biochemistry Department for doing molecular dynamics simulations on spike protein. Further acknowledgements go to post-doctoral fellow Kesavan Babu, director the University of Oklahoma Protein Purification and Production Core Dr. Philip Bourne, and graduate student Tamara Flusche from the University of Oklahoma Chemistry and Biochemistry Department for cloning and purifying some of the SpyCas9 variants for nitroxide labeling. There are also multiple undergraduate students who have worked alongside me throughout the past few years, and I would like to thank Christian DeMaio, Michael Ting, and Yadin Joseph for being good learners, diligent workers, and very helpful collaborators! Other lab mates I am privileged to have worked with over the years are Dr. Ramya Sundaresan, Dr. Mason Van Orden, Bailey Lacey, Lindsie Martin, Chhandosee Ganguly, and Saadi Rostami. I would also like to acknowledge Dr. Duen-Shian Wang from the Liu lab at the University of North Texas Health Sciences Center for the molecular dynamics simulations with SpyCas9<sup>H982A</sup>, and Dr. Peter Qin and the University of Southern California for sharing Matlab scrips for kinetics analysis of H982A. I would like to thank my committee members (including former committee members Dr. Ann West and Dr. Bradley Stevenson) for keeping me on track throughout my graduate education at OU. Specific thanks to Dr. Laura-Isobel McCall and Dr. Krithi Sankaranarayanan for being my first-year lab rotation mentor and bioinformatics professor respectively. I would also like to say thanks to Dr. Shanteri Singh and Dr. Susan Schroeder for the opportunity to be their teaching assistant. Heartfelt thanks go out to Dr. Fares Najar, who is the former director of the University of Oklahoma Chemistry & Biochemistry Bioinformatics Core and currently a bioinformatics research scientist at Oklahoma State University, for teaching me to do bioinformatics and always

being available to advise me on my projects. I would like to thank two friends, Bob Petron and Cindy Kimmell for their encouraging words. Most of all, I would like to thank Dr. Rakhi Rajan, my principal investigator at the University of Oklahoma Chemistry and Biochemistry Department, for the opportunity to learn and work together since the Summer of 2017. It has been wonderful!

## Table of Contents

|                                                                                                                                                                                                              |      |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------|
| Acknowledgments.....                                                                                                                                                                                         | iv   |
| List of abbreviations.....                                                                                                                                                                                   | vii  |
| Abstract.....                                                                                                                                                                                                | viii |
| <a href="#">Chapter 1</a> : Introduction to the Roles of CRISPR-Cas in Adaptive Immunity and Various Aspects of Bacterial Physiology, and its Repurposing for Nucleic Acid Manipulation Biotechnologies..... | 1    |
| 1.1.0 Copyright information.....                                                                                                                                                                             | 1    |
| 1.2.0 Overview.....                                                                                                                                                                                          | 1    |
| 1.3.0 Defense against mobile genetic elements (MGEs) using CRISPR-Cas.....                                                                                                                                   | 2    |
| 1.4.0 Physiological roles of CRISPR-Cas observed in bacteria.....                                                                                                                                            | 11   |
| 1.5.0 CRISPR-Cas Biotechnology Tools.....                                                                                                                                                                    | 15   |
| 1.6.0 Outlook.....                                                                                                                                                                                           | 24   |
| 1.7.0 Appendix.....                                                                                                                                                                                          | 26   |
| <a href="#">Chapter 2</a> : Engineering a safer SpyCas9-based gene therapy tool by a H982A substitution in the RuvC active site to reduce guide RNA-free DNA cleavage.....                                   | 27   |
| 2.1.0 Copyright information.....                                                                                                                                                                             | 27   |
| 2.2.0 Overview.....                                                                                                                                                                                          | 27   |
| 2.3.0 Introduction.....                                                                                                                                                                                      | 28   |
| 2.4.0 Methods.....                                                                                                                                                                                           | 32   |
| 2.4.1 Protein variant construction and purification.....                                                                                                                                                     | 32   |
| 2.4.2 sgRNA transcription.....                                                                                                                                                                               | 35   |
| 2.4.3 Substrate production for gRNA-dependent DNA cleavage assays.....                                                                                                                                       | 36   |
| 2.4.4 <i>In vitro</i> SpyCas9 activity assays.....                                                                                                                                                           | 36   |
| 2.4.5 Quantification of DNA cleavage.....                                                                                                                                                                    | 37   |
| 2.4.6 Setup of SpyCas9-ntDNA complex models for molecular dynamics (MD) simulations.....                                                                                                                     | 39   |
| 2.4.7 Procedure of MD simulations.....                                                                                                                                                                       | 40   |
| 2.4.8 Protein sequence conservation analysis of SpyCas9's H982-H985 region.....                                                                                                                              | 42   |
| 2.4.9 Statistical analyses.....                                                                                                                                                                              | 43   |

|                                                                                                                                                                                                                    |                                                                                                                                                                              |    |
|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 2.5.0                                                                                                                                                                                                              | Results.....                                                                                                                                                                 | 43 |
| 2.5.1                                                                                                                                                                                                              | Residues in the RuvC active site of SpyCas9 contributes differently towards guide RNA-free DNA cleavage.....                                                                 | 43 |
| 2.5.2                                                                                                                                                                                                              | His982 of SpyCas9's RuvC active site is a major enhancer of gRNA-free DNA cleavage.....                                                                                      | 47 |
| 2.5.3                                                                                                                                                                                                              | Substitution of His982 does not severely impair guide gRNA-dependent DNA cleavage.....                                                                                       | 56 |
| 2.5.4                                                                                                                                                                                                              | Substitution of His982 enhances DNA mismatch sensitivity of SpyCas9.....                                                                                                     | 62 |
| 2.5.5                                                                                                                                                                                                              | H982A proposedly modulates gRNA-free DNA cleavage by preventing proper positioning of single-stranded DNA and by weakening the interaction between $Mn^{2+}(A)$ and D10..... | 64 |
| 2.6.0                                                                                                                                                                                                              | Discussion.....                                                                                                                                                              | 67 |
| 2.6.1                                                                                                                                                                                                              | $Mn^{2+}$ binding in RuvC active site allosterically improves the interaction between SpyCas9 and ntDNA.....                                                                 | 67 |
| 2.6.2                                                                                                                                                                                                              | Phylogenetic conservation of histidine residues in Cas9 active sites.....                                                                                                    | 68 |
| 2.6.3                                                                                                                                                                                                              | RuvC is organized as an autonomous endonuclease that can access and cleave DNA without gRNA-dependent conformational changes.....                                            | 71 |
| 2.6.4                                                                                                                                                                                                              | RuvC active site modifications can potentially remove gRNA-free and off-target DNA cleavages.....                                                                            | 72 |
| 2.6.5                                                                                                                                                                                                              | Fine-tuning the SpyCas9 active site divalent metal ion interactions for a safer gene editing tool.....                                                                       | 74 |
| <a href="#">Chapter 3</a> : Distinguishing the roles of helix-integrity <i>versus</i> protein-nucleic acid interactions of the SpyCas9 bridge helix in modulating RNA-guided DNA cleavage target selectivity ..... |                                                                                                                                                                              | 75 |
| 3.1.0                                                                                                                                                                                                              | Authorship.....                                                                                                                                                              | 75 |
| 3.2.0                                                                                                                                                                                                              | Overview.....                                                                                                                                                                | 75 |
| 3.3.0                                                                                                                                                                                                              | Introduction.....                                                                                                                                                            | 76 |
| 3.4.0                                                                                                                                                                                                              | Methods.....                                                                                                                                                                 | 80 |
| 3.4.1                                                                                                                                                                                                              | Construction of BH loop variants to measure contributions of bridge helix integrity and sidechain interaction in target DNA cleavage sensitivity.....                        | 80 |
| 3.4.2                                                                                                                                                                                                              | Designing SpyCas9 constructs for nitroxide labeling for SDSL EPR studies.....                                                                                                | 80 |
| 3.4.3                                                                                                                                                                                                              | Cloning and purifying SpyCas9 protein variants.....                                                                                                                          | 81 |

|                                                                                                                                             |                                                                                                                                                                                                                                          |     |
|---------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 3.4.4                                                                                                                                       | <i>In vitro</i> gRNA-guided double stranded DNA cleavage assays.....                                                                                                                                                                     | 84  |
| 3.4.5                                                                                                                                       | Quantification of <i>in vitro</i> gRNA-guided double stranded DNA cleavage assays.....                                                                                                                                                   | 84  |
| 3.5.0                                                                                                                                       | Results.....                                                                                                                                                                                                                             | 85  |
| 3.5.1                                                                                                                                       | SpyCas9 <sup>2Ala</sup> has a higher RNA-guided DNA linearization rate than both SpyCas9 <sup>WT</sup> and SpyCas9 <sup>2Pro</sup> .....                                                                                                 | 85  |
| 3.5.2                                                                                                                                       | Integrity of BH is essential for Cas9 to cleave PAM-proximal mismatch containing DNA.....                                                                                                                                                | 90  |
| 3.5.3                                                                                                                                       | DNA linearization is reduced in SpyCas9 <sup>2Pro</sup> with PAM-proximal mismatches indicating better performance for gene editing.....                                                                                                 | 93  |
| 3.5.4                                                                                                                                       | EPR variants are active for RNA-guided DNA cleavage.....                                                                                                                                                                                 | 96  |
| 3.6.0                                                                                                                                       | Discussion and future directions.....                                                                                                                                                                                                    | 98  |
| 3.6.1                                                                                                                                       | BH integrity plays a more important role than sidechain interaction in DNA mismatch sensitivity.....                                                                                                                                     | 98  |
| 3.6.2                                                                                                                                       | SpyCas9 <sup>2Ala</sup> 's high DNA cleavage rate and comparable selectivity for on-target cleavage rate relative to SpyCas9 <sup>WT</sup> may lend it to certain biotechnology applications.....                                        | 100 |
| 3.6.3                                                                                                                                       | Mechanistic insights into the effects of the two-proline and two-alanine substitutions on SpyCas9's on-target DNA linearization selectivity reveal more about how the actions of HNH and RuvC are coordinated for DNA linearization..... | 101 |
| 3.6.4                                                                                                                                       | Future prospects.....                                                                                                                                                                                                                    | 101 |
| <a href="#">Chapter 4</a> : Predicting Potential Antiviral targets in Coronavirus-2 Spike Protein Using Sequence Conservation Analysis..... |                                                                                                                                                                                                                                          | 103 |
| 4.1.0                                                                                                                                       | Authorship.....                                                                                                                                                                                                                          | 103 |
| 4.2.0                                                                                                                                       | Overview.....                                                                                                                                                                                                                            | 103 |
| 4.3.0                                                                                                                                       | Introduction.....                                                                                                                                                                                                                        | 104 |
| 4.3.1                                                                                                                                       | Coronavirus taxonomy.....                                                                                                                                                                                                                | 105 |
| 4.3.2                                                                                                                                       | CoV-2 genes and proteins.....                                                                                                                                                                                                            | 106 |
| 4.3.3                                                                                                                                       | Spike protein structure and function.....                                                                                                                                                                                                | 106 |
| 4.4.0                                                                                                                                       | Methods.....                                                                                                                                                                                                                             | 110 |
| 4.4.1                                                                                                                                       | Curating the spike protein sequence database.....                                                                                                                                                                                        | 110 |
| 4.4.2                                                                                                                                       | Measuring spike protein conservation.....                                                                                                                                                                                                | 113 |

|                                                                                                                                                                         |     |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.4.3 Identifying small-molecule-binding pockets in spike protein using molecular dynamics simulations and molecular docking.....                                       | 113 |
| 4.4.4 Analyzing biochemical data on spike expression and spike:ACE2 binding.....                                                                                        | 114 |
| 4.5.0 Results.....                                                                                                                                                      | 115 |
| 4.5.1 Spike protein regions vary in amino acid conservation across SARS, MERS, and CoV-2 strains.....                                                                   | 115 |
| 4.5.2 Molecular docking studies reveal small molecules that can bind to different pockets in the S1 domain of spike protein.....                                        | 119 |
| 4.5.3 Mapping of the binding pockets onto the spike protein trimer shows different interfaces of RBD as effective drug binding sites.....                               | 121 |
| 4.5.4 Conservation of spike protein residues in small-molecule-binding-pockets.....                                                                                     | 124 |
| 4.5.5 Conserved spike residues that are important for expression and spike:ACE2-binding are found in small-molecule-binding pockets in the receptor binding domain..... | 126 |
| 4.5.6 Conservation of some spike residues changes when zoonotic-isolate spike sequences are added to a dataset of human-infecting CoV spikes.....                       | 132 |
| 4.5.7 Conserved residues have been shown to be important for biological function, viral evolution, and antibody binding.....                                            | 139 |
| 4.6.0 Discussion and future directions.....                                                                                                                             | 147 |
| 4.6.1 Reasoning behind targeting highly conserved <i>versus</i> variable proteins or amino acid residues.....                                                           | 147 |
| 4.6.2 Future Prospects.....                                                                                                                                             | 148 |
| <a href="#">Chapter 5</a> : Outlook.....                                                                                                                                | 149 |
| 5.1.0 Summary.....                                                                                                                                                      | 149 |
| 5.1.1 CRISPR-Cas-related projects.....                                                                                                                                  | 149 |
| 5.1.2 Coronavirus spike protein project: Prioritizing antiviral targets in coronavirus-2 (CoV-2) spike using sequence conservation analysis.....                        | 152 |
| 5.2.0 Future directions.....                                                                                                                                            | 153 |
| 5.2.1 High fidelity Cas9 variants for safer gene therapy tools.....                                                                                                     | 153 |
| 5.2.2 Small molecule drugs to prevent infection from present and future CoV-2 strains.....                                                                              | 155 |
| References.....                                                                                                                                                         | 157 |
| Supplementary file: Cas9-seq.csv.....                                                                                                                                   | 179 |

## List of abbreviations

ds – double stranded; ss - single stranded; CRISPR – clustered, regularly interspaced short palindromic repeats; crRNA – CRISPR RNA (crRNA); Cas – CRISPR-associated, HD – histidine and aspartate; tracrRNA – *trans*-activating crRNA; HEPN – higher eukaryotes and prokaryotes nucleotide-binding; BLP – bacterial lipoprotein; scaRNA – small, CRISPR-Cas-associated RNA; TLR – toll-like receptor; LsrF – 3-hydroxy-5-phosphonooxypentane-2,4-dione thiolase; CovR/S – control of virulence regulator and sensor; FasA/FasB – fibronectin/fibrinogen-binding / hemolytic activity / streptokinase regulatory (Fas) system; TCS – two-component system; VicR/K – a TCS with a putative histidine kinase (Vick) and a response regulator (VicR); Mga – multigene regulator in GAS; ScpA – streptococcal C5a peptidase; SIC – streptococcal inhibitor of complement; QS – quorum sensing; PCD – programmed cell death; cOA – cyclic oligo-adenylate; HGT – horizontal gene transfer; SpyCas9 – *Streptococcus pyogenes* Cas9; sgRNA – single guide RNA; TALENS – Transcription activator-like effector nucleases; NHEJ – non-homologous end joining; HDR – homology directed repair; dCas – nuclease inactivated dead Cas protein; MGEs – Mobile genetic elements; crRNAs – CRISPR RNAs; gRNA – guide RNA; SpyCas9H982A – SpyCas9 H982A variant; SpyCas9WT – SpyCas9 wild-type; CRISPR – clustered regularly interspaced short palindromic repeats; Cas – CRISPR associated; PAM – protospacer adjacent motif; PFS – protospacer-flanking site; sgRNA – single guide RNA; tracrRNA – trans-activating crRNA; tDNA – complementary strand target strand DNA; ntDNA – non-complementary strand non-target strand DNA; dsDNA – double stranded DNA; REC – recognition; NUC – nuclease; binary complex – Cas9-sgRNA; ternary complex – Cas9-sgRNA-DNA; ntDNA – non-target DNA; R-I – RNA-independent; CjeCas9 – *Campylobacter jejuni* Cas9; SDM – site-directed mutagenesis; BH – bridge helix; SpyCas9<sup>2Pro</sup> – SpyCas9 L64P/K65P variant; SpyCas9<sup>2Ala</sup> – SpyCas9 L64A/K65A variant; EPR – electron paramagnetic resonance; SDSL – site-directed spin labeling; MCH, MM3, MM5, and MM18 – double stranded DNA plasmid substrate completely matched or with a mismatch to the gRNA at the 3<sup>rd</sup>, 5<sup>th</sup>, or 18<sup>th</sup> nucleotide position respectively for



SpyCas9 RNA-guided DNA cleavage;  $k_{\text{obs}}$  – observed rate constant;  $k_{\text{fast}}$  – rate constant in the fast phase of the reaction;  $k_{\text{slow}}$  – rate constant in the slow phase of the reaction;  $\alpha$  – active fraction;  $\alpha_{\text{fast}}$  – active fraction in the fast phase of the reaction;  $\alpha_{\text{slow}}$  – active fraction in the slow phase of the reaction; MSE – mean squared error; SEM – standard error of the mean; ACE2 – Angiotensin-converting enzyme 2; 229E – human coronavirus-229E; OC43 – human coronavirus-OC43; NL63 – human coronavirus-NL63, HKU1 – coronavirus-HKU1, SARS – severe acute respiratory syndrome coronavirus; MERS – Middle East Respiratory Syndrome coronavirus; CoV-2 – SARS-CoV-2; RBD – receptor binding domain.

## Abstract

Clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR-associated (Cas) proteins are adaptive immune systems that protect bacteria and archaea from mobile genetic elements. These systems have been repurposed for gene editing biotechnologies. CRISPR-Cas9-based gene therapies are in clinical trials, but a safety concern is that *Streptococcus pyogenes* Cas9 can trigger non-specific DNA damage when not bound to guide RNA (gRNA). Such gRNA-free damage to genomic DNA in human nuclei was reported recently with *Streptococcus pyogenes* Cas9 (SpyCas9) in human cells transfected with plasmids carrying the Cas9 but devoid of a gRNA, signifying the prevalence of such promiscuous DNA damage under cellular conditions. To eliminate these non-specific cleavage activities, SpyCas9 variants are being developed through our research. One mutation that successfully removes Mn<sup>2+</sup>-dependent gRNA-free DNA cleavage is SpyCas9<sup>H982A</sup>. Elucidating the fidelity-increasing DNA cleavage mechanisms of SpyCas9<sup>H982A</sup> will advance the development of safe Cas9-based gene therapies.

Another safety concern is the finding that SpyCas9 cuts the DNA at unintended sites that partially hybridize to the guide RNA. Effects of such off-targeting include increased risks of cancer and other health problems. The SpyCas9 L64P/K65P substitution variant has been shown to have higher selectivity such that it is less likely to cleave DNA when there is a mismatch to the guide RNA. In this research, a SpyCas9 L64A/K65A variant was constructed and assayed for RNA-guided DNA cleavage to determine whether the loss of intra-protein interactions or compromised structure of the helix containing L64/K65 has the primary role in eliciting this increase in specificity. In vitro DNA cleavage assays showed that the compromised helical structure causes the increase in selectivity for on-target as opposed to PAM-proximal mismatched DNA cleavage in terms of linearization efficiency.

A novel coronavirus outbreak of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) occurred in December 2019, and variants of the virus have been emerging and causing surges in cases, hospitalizations, deaths, and societal

disruptions over the past few years. Vaccines have been developed, but research on a variety of vaccines and potentially antiviral targets for small-molecule drugs is helpful for preventing infections from future variants. Initial contact between human cells and SARS-CoV-2 is mediated by the viral Spike (S) protein. The objective of this research is to identify antiviral target sites in S protein and categorize them into sites that are highly evolving and that are highly conserved in a range of human-infecting coronaviruses. The rationale is that targeted drug discovery to these two types of sites will enable better preparedness in tackling coronavirus as it continues to evolve. S protein sequence conservation was analyzed among severe disease causing human coronaviral strains (including MERS, SARS-CoV, and SARS-CoV-2), along with molecular docking of the S protein receptor binding domain with a compound library. Our results identified four pockets and six compounds that are promising for rational drug development.

# 1 Chapter 1: Introduction to the Roles of CRISPR-Cas in Adaptive Immunity and Various Aspects of Bacterial Physiology, and its Repurposing for Nucleic Acid Manipulation Biotechnologies

## 1.1.0 Copyright information

This introductory chapter was originally published in *Frontiers Cellular and Infection Microbiology* on January 28th, 2021 (Newsom et al., 2021). This *Frontiers* article, copyright © 2021 Newsom, Parameshwaran, Martin, and Rajan, is under the Creative Commons Attribution License (CC BY) which states that the article contents can be shared (copied or redistributed) and adapted as long as appropriate credit and a [link to the license](#) are given. Terms of the license are that changes must be indicated and no additional restrictions may be applied. Differences between the *Frontiers* article and this dissertation are noted in the appendix (section 1.7.0).

## 1.2.0 Overview

Bacterial and archaeal CRISPR-Cas systems offer adaptive immune protection against foreign mobile genetic elements (MGEs). This function is regulated by sequence specific binding of CRISPR RNA (crRNA) to target DNA/RNA, with an additional requirement of a flanking DNA motif called the protospacer adjacent motif (PAM) in certain CRISPR systems. This chapter presents how the same fundamental mechanism of RNA-DNA and/or RNA-RNA complementarity is utilized by bacteria to regulate two distinct functions: to ward off intruding genetic materials and to modulate diverse physiological functions. The best documented examples of alternate functions are bacterial virulence, biofilm formation, adherence, programmed cell death, and quorum sensing. While extensive complementarity between the crRNA and the targeted DNA and/or RNA seems to constitute an efficient phage protection system, partial complementarity seems to be the key for several of the characterized alternate functions. Cas proteins are also involved in sequence-specific and non-sequence specific RNA cleavage

and control of transcriptional regulator expression, the mechanisms of which are still elusive. Over the past decade, the mechanisms of RNA-guided targeting and auxiliary functions of several Cas proteins have been transformed into powerful gene editing and biotechnological tools. This chapter describes the biology of CRISPR-Cas and provides a synopsis of CRISPR-based technologies. Even with the abundant mechanistic insights and biotechnology tools that are currently available, the discovery of new and diverse CRISPR types holds promise for future technological innovations, which will pave the way for precision genome medicine.

#### 1.3.0 Defense against mobile genetic elements (MGEs) using CRISPR-Cas

In the microscopic world, bacteria and archaea are invaded by various nucleic acids called mobile genetic elements (MGEs), such as viruses and plasmids, that may co-opt cellular machinery and damage the host (Jackson et al., 2017; Koonin and Wolf, 2008). To defend themselves against MGEs, about 50% of bacteria and about 87% of archaea have an adaptive immune system called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) and CRISPR-associated (Cas) proteins (Jackson et al., 2017). These systems use a guide RNA (gRNA) to guide a nuclease to the MGE *via* base complementarity (Bolotin et al., 2005b; Makarova et al., 2006; Mojica et al., 2005). The guide region of the gRNA is integrated into the self-genome through an adaptation process that occurred when the organism was previously exposed to the MGE (Makarova et al., 2006). The Cas ribonuclear protein (RNP) follows certain rules to distinguish self *versus* non-self nucleic acid using mechanisms such as protospacer adjacent motif (PAM) recognition or protospacer flanking site (PFS) discrimination (Abudayyeh et al., 2016b; Mojica et al., 2009). The RNP locates a specific region in the intruder genome based on complementarity with the gRNA. This leads to a nuclease, which is either a part of the RNP or an independent Cas or another cellular nuclease, cleaving and inactivating the MGE, offering protection to the host organism (Gasiunas et al., 2012; Jinek et al., 2012). There is also evidence that besides this adaptive immunity role, CRISPR-Cas systems

may have an additional role in endogenous gene regulation (Faure et al., 2019; Liu and Doudna, 2020; Murugan et al., 2017).

#### 1.3.1 Discovery of CRISPRs, spacer origins, and CRISPR-associated (Cas) gene functions

CRISPR-Cas systems are encoded in bacterial and archaeal genomes, and they were first discovered by scientists analyzing an archaeal genome for clues about how it adapts to saline environments (Mojica et al., 1993). It was noticed that many archaeal and bacterial genomes have clusters of regularly interspaced short palindromic repeats (Bolotin et al., 2005b; Makarova et al., 2006; Mojica et al., 2005). These clusters of regularly interspaced short palindromic repeats are known as the CRISPR arrays, which are transcribed and processed into individual CRISPR RNAs (crRNAs) each containing a “repeat” region that facilitates complex formation with the Cas nuclease and a “spacer,” harvested by the CRISPR-Cas system during a previous encounter with the MGE, that base pairs to the target MGE (Charpentier et al., 2015; Gasiunas et al., 2012; Jinek et al., 2012). Bioinformatics analysis revealed that many of the spacers are fragments of MGE sequences (Horvath et al., 2008) and that a PAM sequence is needed adjacent to that fragment in the MGE to make that MGE a target for CRISPR-Cas-mediated defense (Mojica et al., 2009). It was also observed that certain CRISPR-associated (Cas) genes are encoded near the CRISPR array, but the function of these genes was unknown (Makarova et al., 2006). There were hypotheses that these proteins had roles in gene repair (Makarova et al., 2002), but their roles in CRISPR-Cas adaptive immunity, such as nucleases to destroy MGEs and integrases to insert spacers from MGEs into the CRISPR array, were eventually discovered (Barrangou et al., 2007; Beloglazova N, 2008; Brouns et al., 2008; Marraffini and Sontheimer, 2008; Nunez et al., 2014).

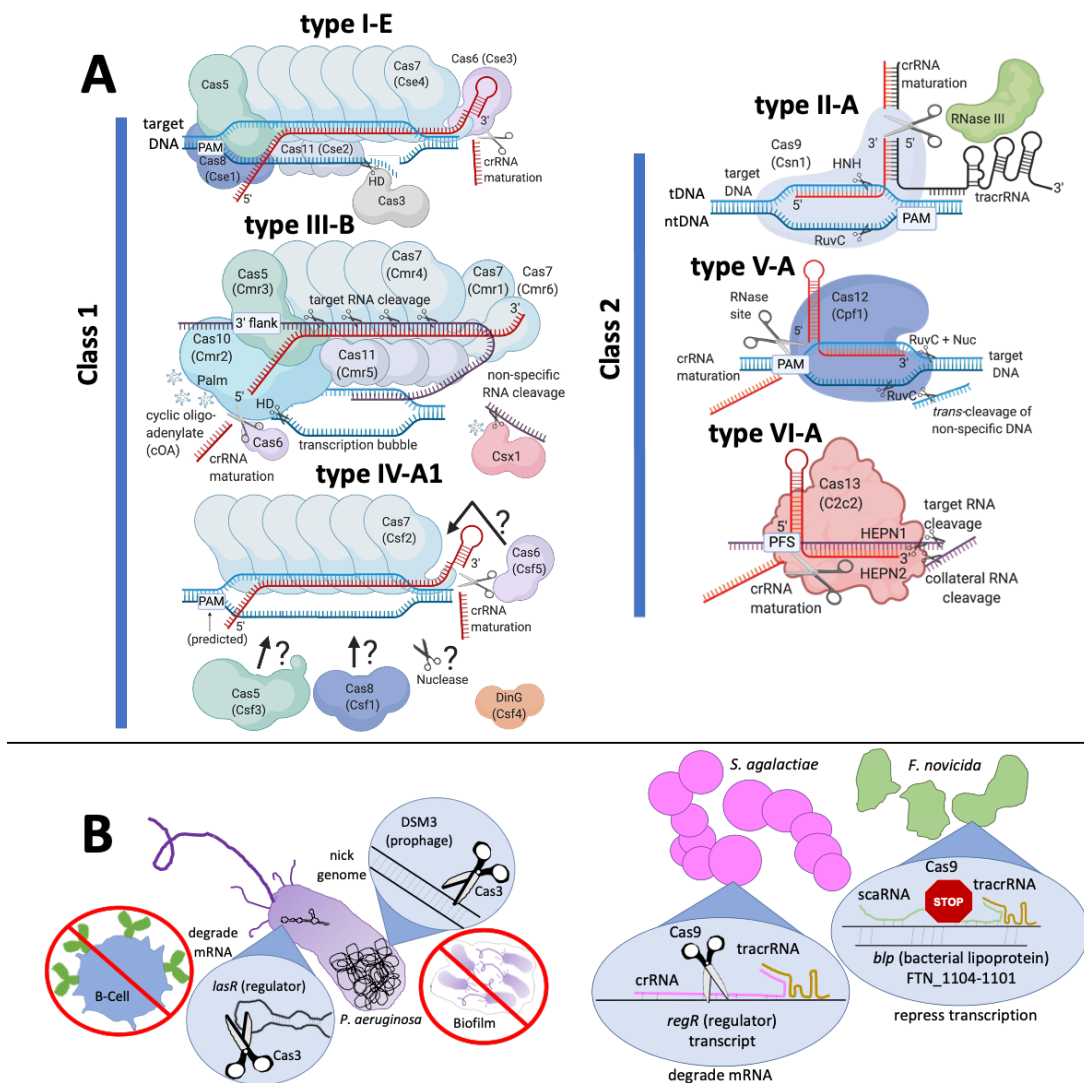
Finally, the three stages of the CRISPR-Cas defense mechanism were established as follows:

1. Adaptation – Adaptation is the process by which bacteria and archaea use Cas1 and Cas2 proteins to integrate new spacer sequences into their CRISPR arrays to offer future immunological protection against a specific intruder (Jackson et al., 2017). Cas1 and Cas2 are conserved in most CRISPR-Cas systems characterized so far (Makarova et al., 2020). The Cas1-Cas2 heterohexameric complex (4 Cas1-2 Cas2) binds to pieces of short DNA produced by the cellular machinery and integrates into the CRISPR array following several mechanisms that maintain site-specific integration (Alkhnbashi et al., 2016; Amitai and Sorek, 2016; Flusche and Rajan; Heler et al., 2014; Mosterd et al., 2020). For type II-A CRISPR-Cas systems, the integration location within the CRISPR array is determined differently for three distinct subgroups type II-A adaptation machinery (Van Orden et al., 2017; Van Orden et al., 2020). In each CRISPR-Cas type, these spacers lead to individual crRNAs that possess the ability to sequence-specifically target a plethora of unique complementary regions.
2. CRISPR-associated RNA (crRNA) processing – Cas proteins process the CRISPR array transcript into individual crRNAs that can complex with Cas nucleases to guide them to target MGEs (Behler and Hess, 2020). This crRNA processing is done by specific Cas proteins and sometimes in association with cellular RNases to create unique sequence regions in the crRNAs that are essential for their functionality (Behler and Hess, 2020).
3. Interference – The Cas nuclease in complex with the crRNA sequence specifically cleave the target nucleic acid (Plagens et al., 2015). In certain CRISPR systems, the sequence-specific cleavage can turn on non-sequence specific nucleic acid cleavage that causes host cell death or cell arrest that is a mechanism to cope with the intruder MGE (Abudayyeh et al., 2016b; Kazlauskienė et al., 2017).

### 1.3.2 Classification of CRISPR-Cas systems

There are a variety of CRISPR-Cas systems each with a different approach to interfering with MGEs. There are two classes distinguished based on whether

interference is performed by a multi-protein complex (Class 1) or a single protein (Class 2) (Koonin et al., 2017). Furthermore, there are types within each class that are differentiated based on the signature Cas protein that is essential for interference (Makarova and Koonin, 2015). There are also subtypes within each type that are differentiated based on the molecular phylogeny of the CRISPR locus (Makarova and Koonin, 2015). An overview of this classification scheme and examples of subtype-specific crRNA processing and interference mechanisms are shown in Figure 1.1.





**Figure 1.1. CRISPR-Cas classification, interference complexes, and**

**alternate functions. A)** Interference complexes, crRNA maturation, and nucleic acid targeting. Type I-E CASCADE from *Escherichia coli* is shown (Brouns et al., 2008; Jore et al., 2011). The interference complex comprises six Cas7s, two Cas11s, and one each of Cas5, Cas8, and Cas6. This complex recruits Cas3, which uses a histidine aspartate (HD) nuclease domain to nick the target DNA, followed by unwinding of the DNA by helicase domain and long-range degradation of DNA by the HD domain. Type III is an RNA-targeting system with a historical distinction into two groups of effector complexes, “Csm” (Cas subtype Mtube, includes subtypes A, D, E and F) and “Cmr” (Cas module RAMP, includes subtypes B and C), differentiated based on the sequences of Cas11 subunit, called as Csm2 and Cmr5 respectively in Csm and Cmr complexes (Tamulaitis et al., 2017). Type III-B from *Pyrococcus furiosus* is shown (Cocozaki et al., 2012). The interference complex comprises six Cas7s, coded by three distinct genes (*cmr4*, *cmr1* and *cmr6*), three Cas11s, and one each of Cas5 and Cas10. While Cas7 cleaves target RNA with a 6-nucleotide periodicity, Cas10 cleaves ssDNA in the transcription bubble using its HD domain (Kazlauskienė et al., 2016; Mogila et al., 2019; Taylor et al., 2015; You et al., 2019). The Palm domain of Cas10 produces cyclic oligo-adenylate (cOA) that activates Csx1 for non-specific RNA cleavage (Kazlauskienė et al., 2017). Type IV is represented by subtype A1 from *Aromatoleum aromaticum* (Özcan et al., 2019). The interference complex comprises Cas5, Cas6, Cas8, and multiple Cas7s, although the exact subunit stoichiometry is unknown. DinG is a helicase necessary for interference (Crowley et al., 2019; Özcan et al., 2019; Pinilla-Redondo et al., 2020). Question marks and arrows in type IV-A1 indicate that subunit organization and nuclease identity is unknown. Cas6, the protein required for crRNA processing may (for example, CasE, alternatively known as Cas6e (Uniprot accession number Q46897), in the type I-E CASCADE complex (Brouns et al., 2008)) or may not (for example, Cas6 in mature type III-B Cmr complexes (Mogila et al., 2019)) be part of the effector complex in Class I systems. Each class II system is represented by its respective subtype “A”. Type II-A Cas9 uses

its HNH endonuclease domain to cleave the DNA strand hybridized to crRNA and RuvC endonuclease domain to cleave the other strand (Jinek et al., 2012). RNase III is required for crRNA processing in type II systems (Deltcheva et al., 2011). Type V-A Cas12a (previously called Cpf1) uses RuvC domain to cleave both strands of the target DNA, even though assistance of the Nuc domain is needed to cleave the strand hybridized to crRNA (Zetsche et al., 2015a). Cas12a also possesses non-specific DNA cleavage, called *trans*-cleavage (Li et al., 2018a). Type VI-A Cas13a uses two higher eukaryotes and prokaryotes nucleotide-binding (HEPN) domains to elicit sequence-specific and collateral (non-specific) cleavage of target RNA (Abudayyeh et al., 2016a). Type V-A and type VI systems possess specific sites within their effector protein for crRNA processing (East-Seletsky et al., 2016; Fonfara et al., 2016). Secondary trimming for crRNA maturation is not shown. Protein legacy names are given in parentheses. PAM stands for protospacer adjacent motif, PFS stands for protospacer flanking site, and 3' flank stands for 3' flanking sequence of the protospacer that is non-complementary to the 5'-tag of crRNA; each of which are used for self *versus* non-self recognition. Nucleic acids are not to scale with proteins. Target or non-specifically cleaved RNA is in purple, crRNA is red, tracrRNA is in black and DNA is blue. DNA strand complementary to the crRNA (target strand DNA, tDNA) and the non-complementary strand of the double stranded DNA (non-target strand DNA, ntDNA) are labeled for type II-A. **B)** Left: *Pseudomonas aeruginosa* (type I-F) Cas3-mediated RNA and DNA targeting affect host immune evasion and biofilm formation respectively. Right: Comparison between *Streptococcus agalactiae* (type II-A) and *Francisella novicida* (type II-B) mechanisms of Cas9-mediated down-regulation of endogenous genes. The differences include the type of guide RNA (crRNA *versus* scaRNA), nucleic acid target (mRNA *versus* genomic DNA), mode of repression (transcript degradation *versus* steric hindrance to RNA-polymerase), and endogenous gene target (master regulatory protein *versus* regulon affecting virulence). Created with BioRender.com in the original publication (Newsom et al., 2021).

Some interesting points about each subtype are as follows:

#### Type I – the plentiful type

Type I systems are the most abundant in nature (Koonin et al., 2017; Zheng et al., 2020). The interference mechanism of type I adaptive immunity involves Cas3 (a Cas protein recruited by the crRNA-guided interference complex) unwinding target DNA and then cutting and degrading the DNA strand that is single stranded in the R-loop formed between the DNA and the crRNA (Figure 1.1A) (Zheng et al., 2020).

#### Type II – the pioneer type

Type II systems are the best characterized for use in gene editing biotechnologies, and Cas9 from *Streptococcus pyogenes* (SpyCas9) is currently the most widely used CRISPR-Cas-based gene editing biotechnology (Jinek et al., 2012; Liu et al., 2020). The interference mechanism of type II adaptive immunity involves Cas9 complexing with crRNA and *trans*-acting crRNA (tracrRNA) (Figure 1.1A), which aids in crRNA processing and RNP complex formation with Cas9 (Charpentier et al., 2015; Deltcheva et al., 2011), to make a double stranded break in the target DNA (Jinek et al., 2012). Cas9 has two endonuclease domains, HNH cleaving the target strand and RuvC cleaving the non-target strand creating a double stranded cut that is trimmed further by the RuvC domain to produce 3' recessed DNA ends in the protospacer region that flanks the PAM (Jinek et al., 2012; Stephenson et al., 2018). Cas9-induced DSBs are repaired in mammalian cells primarily through non-homologous end joining (NHEJ) though it is possible to induce some level of homology directed repair (HDR) through editing cells when they are in the S phase of the cell cycle (Brandsma and van Gent, 2012; Engstrom and Kmiec, 2008) (Sansbury et al., 2019). Another way to favor HDR is by suppressing the NHEJ pathway (Chu et al., 2015). NHEJ tends to introduce frame shifts in gene sequences containing the DSB (Kim and Kim, 2014), which makes Cas9 a suitable tool for knocking out

genes; however, its ability to induce HDR makes it a suitable tool for gene replacement as well (Chu et al., 2015).

#### Type III – the versatile type

Type III systems use multiple activities to defend against MGEs:

Spatiotemporally controlled DNA-targeting, both specific and non-specific RNA-targeting, and small-molecule signaling between the Cas proteins to name a few (Figure 1.1A) (Kolesnik et al., 2021). The interference mechanism of type III adaptive immunity starts with RNA-guided RNA-cleavage by Cas7 in the interference complex (which contains multiple other Cas proteins as well as crRNA) (Kazlauskienė et al., 2016). When the target-RNA is in a transcription bubble, the Cas10 in the interference complex degrades the single stranded DNA in the transcription bubble in a spatiotemporally controlled manner since this degradation is activated when the interference complex binds to the RNA target (Kazlauskienė et al., 2016; Mogila et al., 2019; Taylor et al., 2015; You et al., 2019). The Cas10 also produces cyclic oligo adenylates that signal for Csx1 to start non-specific RNA-cleavage that leads to cell death/arrest because it does not differentiate between intruder and host RNA (Kazlauskienė et al., 2017).

#### Type IV – the mysterious type

Type IV systems are the least characterized and the function or mechanism of many of the protein components is unknown (Figure 1.1A) (Taylor et al., 2021). The interference mechanism of type IV adaptive immunity involves an unknown nuclease doing crRNA-guided DNA targeting.

#### Type V – the sidekick type

Type V systems are the second most characterized for gene editing biotechnologies, and the potential uses are somewhat complementary to those of type-II (Liu et al., 2020). The interference mechanism of type V adaptive immunity involves the Cas12-crRNA RNP making a double stranded break in the target DNA (Figure 1.1A) (Zetsche et al., 2015a). The Cas12a DSB has

staggered ends (Zetsche et al., 2015a). Relative to that of Cas9, the Cas12a DSB-induced genome repair response in mammalian cells tends to be HDR as opposed to NHEJ (Sansbury et al., 2019). When template DNA is delivered along with the Cas12a RNP components to the mammalian cell, HDR replaces the gene near the double stranded break with the template DNA (Kim and Kim, 2014), which makes Cas12a a suitable tool for knocking in genes. Cas12a has a non-specific DNA cleavage activity that is activated after binding to the target DNA (Chen et al., 2018; Li et al., 2018a; Parameshwaran et al., 2021), and this activity is useful for diagnostic tools (Table 1.1.F.1)

#### Type VI – the selfless type

Type VI systems have a “self-destruct” function that kills the host cell possibly as an altruistic attempt to protect other bacteria from viral infection (Faure et al., 2019; Koonin and Zhang, 2017). The interference mechanism of type VI adaptive immunity involves the Cas13-crRNA RNP doing RNA-guided RNA targeting when the protospacer flanking site (PFS) requirements are met (Figure 1.1A) (Abudayyeh et al., 2016a; O'Connell, 2019). PFS is similar to PAM, but it tends to be less of a requirement for a specific motif and more of a requirement for certain nucleobases to be absent near the target site (Gleditsch et al., 2019). Target recognition triggers a secondary activity called “collateral cleavage” which is non-sequence specific RNA-degradation activity responsible for “self-destruction” of the infected cell (Abudayyeh et al., 2016a). Type VI systems lend themselves to RNA-targeting knockdown biotechnology applications (Abudayyeh et al., 2017; Konermann et al., 2018), which, apart from the non-specific RNA-cleavage activity inherent to this CRISPR-Cas type (Kelley et al., 2022; Konermann et al., 2018), could be considered safer than the DNA-targeting knockout applications because the effect is more transient since only the transcripts, not the template DNA, are destroyed (Cox et al., 2017; Kushawah et al., 2020). Interestingly, it has been observed that collateral cleavage does not occur when *Leptotrichia shahii* Cas13a is used for RNA knockdown in mammalian cells (Abudayyeh et

al., 2017), so Cas13's collateral cleavage activity does not disqualify it from gene therapy biotechnology applications.

#### 1.4.0 Physiological roles of CRISPR-Cas observed in bacteria

Some bacteria regulate physiological processes using RNA-guided Cas proteins that target the self-genome. While partial complementarity between the guide RNA and the self-genome or endogenous transcripts seems to drive gene regulation, complete complementarity with the genome has been shown to regulate prophage and temperate phage life-cycles and trigger bacterial evolution through recombination to avoid lethal self-targeting (Aklujkar and Lovley, 2010; Goldberg et al., 2014; Vercoe et al., 2013). Selected examples CRISPR systems regulating bacterial pathogenicity and survival are presented in this section.

#### 1.4.1 Virulence

##### 1.4.1.1 Host immune evasion

CRISPR-Cas systems can regulate gene expression to conceal bacteria from host toll-like receptors (TLR). The best studied example is the *Francisella novicida* U112 type II-B system, which downregulates a bacterial lipoprotein (BLP) on the cell envelope to evade host immune responses (Ratner et al., 2019; Sampson et al., 2013). This mechanism requires the association of Cas9 with tracrRNA and small, CRISPR-Cas-associated RNA (scaRNA), which specifically represses transcription of a regulon containing *blp* (FTN\_1104-1101) (Figure 1.1B). Transcriptomic studies have revealed similar trends where *blps* are upregulated in *cas9* deletion mutants ( $\Delta cas9$ ) of *Streptococcus agalactiae* GD201008-001 (type II-A) and *Riemerella anatipestifer* (type II-C) (Ma et al., 2018; Wang et al., 2019). In *Streptococcus pyogenes* GAS-M1T1-5448 (type II-A),  $\Delta cas9$  produces less master regulator protein Mga, which in turn downregulates ScpA and SIC (streptococcal inhibitor of complement-mediated lysis protein) proteins, which are essential to inactivate the host's complement immune defense mechanism (Gao et al., 2019). *S. pyogenes* Cas9 also downregulates the CovR/S two component system (TCS) which regulates

capsule genes that confer antiphagocytic properties (Gao et al., 2019; Sarkar and Sumbly, 2017). Involvement of a type I-F CRISPR-system in virulence was demonstrated in *Pseudomonas aeruginosa* UCBPP-PA14, where Cas3 degrades the mRNA of the master regulator protein LasR to evade the TLR-initiated host immune response (Figure 1.1B, (Li et al., 2016)).

#### 1.4.1.2 Bacterial adherence and cytotoxicity

CRISPR-Cas systems modulate bacterial-host interactions, the mechanisms of which are currently evolving. Relative to the wild-type, *Neisseria*, *Streptococcus*, and *Campylobacter*,  $\Delta cas9$  strains adhere poorly to mammalian host cells (Gao et al., 2019; Heidrich et al., 2019; Ma et al., 2018; Shabbir et al., 2018; Spencer et al., 2019). The *S. pyogenes* mechanism involves Cas9-mediated down-regulation of the FasA/FasB TCS that downregulates adhesion factors (Gao et al., 2019). *Campylobacter jejuni* and *S. agalactiae* Cas9s can increase bacterial cytotoxicity, adhesion, and survival within host cells (Louwen et al., 2013; Ma et al., 2018; Saha et al., 2020a; Shabbir et al., 2018). In *S. agalactiae*, hyaluronidase activity essential for blood-brain barrier penetration is regulated via Cas9-mediated cleavage of the master regulator *regR* transcript (Figure 1.1B, (Ma et al., 2018)).

#### 1.4.1.3 Group behavior

There is growing evidence that CRISPR systems can regulate group behaviors to increase bacterial virulence. Based on *Salmonella enterica* serovar Enteritidis  $\Delta cas3$  (type I-E) transcriptome studies, Cas3 downregulates LsrF production, preventing degradation of quorum sensing (QS) signaling molecules by LsrF, ultimately increasing expression of the *Lsr* QS operon (Cui et al., 2020).

Transcriptomic studies of  $\Delta cas3$  *Streptococcus mutans* UA159 (type I-C) showed downregulation of biofilm formation genes controlled by the VicR/K TCS known to influence streptococcal tissue specificity (Tang et al., 2019; Vega et al., 2016). Phage-dependent biofilm regulation was observed in *Pseudomonas aeruginosa* strain UCBPP-PA14, where partial crRNA hybridization to lysogenized DMS3

phage recruits Cas3 (type I-F) to nick the bacterial genome, initiating RecA-dependent SOS response that inhibits biofilm formation and swarming (Figure 1.1B, (Heussler et al., 2015)). Experiments have implicated Cas9-based virulent group behaviors in *C. jejuni* (type II-C) (Shabbir et al., 2018), *S. pyogenes* (type II-A) (Gao et al., 2019), *Neisseria meningitidis* (type II-C) (Sampson et al., 2013), and *S. agalactiae* (type II-A) (Ma et al., 2018) where  $\Delta cas9$  strains had decreased biofilm formation.

#### 1.4.2 Relation to Genome Repair and Involvement in Non-Virulent Gene Regulation and Programmed Cell Death

Current research also shows CRISPR-Cas systems regulating genes not directly involved in bacterial virulence. Experimental evidence has demonstrated Cas1's ability to cleave branched DNA substrate and revealed enhanced sensitivity of  $\Delta cas1$  *E. coli* to DNA damage (Babu et al., 2011). A recent study has shown that non-homologous end-joining (NHEJ) does not co-exist with type II-A systems due to the competition between Csn2 (type II-A) and NHEJ proteins for DNA substrates (Bernheim et al., 2017). Another DNA-repair system, RecBCD, is essential for spacer acquisition in *E. coli* (type I-E) (Levy et al., 2015). These studies suggest complex interdependence of CRISPR and DNA repair systems, the mechanisms of which are still elusive. Transcriptomic and proteomic studies on  $\Delta cas9$  *S. pyogenes* (type II-A) (Gao et al., 2019), Group B *Streptococcus* (type II-A) (Spencer et al., 2019), and *Riemerella anatipestifer* (type II-C) (Wang et al., 2019) have indicated both up- and down-regulation of several endogenous genes, the implications of which are not clear. Several studies in *Myxococcus xanthus* have shown that CRISPR systems are involved in cell-stress dependent sporulation (type I-C) and fruiting body development (type III-B) (Viswanathan et al., 2007; Wallace et al., 2014).

Type III and type VI CRISPR systems promote indiscriminate RNA cleavage leading to programmed cell death (PCD), a strategy employed when immune protection has failed (Faure et al., 2019). In type III systems, crRNA binding to



target RNA activates production of cyclic oligoadenylate (cOA) by Cas10. The cOA molecules initiate Csm6/Csx1's indiscriminate RNase activity resulting in cell death (Kazlauskienė et al., 2017; Niewoehner et al., 2017). In type VI, crRNA binding to target RNA activates Cas13 for promiscuous RNase cleavage (Abudayyeh et al., 2016a). Interestingly, the resemblance of Cas2 to the toxin component of toxin-antitoxin systems and the genomic co-localization of CRISPR and PCD elements have led to the idea of their co-dependence in bacterial physiology and are awaiting experimental confirmation (reviewed in Faure et al., 2019).

#### 1.4.3 Diversity of CRISPR-Cas Mechanisms

Current studies indicate that CRISPR has functions beyond adaptive immunity, primarily in regulating genome content and gene expression. Horizontal Gene Transfer (HGT) can be negatively impacted in CRISPR-containing bacteria since the acquired DNA can be targeted by CRISPR. The effects of CRISPR on HGT are evolving. Some bacteria compensate limitations on HGT by maintaining a defective CRISPR locus or establishing mechanisms for CRISPR-tolerance (Sampson and Weiss, 2013; Wimmer and Beisel, 2020). A different perspective on offsetting this disadvantage has been demonstrated by CRISPR-mediated enhancement of transduction by phages which then enables transfer of genetic materials between bacteria (Watson et al., 2018).

Based on studies of streptococcal physiology, we speculate that CRISPR-mediated gene regulation promotes strain differentiation via genome remodeling and regulatory changes. CRISPR changes expression of *S. pyogenes* master regulators Mga and VicR and the TCS CovR/CovS, which then regulate immunomodulatory virulence factors that drive development of strains with different host tissue preferences and physiologies ranging from hypervirulent to carrier status (Sarkar and Sumby, 2017; Vega et al., 2016). Similarly, natural mutation in *P. aeruginosa* master regulator LasR increases bacterial fitness in the cystic fibrosis lung (Smith et al., 2006). A common theme in these diverse

mechanisms is that Cas proteins target genes that bring about differential phenotypes in bacteria. *Legionella pneumophila*'s requirement of Cas2 to infect amoebas indicates the involvement of diverse Cas proteins in endogenous gene regulation, requiring mechanistic characterizations of CRISPR in bacterial physiology (Gunderson and Cianciotto, 2013). Interestingly, the versatile mechanisms derived so far for both adaptive and physiological functions have been harnessed to develop an impressive set of molecular tools for nucleic acid manipulation.

#### 1.5.0 CRISPR-Cas Biotechnology Tools

The RNA-guided nucleic acid targeting of Cas proteins offer several biotechnology tools dependent on the diverse CRISPR mechanisms as shown in Table 1.1 and in reviews (Pickar-Oliver and Gersbach, 2019; Xu and Qi, 2019).

| <b>Biotechnology Application</b> | <b>Cas Protein and CRISPR-Cas type, relevant activities, modifications</b>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             | <b>References</b>                                                                                                                                                                                                                                                                                                                                                                                                                    |
|----------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| A. Gene editing                  | <p>1a. SpyCas9 (type II-A), sequence specific DNA cleavage producing a mixture of staggered and blunt ends</p> <p>1b. Hyper-accurate Cas9 (HypaCas9), derivative of SpyCas9 holding four amino acid mutations near the RNA-DNA interface</p> <p>1c. High-fidelity SpyCas9 (SpCas9-HF), mutation of four amino acids that non-specifically contact target DNA</p> <p>1d. Enhanced specificity SpyCas9 (eSpCas9), three amino acid mutations in a groove that stabilizes the non-target strand of the target DNA</p> <p>1e. Light activated Cas9, split into two fragments and fused to dimerization domains</p> <p>1f. sgRNA regulation to control Cas9</p> <p>1g. chemically inducible Cas9, various modifications</p> | <p>1a. (Cho et al., 2013; Chu et al., 2015; Jiang et al., 2013; Jinek et al., 2012)</p> <p>1b. (Chen et al., 2017)</p> <p>1c. (Kleinstiver et al., 2016)</p> <p>1d. (Slaymaker et al., 2016)</p> <p>1e. (Nihongaki et al., 2015; Yu et al., 2020)</p> <p>1f. (Ferry et al., 2017; Lee et al., 2016b; Liu et al., 2016; Maji et al., 2017; Tang et al., 2017)</p> <p>1g. (Davis et al., 2015; Dow et al., 2015; Liu et al., 2016;</p> |

|  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                    |
|--|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|  | <p>1h. alternative Cas9 orthologs: <i>Neisseria meningitidis</i>, <i>Staphylococcus aureus</i>, and <i>Francisella novicida</i></p> <p>2. Cas12a (type V-A), sequence specific dsDNA cleavage producing staggered ends</p> <p>3. CASCADE (type 1-E), sequence specific DNA binding, fused to FokI nuclease</p> <p>4. CASCADE and Cas3 (type I-A), sequence specific DNA cleavage and degradation of one of the strands of a DNA</p> <p>5. Cmr (type III-B), sequence specific interference of transcriptionally active DNA</p> <p>6. Direct injection of CRISPR-Cas-based drug into the retina to treat hereditary blindness</p> <p>7. Cas10-Csm and Csm6 (type III-A), viral DNA and RNA cleavage that increases viral gene editing efficiency by selecting against unedited phage</p> | <p>Maji et al., 2017; Zetsche et al., 2015b)</p> <p>1h. ((Lee et al., 2016a); (Ran et al., 2015); (Acharya et al., 2019))</p> <p>2. (Ferenczi et al., 2017; Hur et al., 2016; Yan et al., 2017)</p> <p>3. (Cameron et al., 2019)</p> <p>4 &amp; 5. (Li et al., 2015)</p> <p>6. ClinicalTrials.gov Identifier: NCT03872479</p> <p>7. (Nayeemul Bari and Hatoum-Aslan, 2019)</p> <p>8. (Mayo-Muñoz et al., 2018)</p> |
|--|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

|                                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                            |
|--------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
|                                      | <p>8. Anti-CRISPR proteins, CRISPR-Cas inhibition as a selectable marker for editing viral genomes</p> <p>9. Anti-CRISPR proteins, CRISPR-Cas inhibition quenches gene editing reactions to minimize off-targeting</p>                                                                                                                                                                                                                                                                                                                                                            | 9. (Marino et al., 2020)                                                                                                   |
| B. Base editing                      | <p>1. SpyCas9 nickase (type II-A), sequence specific DNA binding, fused to both cytidine deaminase and uracil DNA glycosylase inhibitor</p> <p>2. SpyCas9 (type II-A), sequence specific DNA binding, nuclease inactivated and fused to an engineered adenosine to inosine deaminase capable of targeting DNA</p> <p>3. nuclease inactivated Cas12a (type V-A), sequence specific DNA binding, fused to cytidine deaminase and uracil DNA glycosylase inhibitor</p> <p>4. nuclease inactivated Cas13 (type VI-B), sequence specific RNA binding, fused to adenosine deaminase</p> | <p>1. (Komor et al., 2016)</p> <p>2. (Gaudelli et al., 2017)</p> <p>3. (Li et al., 2018b)</p> <p>4. (Cox et al., 2017)</p> |
| C. Gene knockdown via ssRNA cleavage | 1. <i>Staphylococcus aureus</i> Cas9 (II-A), ssRNA cleavage                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | 1. (Strutt et al., 2018)                                                                                                   |

|                               |                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                                                                            |
|-------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                               | <p>2. <i>Neisseria meningitidis</i> Cas9 (II-C), ssRNA cleavage</p> <p>3. Cmr (III-B), ssRNA cleavage</p> <p>4. Cas13 (VI-A and D), crRNA maturation and ssRNA cleavage</p> <p>5. nuclease inactivated SpyCas9 (type II-A) fused to RNA endonuclease domain, sequence specific ssRNA binding</p>                                                                                                                                                      | <p>2. (Rousseau et al., 2018)</p> <p>3. (Zebec et al., 2014)</p> <p>4. (Abudayyeh et al., 2017; Konermann et al., 2018)</p> <p>5. (Batra et al., 2017)</p> |
| D. Transcriptional repression | <p>1. nuclease inactivated SpyCas9 (type II-A), sequence specific DNA binding</p> <p>2. nuclease inactivated SpyCas9 (type II-A) fused to an epitope tag that recruits antibody-fused methyltransferase that methylates the genomic target to repress transcription, sequence specific DNA binding</p> <p>3. nuclease inactivated Cas12a (type V-A), crRNA maturation and sequence specific DNA binding</p> <p>4. CASCADE (type I-E), DNA binding</p> | <p>1. (Qi et al., 2013)</p> <p>2. (Huang et al., 2017)</p> <p>3. (Zhang et al., 2017)</p> <p>4. (Rath et al., 2014)</p>                                    |

|                           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |                                                                                                                                                 |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------|
| E. Gene activation        | <p>1a. nuclease inactivated SpyCas9 (type II-A) fused to a transactivation domain, sequence specific DNA binding</p> <p>1b. Light activated SpyCas9, nuclease inactivated and fused to a light-inducible heterodimerizing protein</p> <p>2. Cas12a (type V-A), crRNA maturation and sequence specific DNA binding, fused to a transactivation domain</p> <p>3. CASCADE (type I-E), crRNA maturation and sequence specific DNA binding, fused to plant transcriptional activation domain</p> | <p>1a. (Perez-Pinera et al., 2013)</p> <p>1b. (Polstein and Gersbach, 2015)</p> <p>2. (Breinig et al., 2019)</p> <p>3. (Young et al., 2019)</p> |
| F. Nucleic acid detection | <p>1. Cas12a (type V-A), collateral ssDNA cleavage activity upon target binding</p> <p>2. Cas13 (type VI-A and B), collateral ssRNA cleavage activity upon target binding</p> <p>3. Csm6 (type III-A), non-specific nuclease</p>                                                                                                                                                                                                                                                            | <p>1-3. (Gootenberg et al., 2018)</p> <p>1. (Chen et al., 2018)</p>                                                                             |
| G. Antimicrobial          | <p>1. SpyCas9 (type II-A), dsDNA interference</p> <p>2. CASCADE and Cas3 (type I-E), dsDNA interference</p>                                                                                                                                                                                                                                                                                                                                                                                 | <p>1. (Citorik et al., 2014)</p> <p>2. (Gomaa et al., 2014)</p>                                                                                 |

|                                                 |                                                                                                                                                                        |                                                                 |
|-------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
|                                                 | <p>3. Cas13a (type VI-A), promiscuous ssRNA cleavage</p> <p>4. Anti-CRISPR proteins, CRISPR-Cas inhibition to expand the host range for phage-derived therapeutics</p> | <p>3. (Kiga et al., 2020)</p> <p>4. (Marino et al., 2020)</p>   |
| H. Transposition or genome tagging              | <p>1. CASCADE (type I-F), sequence specific DNA binding, fused to transposition protein</p> <p>2. Cas1 and Cas2 (type I-E), sequence-selective DNA integration</p>     | <p>1. (Klompe et al., 2019)</p> <p>2. (Wright et al., 2017)</p> |
| I. Protect plants from both RNA and DNA viruses | <p>1. Cas13 (type VI-A), crRNA maturation and sequence specific ssRNA cleavage</p> <p>2. SpyCas9 (type II-A), sequence specific DNA interference</p>                   | <p>1. (Aman et al., 2018)</p> <p>2. (Ali et al., 2015)</p>      |
| J. Decrease bacterial pathogenicity             | Anti-CRISPR proteins to inhibit CRISPR-Cas systems involved in bacterial pathogenicity                                                                                 | (Marino et al., 2020)                                           |
| K. Record transcription                         | Cas1:Cas2 complex (type III-D), natural fusion to reverse transcriptase, reverse transcription and sequence selective DNA integration                                  | (Schmidt et al., 2018)                                          |
| L. Store digital information                    | Cas1:Cas2 complex (type I-E), sequence selective DNA integration                                                                                                       | (Shipman et al., 2017)                                          |



|                                                                                                                                                                                                                                           |                                                                                                                                                                                                         |                                                                                       |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| M. Visualizing nucleic acid localization                                                                                                                                                                                                  | <p>1. nuclease inactivated SpyCas9 (type II-A) fused to GFP, sequence specific binding to target DNA</p> <p>2. nuclease inactivated Cas13 (type VI-A) fused to GFP, sequence specific ssRNA binding</p> | <p>1. (Chen et al., 2013; Nelles et al., 2016)</p> <p>2. (Abudayyeh et al., 2017)</p> |
| N. Mediate chromatin immunoprecipitation for downstream analysis of epiproteome (interacting proteins, histone post translational modifications, RNAs, and neighboring genomic regions) via mass spectrometry and nucleic acid sequencing | nuclease inactivated and affinity tagged SpyCas9 (type II-A), sequence specific DNA binding                                                                                                             | (Fujita and Fujii, 2015; Waldrip et al., 2014)                                        |
| O. Subnuclear proteomic profiling                                                                                                                                                                                                         | nuclease inactivated SpyCas9 (type II-A) fused to APEX2 (an engineered soybean ascorbate peroxidase that labels near-by proteins with biotin), sequence specific DNA binding                            | (Gao et al., 2019)                                                                    |

|                                                                                                                                             |                                                                                |                          |
|---------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------|
| P. Influence mRNA splicing outcomes without hindering translation to protein (potential treatment for diseases caused by gene mis-splicing) | Cas13d (CasRx, type VI-D), sequence specific DNA binding, nuclease inactivated | (Konermann et al., 2018) |
|---------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------|

**Table 1.1.** CRISPR-Cas biotechnology applications

### 1.5.1 Gene Editing

The revolutionary characteristic of CRISPR-Cas-based gene editing is the ability to introduce heritable genome modifications by programming Cas proteins using guide RNAs, instead of modifying the editing protein, which was the limiting aspect of zinc fingers and TALENS (Gaj et al., 2013). Discovery of CRISPR-Cas-based genome editing tools was awarded the Nobel prize in 2020.

The most widely used CRISPR-based gene editing system is *Streptococcus pyogenes* Cas9 (SpyCas9) with a single guide RNA (sgRNA), where the crRNA and tracrRNA are fused into sgRNA for easier delivery options (Table 1.1A). The most prominent application is modification of eukaryotic genes utilizing efficient DNA-repair pathways of eukaryotes to repair Cas9 induced DNA breakage by NHEJ (creates gene knock-outs) or homology directed repair (HDR, creates gene knock-ins). An approach to fix genetic mutations while eliminating unwarranted effects from off-target DNA cleavage is fusing Cas9 with base editors (Table 1.1B, (Komor et al., 2016)). Cas9-based gene editing in prokaryotes is limited due to inefficient DNA repair mechanisms; however, Cas9 cleavage-induced killing of unedited cells can be used to increase efficiency of other bacterial gene-editing methods (Wilson et al., 2003).

Cas12a and CASCADE gene editing systems complement those of Cas9 as their DNA cleavage mechanisms produce HDR-enhancing staggered ends and instill long-range deletions respectively (Dolan et al., 2019; Paul and Montoya, 2020).

#### 1.5.2 Transcript and Epigenetic Manipulation

Transcripts can be targeted and cleaved by Cas13 and some Cas9 orthologs including *Staphylococcus aureus* Cas9 and *Neisseria meningitidis* Cas9 for transcript level gene knock-down (Table 1.1C, (Konermann et al., 2018; Rousseau et al., 2018; Strutt et al., 2018)). Nuclease inactivated dead Cas (dCas) proteins can block or activate transcription *via* steric hindrance of RNA-polymerase or fusion to transcriptional activators respectively (Table 1.1.D and E, (Perez-Pinera et al., 2013; Qi et al., 2013)).

#### 1.5.3 Nucleic Acid Detection Kits

Cas12a and Cas13 inherently cleave non-specific nucleic acids after RNA-mediated target binding (Figure 1.1A), enabling several fluorescent-based detection kits (Table 1.1F, (Gootenberg et al., 2018)). Notably, STOP (SHERLOCK Testing in One Pot) for COVID-19 diagnostics (STOPCovid) uses this technology (Joung et al., 2020).

#### 1.5.4 Biomedical Tools and Future Prospects in Genome Medicine

The site-specific genetic modulation facilitated by Cas proteins has been monumental in quickly advancing gene therapy with several ongoing clinical trials for cancer immunotherapies, blood disorders, blindness etc. (Table 1.1.A.6, [clinicaltrials.gov](https://clinicaltrials.gov)). To facilitate progress, SpyCas9 off-targeting concerns are being addressed using highly stringent Cas proteins, temporal control of protein expression, and gene editing reaction activators/quenchers (Table 1.1.A, (Chen et al., 2017)). As new CRISPR systems with diverse mechanisms are being characterized, new biomedical applications will follow.

#### 1.6.0 Outlook

As described here, CRISPR-Cas is a two-in-one mechanism for protection against intruding nucleic acids as well as for regulating bacterial physiology, including pathogenicity. Following the current discovery trends, genomic analyses will keep unearthing new CRISPR-Cas systems, sometimes even re-writing the existing rules (e.g. discovery of first known Cas9 in nanoarchaea, (Burstein et al., 2017)). The mechanistic basis of CRISPR-based physiology is still evolving. Interestingly, most spacer targets are yet to be identified, (Shmakov et al., 2017) and these target locations may hold the key to mechanisms of more CRISPR-mediated alternate functions.

Currently, Cas9, Cas1, Cas2, and Cas3 have been directly implicated in several physiological functions, implicating potential comparable functions in other Cas proteins. So far, Cas9 has been the most divergent in function ranging from phage defense to gene regulation to directly acting as a virulence factor in *C. jejuni* (Faure et al., 2019). Future research is crucial in understanding the structure-function relations of these diverse Cas9 mechanisms. The recent finding that secretion of guide-free Cas9 by *C. jejuni* can inflict host cell DNA damage points to Cas9 mechanisms that are independent of guide RNA (Saha et al., 2020c; Sundaresan et al., 2017).

The arms race between bacteria and viruses has delivered an arsenal of anti-CRISPR proteins (Marino et al., 2020). In an interesting twist, the use of CasF (also known as Cas12j) by huge phages showed adaptation of CRISPR-Cas mechanisms to ward off virophages (Pausch et al., 2020). The fitness cost benefits of an active CRISPR system and the maintenance of degenerated CRISPR-Cas systems in several bacteria provide interesting future research avenues. CRISPR-Cas systems have been unique because of their repurposing into powerful diagnostic, therapeutical and experimental tools. While there are several aspects such as off-target effects and promiscuous DNA damage that need to be fixed by future research (Haapaniemi et al., 2018; Wang et al.,

2020c), the days of personal precision genome medicine are a step closer with this powerful technology.

#### 1.7.0 Appendix

The title of the *Frontiers* review article, The CRISPR-Cas Mechanism for Adaptive Immunity and Alternate Bacterial Functions Fuels Diverse Biotechnologies, was modified to be the title of this dissertation chapter: Introduction to the Roles of CRISPR-Cas in Adaptive Immunity and Various Aspects of Bacterial Physiology, and its Repurposing for Nucleic Acid Manipulation Biotechnologies. Furthermore, sections 1.3.0 – 1.3.2 replaced similar content in the original *Frontiers* article. Figure 1.1 was created together with the co-authors of the *Frontiers* review article and Figure 1.1.A has been used in a modified form in Hari Priya Parameshwaran's Ph.D. dissertation as well.

## 2 Chapter 2: Engineering a safer SpyCas9-based gene therapy tool by a H982A substitution in the RuvC active site to reduce guide RNA-free DNA cleavage

### 2.1.0 Copyright information

This chapter is unpublished biochemical research from the Rajan lab at the University of Oklahoma Department of Chemistry and Biochemistry and *in silico* molecular dynamics analyses by Dr. Duen-Shian, a member of Dr. Jin Liu's laboratory at the University of North Texas Health Sciences Center. Figures 2.1, 2.13, 2.14, and 2.15 were made by Dr. Duen-Shian.

### 2.2.0 Overview

To protect against mobile genetic elements (MGEs), some bacteria and archaea have CRISPR-Cas adaptive immune systems. CRISPR RNAs (crRNAs) bound to Cas nucleases guide the nucleases to MGEs, enabling hybridization of the crRNA and MGE based on sequence complementarity, followed by degradation of the MGE. This programmable sequence-specific DNA cleavage activity has been harnessed for gene editing. CRISPR-based applications possess safety concerns such as off-target and guide RNA (gRNA)-free DNA cleavages, both of which are observed in the Cas nuclease commonly used for gene editing, *Streptococcus pyogenes* Cas9 (SpyCas9). We focused on developing SpyCas9 variants devoid of gRNA-free DNA cleavage activity, while retaining efficient gRNA-dependent DNA cleavage. Towards this, we designed RuvC active site variants based on divalent metal coordination in the active site and developed the SpyCas9 H982A variant (SpyCas9<sup>H982A</sup>) which, compared to SpyCas9 wild-type (SpyCas9<sup>WT</sup>), has ~167-fold reduction in Mn<sup>2+</sup>-dependent gRNA-free DNA cleavage and only ~1.2-fold reduction in the efficiency of gRNA-dependent DNA cleavage. Mechanistic analysis by molecular dynamics simulations shows that Mn<sup>2+</sup>, but not Mg<sup>2+</sup>, produces a gRNA-free DNA cleavage competent state that is disrupted by the H982A substitution. Further assays of DNA cleavage target specificity showed that SpyCas9<sup>H982A</sup> is more selective than is SpyCas9<sup>WT</sup> against DNA substrates that possess a PAM-proximal mismatch with the gRNA.

Phylogenetic analysis of Cas9 protein sequences show that H982 is prevalent in multiple orthologs suggesting that this protein engineering approach can be applied to multiple Cas9 orthologs for developing safer genome tools. Our results demonstrate the feasibility of selectively modulating cation:protein interactions to engineer superior gene editing tools.

### 2.3.0 Introduction

CRISPR-Cas systems are RNA-directed adaptive immune systems present in bacteria and archaea (Bolotin et al., 2005a; Makarova et al., 2006; Mojica et al., 2005). The CRISPR locus retains immunological memory of past mobile genetic element infections as short pieces of intruder genome that are transcribed into CRISPR RNA (crRNA). The crRNA assembles with Cas proteins and the RNA-protein complex is guided, sequence-specifically, to the intruder genome for cleavage and inactivation upon subsequent attacks. The sequence-specificity is established by the complementarity between the guide region of the crRNA and the invading DNA or RNA. CRISPR systems possess inherent mechanisms such as the protospacer adjacent motif (PAM) or the protospacer-flanking site (PFS), which are adjacent to the target and interact with the protein to differentiate self and intruding nucleic acid (Gleditzsch et al., 2019). The gRNA-dependent nucleic acid targeting and cleavage features of Cas proteins have been repurposed into powerful genome and biotechnological applications (Newsom et al., 2020). Of the numerous CRISPR systems (Makarova et al., 2020), Cas9, the signature protein of the type II CRISPR system, is being widely used for gene editing and gene therapy applications due to the simplistic requirements of a single multi-domain Cas9 protein and a single guide RNA (sgRNA, fusion of crRNA and trans-activating crRNA (tracrRNA)) for sequence-specific DNA cleavage (Jinek et al., 2012). Cas9 uses HNH and RuvC endonucleases, respectively, to cleave the complementary strand (target strand DNA, tDNA) and non-complementary strand (non-target strand DNA, ntDNA) of a double stranded DNA (dsDNA) that exists as an R-loop with the sgRNA (Jinek et al., 2012; Nishimasu et al., 2014).

One of the hallmarks of Cas9's DNA cleavage is the intricate and orchestrated conformational changes including the different domains of Cas9, sgRNA and the DNA to establish site-specific DNA cleavage (Jiang and Doudna, 2017). Cas9 has a bilobed architecture consisting of a recognition (REC) lobe and a nuclease (NUC) lobe, which are connected by an arginine-rich bridge helix (Jinek et al., 2014; Nishimasu et al., 2014). The apo-Cas9 crystal structure shows a self-inhibitory form of the protein where the endonuclease domains are not easily accessible for DNA cleavage (Jinek et al., 2014), even though studies have shown that Cas9 can bind non-specifically to DNA in the absence of sgRNA (Sternberg et al., 2014). Upon binding to sgRNA, the REC lobe, especially the REC3 domain, undergoes a large structural rearrangement consisting of ~65 Å rotation towards the HNH domain (Jiang and Doudna, 2017; Jiang et al., 2015; Jinek et al., 2014; Nishimasu et al., 2014). The binary complex conformation has a preordered seed region that is essential for searching complementarity in the target DNA and has pre-formed grooves for binding nucleic acid (Jiang et al., 2015; Jinek et al., 2014). Compared to structural rearrangements in the binary complex, the extent of conformational reorientation in the ternary complex (Cas9-sgRNA-DNA) is smaller. These changes include slight shifts in the positions of REC2 and REC3 and the movement of HNH domain towards the target strand of the DNA (Jiang and Doudna, 2017; Jiang et al., 2016; Jiang et al., 2015). The DNA binding is initiated by searching and locating a cognate PAM, upon which RNA-DNA base pairing occurs leading to the formation of an R-loop (Sternberg et al., 2014; Szczelkun et al., 2014). To further ensure sequence-specificity in DNA cleavage, the endonuclease domains, HNH and RuvC, undergo further conformational transitions that are tightly regulated by RNA:DNA base pairing at the PAM-distal end of the R-loop (Josephs et al., 2016; Sternberg et al., 2015). The RuvC domain is conformationally dependent on HNH transition for activation for DNA cleavage, and once the transitions of HNH and RuvC occurs placing them near the scissile phosphates (Jiang et al., 2016; Sternberg et al., 2015), DNA cleavage occurs in a concerted manner as demonstrated by kinetic analysis of dsDNA scissions (Gong et al., 2018; Raper et al., 2018). Molecular dynamics



simulations revealed allosteric mechanisms of Cas9 activation including crosstalk between different domains of Cas9, activation by PAM-binding and role of the non-target DNA (ntDNA) strand in coordinating the conformational transitions essential for DNA cleavage (Babu et al., 2021; Casalino et al., 2020; East et al., 2020; Nierzwicki et al., 2021a; Palermo, 2019; Palermo et al., 2017b; Zuo and Liu, 2020).

In spite of the requirement of sgRNA-binding and intensive domain reorganizations to activate conformational changes essential for sequence-specific DNA cleavage, we previously showed that Cas proteins can promote promiscuous DNA cleavage when cognate guide-RNA is absent (Sundaresan et al., 2017). This activity, called guide RNA-free (also known as RNA-independent (R-I) and guide-free) DNA cleavage, requires specific divalent metals (such as  $Mn^{2+}$  or  $Co^{2+}$ ) and manifests differently in Cas orthologs. For example, double stranded plasmid DNA is nicked by *Francisella novicida* (Fno) Cas12a and FnoCas9, while this activity is minimal in *Streptococcus pyogenes* (Spy) Cas9 (Sundaresan et al., 2017). With a single stranded circular plasmid, SpyCas9 and FnoCas12a completely degrades the substrate, while this activity is nominal for FnoCas9 (Sundaresan et al., 2017). FnoCas9 uses its HNH domain, SpyCas9 uses its RuvC domain, and FnoCas12a uses the coordinated activities of both RuvC and Nuc domains for gRNA-free DNA cleavage. Interestingly the presence of even one of the cognate RNAs (crRNA or tracrRNA) does not abolish gRNA-free DNA cleavage of SpyCas9 and the gRNA-free DNA cleavage is PAM-independent (Sundaresan et al., 2017). Another study showed that several Cas12a orthologs possess gRNA-free single and dsDNA cleavage activities with diverse divalent cations and that modifying DNAs with phosphorothioate links prevent this promiscuous DNA cleavage (Li et al., 2020). These features of gRNA-free DNA cleavage point to the importance of the consistent presence of equimolar amounts of sgRNA and Cas9 to prevent unwarranted DNA cleavage in biotechnology applications such as gene editing.

A recent study demonstrated that gRNA-free DNA cleavage occurs under cellular conditions (Saha et al., 2020c). This study that characterized gRNA-free activities of SpyCas9 and *Campylobacter jejuni* (Cje) Cas9 in human cell lines demonstrated that Cas9 can localize to the human cell nucleus using a native helix, called the bridge helix, which serves as a nuclear localization signal (NLS). Cas9 induced non-specific DNA damage in human cells which recruited p53 and gH2AX markers when transfected with Cas9 in the absence of sgRNA. A genome wide sequencing analysis showed accumulation of dsDNA breaks with CjeCas9 induction and the position of the break points were PAM independent. Inactivating the endonuclease domains of Cas9 or co-transfecting Cas9 and sgRNA prevented recruitment of DNA damage markers (Saha et al., 2020c). This study establishes that cellular conditions (both divalent metal concentrations and forms of DNA) can promote gRNA-free DNA cleavage activity. Studies have also indicated that Cas9 is a virulence factor in *C. jejuni*, which is secreted out of the bacterium and that CjeCas9 acts as a cytotoxic agent that is essential for a successful infection of human tissue (Saha et al., 2020b).

The delivery approaches for genome applications such as the use of separate plasmids for Cas9 and sgRNA, and possible differences in the half-lives of Cas9 and sgRNA when delivered as a purified Cas9-sgRNA ribonucleoprotein complex can potentially create conditions where Cas9 is free of its sgRNA. Several studies have shown that unexpected effects are possible following Cas9-based gene editing, the reasons of which are unclear currently. Examples include p53 upregulation as a response to DNA breaks (Ihry et al., 2018; Merkle et al., 2017), chromosomal truncations (Cullot et al., 2019; Zuccaro et al., 2020), unintended gene regulations in neurons (Yang et al., 2018) *etc.* gRNA-free DNA cleavage is different from off-target effects of gene editing. Off-target DNA cleavage is associated with sgRNA-guided DNA cleavage of Cas9 where dsDNA breaks are introduced at genomic regions that are only partially complementary to the guide region of the sgRNA (Fu et al., 2013). While off-target effects can be mitigated by sgRNA design strategies (Allen et al., 2020; Rahdar et al., 2015), such efforts

would not remove the potentially deleterious effects of gRNA-free DNA cleavage. Under this scenario, it is essential that we develop Cas protein variants that are devoid of gRNA-free cleavage, while maintaining efficient sgRNA-guided, sequence-specific, DNA cleavage. Furthermore, the molecular mechanism of gRNA-free DNA cleavage should be elucidated to discover better ways of differentially modulating the gRNA-free and gRNA-dependent DNA cleavage activities to engineer safer and more efficient SpyCas9-based gene editing tools.

Towards this goal, a gRNA-free DNA cleavage resistant SpyCas9<sup>H982A</sup> variant was engineered using the following reasoning. Since SpyCas9 uses the RuvC catalytic center to catalyze gRNA-free single stranded DNA (ssDNA) cleavage (Sundaresan et al., 2017), and certain interactions among divalent metal cations, amino acids, and ntDNA might be compensating for the lack of hybridized gRNA:ntDNA to stabilize the RuvC center in a state competent for ntDNA cleavage, removing such interactions could selectively modulate gRNA-free DNA cleavage without significantly reducing the efficiency of sgRNA-guided DNA cleavage. In addition, it has been reported that E188 of *Bacillus halodurans* RNase H, which is spatially and functionally equivalent to H982 in SpyCas9, may play a role in recruiting metal ions to the catalytic center (Nowotny et al., 2005). We therefore hypothesized that the alanine substitutions of H982 and other amino acids in the SpyCas9 RuvC center may disrupt Mn<sup>2+</sup> recruitment thus reducing Mn<sup>2+</sup>-dependent gRNA-free DNA cleavage. Our data show that H982 provides a differential regulation of gRNA-free and sgRNA-guided DNA cleavage activities. The substitution of His 982 to Ala (SpyCas9<sup>H982A</sup>) causes a ~167-fold reduction in Mn<sup>2+</sup>-mediated gRNA-free DNA cleavage with only a slight reduction in sgRNA-guided DNA cleavage. We report methods to further improve the safety of Cas9-based genome editing applications and is highly valuable considering the increased use of Cas9 in gene therapy.

#### 2.4.0 Methods

##### 2.4.1 Protein variant construction and purification

Wild-type SpyCas9 (SpyCas9<sup>WT</sup>) and a H982A substituted variant, SpyCas9<sup>H982A</sup>, were purified using previously published protocols (Jinek et al., 2012; Sundaresan et al., 2017). Briefly, the Ni-NTA column (HisTrap FF Crude, 5 mL) was washed with 15 CV buffer A (50 mM TRIS pH 8, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP) and loaded with clarified cell lysate containing the Cas9 protein. Gradient elution was done using 0-100% Nickel-NTA buffer B (50 mM TRIS pH 8, 500 mM NaCl, 500 mM imidazole, 1 mM TCEP) over 100 mL. The expected size of MBP-SpyCas9 is 202 kDa. Fractions were pooled and dialyzed overnight in cation exchange buffer A (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM DTT, 10% glycerol) at 4°C with a simultaneous incubation with TEV protease to remove the MBP tag. The sample was loaded onto cation exchange column (SP HP, 5 mL) and eluted over 100 mL, 0-100% cation exchange buffer B (20 mM HEPES pH 7.5, 1 M KCl, 2 mM DTT, 10% glycerol). The SpyCas9 protein size after cleavage of MBP tag is 159 kDa. The protein was concentrated using a centrifugal filter and injected onto a size exclusion column (SEC, Hi Prep 16/60 Sephacryl S-300 HR, 120 mL) equilibrated in SEC buffer (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM TCEP) and eluted isocratically with 120 mL of SEC buffer. The protein was quantified using a Nanodrop spectrophotometer, aliquoted, and flash frozen in liquid nitrogen for storage at -80°C. After using site-directed mutagenesis (SDM) (see Table 2.1 for primer sequences) to introduce H982A substitution, SpyCas9<sup>H982A</sup> was purified using the same method as SpyCas9<sup>WT</sup>.

| Name                     | Purpose                                                                                          | Sequence                            |
|--------------------------|--------------------------------------------------------------------------------------------------|-------------------------------------|
| SpyCas9<br>-H982A-<br>FP | Forward<br>PCR primer<br>for SDM to<br>introduce<br>the H982A<br>substitution<br>into<br>SpyCas9 | TGAGATTAACAATTACGCTCATGCCCATGATGCGT |

|                                                    |                                                                                                  |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |
|----------------------------------------------------|--------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| SpyCas9<br>-H982A-<br>RP                           | Reverse<br>PCR primer<br>for SDM to<br>introduce<br>the H982A<br>substitution<br>into<br>SpyCas9 | ACGCATCATGGGCATGAGCGTAATTGTTAATCTCA                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| sgRNA                                              | <i>in vitro</i> DNA<br>cleavage<br>assays                                                        | <p>5'-<br/> <i>gaaattaatacgactcactata</i><b><u>GGG</u></b><b><u>A</u></b><b><u>T</u></b><b><u>T</u></b><b><u>T</u></b><b><u>C</u></b><b><u>T</u></b><b><u>T</u></b><b><u>C</u></b><b><u>T</u></b><b><u>T</u></b><b><u>G</u></b><b><u>C</u></b><b><u>G</u></b><b><u>C</u></b><b><u>T</u></b><b><u>T</u></b><b><u>T</u></b><b><u>T</u></b><b><u>G</u></b><br/> TTTTAGAGCTATGCTGTTTT<b>G</b><b>aaa</b>CAAAACAGCATAGC<br/> AAGTTAAAATAAGGCTAGTCCGTTATCAACTTGAAAAA<br/> GTGGCACCGAGTCGGTGCTT,TTTTAT<b><u>GTCTTC</u></b>-3'</p> <p><i>italics</i> – T7 promoter<br/> <b>BOLD UPPER CASE underlined</b> – extra Gs for enhanced transcription<br/> <b>BOLD UPPER CASE</b> – 20 nt guide region<br/> bold lower case – tetra loop<br/> <u>underlined</u> – BbsI restriction sites<br/> comma (,) – the BbsI cut site</p> |
| Matched<br>(MCH)<br>substrate                      | <i>in vitro</i> DNA<br>cleavage<br>assays                                                        | GATTTCTTCTTGCGCTTTTT                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| Mismatch<br>in<br>position 3<br>(MM3)<br>substrate | <i>in vitro</i> DNA<br>cleavage<br>assays                                                        | <p>GATTTCTTCTTGCGCTT<b>A</b>T</p> <p><b>Bold</b> - mismatch</p>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

|                                                         |                                           |                                                             |
|---------------------------------------------------------|-------------------------------------------|-------------------------------------------------------------|
| Mismatch<br>in<br>position 5<br>(MM5)<br>substrate      | <i>in vitro</i> DNA<br>cleavage<br>assays | GATTTCTTCTTGCGC <b>A</b> TTTT<br><br><b>Bold</b> - mismatch |
| Mismatch<br>in<br>position<br>18<br>(MM18)<br>substrate | <i>in vitro</i> DNA<br>cleavage<br>assays | GA <b>A</b> TTCTTCTTGCGCTTTTT<br><br><b>Bold</b> - mismatch |

**Table 2.1.** Sequences of oligonucleotides used in this study

#### 2.4.2 sgRNA transcription

*In vitro* run-off transcription was used to obtain sgRNA for gRNA-dependent DNA cleavage assays. The transcription template was prepared by linearizing a plasmid (pUC19 backbone, 2.7 kbp) containing the sgRNA sequence (see Table 2.1) using BbsI restriction enzyme (4-hour incubation of 5 µg plasmid template with 20 Units of BbsI in a 100 µL reaction volume at 37°C) and cleaning and concentrating the linearized plasmid using a PCR cleanup kit (E.Z.N.A.® Cycle Pure Kit, Omega Biotech)(Babu et al., 2019). The linearized plasmid DNA template was transcribed using T7-RNA polymerase, the promoter of which is present upstream of the sgRNA gene in the plasmid. The 200 µL transcription contained 1X RNA Polymerase buffer (40 mM TRIS-HCl pH 8.0, 1 mM Spermidine, 50 µg/mL BSA, 20 mM MgCl<sub>2</sub>, 2 mM DTT), 9 mM GTP, 8 mM UTP, 8 mM ATP, 8 mM CTP, 56 Units of RNasin, 1.92 Units of inorganic pyrophosphatase, 2.5 µg of linearized plasmid template, and 40 µg of T7 RNA polymerase (25 µg added initially and 15 µg added at the 2-hour mark). The reaction was incubated for 4 hours at 37°C. The reaction was ethanol precipitated. The pellet was extracted and resuspended in 8 M Urea, 1XTBE (90

mM TRIS-Cl, 90 mM boric acid, 2 mM EDTA) and 0.03% bromophenol blue and purified using a 10% acrylamide gel containing 8M urea. The 120-nucleotide sgRNA band was excised from the gel, crushed, and soaked overnight in RNA elution buffer (50 mM potassium acetate ( $C_2H_3O_2K$ ), 200 mM KCl pH 7), followed by ethanol precipitation and 80% ethanol washes. The RNA pellet was resuspended in nuclease-free water and annealed in 10 mM TRIS pH 8 and 50 mM NaCl by heating to 95 °C for 2 min and then allowing the heat block to cool gradually to room temperature over the course of ~45 min.

#### 2.4.3 Substrate production for gRNA-dependent DNA cleavage assays

Target sequences were ligated into a pUC19 backbone and mutated as described previously (Babu et al., 2019). Plasmids were isolated from DH5 $\alpha$  using a mini-prep kit (E.Z.N.A.® Plasmid DNA Mini Kit I, Omega Biotech, Norcross, GA, D6942). Some plasmid preparations had a higher fraction of non-supercoiled DNA, so for all the substrates, the band corresponding to supercoiled population was gel extracted using a kit (QIAquick Gel Extraction Kit, Qiagen, Hilden, Germany) for use as the substrate.

#### 2.4.4 *In vitro* SpyCas9 activity assays

##### 2.4.4.1 Guide RNA-free single-stranded M13mp18 DNA cleavage

Unless otherwise stated, reaction components contained 1X buffer (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM TCEP), 100 ng (4.5 nM) single stranded plasmid DNA (M13mp18), 10 mM divalent metal, and 100 nM SpyCas9. Reactions were performed at 37°C for 30 min unless otherwise stated. All divalent metals had chloride counterions. We noticed accumulation of reaction products in the wells of the gels when the assay contained  $Co^{2+}$ . To resolve this,  $Co^{2+}$  reactions were stopped with a 5X stock of 50 mM EDTA and 75 mM  $CaCl_2$  (for Proteinase K activity) and treated with 0.8 Units of Proteinase K (New England Biolabs) for 20 min at 50°C. All the reactions were prepared for agarose gel electrophoresis by adding 2X stop dye (100 mM EDTA, 2% SDS, 20% glycerol, and 0.08% Orange G dye). The reaction products were resolved on a

1% agarose gel and the bands were visualized post-gel running using 0.1 µg/mL ethidium bromide.

#### 2.4.4.2 gRNA-dependent double-stranded pUC19 DNA cleavage

Reaction components were 1X cleavage buffer (20 mM Tris pH 7.5, 100 mM KCl, 5% (v/v) glycerol and 0.5 mM TCEP), 100 ng (5 nM) double stranded plasmid DNA (pUC19) containing a target site complementary to the guide region of the sgRNA (for time-course assays) or with some mismatches to the guide region (for off-target DNA cleavage assays), 10 mM divalent metal, 120 nM sgRNA, and 100 nM SpyCas9 incubated at 37°C for 15 min unless otherwise stated. The reaction products were resolved on a 1.1% agarose gel and the bands were visualized post-gel running using 0.1 µg/mL ethidium bromide.

#### 2.4.5 Quantification of DNA cleavage

##### 2.4.5.1 Guide RNA-free DNA cleavage rate constant calculation

The agarose gel images were analyzed using imageJ software (Schneider et al., 2012) to measure the band intensities of full-length M13mp18 single-stranded DNA (substrate), which includes both the circular and linear forms. The M13mp18 is degraded by SpyCas9 over time. For each lane, the intensity of circular M<sub>13mp18</sub> substrate remaining (C) was then normalized to that of the no protein control lane (C<sub>0</sub>) to obtain the fraction of DNA precursor remaining (P<sub>M13mp18</sub>).

$$P_{M13mp18} = \frac{C}{C_0} \quad (\text{Equation 2.1a})$$

Three replications were averaged, and the program Matlab (Natick, MA) was used to fit the following one-exponential decay equation to the P<sub>M13mp18</sub> to calculate the rate constant for guide RNA-free DNA degradation:

$$P_{M13mp18} = e^{-k_{\text{obs}}t} \quad (\text{Equation 2.1b})$$



Where  $t$  is time (min) and  $k_{obs}$  is the observed DNA degradation rate constant ( $\text{min}^{-1}$ ).

#### 2.4.5.2 gRNA-dependent DNA cleavage rate constant calculation

The agarose gel images were analyzed using imageJ software to measure the band intensities of the supercoiled (S), nicked (N), and linear (L) bands.

Normalized fraction supercoiled ( $P_{pUC19}$ ) was calculated for each time point as follows:

$$P_{pUC19} = \frac{S}{N + L + S} \div \left( \frac{S}{N + L + S} \right)_0 \quad (\text{Equation 2.2a})$$

Where  $\left( \frac{S}{N+L+S} \right)_0$  is the fraction of supercoiled in the control lane.

The program Matlab (Natick, MA) was used to fit the following two-exponential decay equation to the average  $P_{pUC19}$  from four replications to calculate the rate constant for gRNA-dependent DNA cleavage:

$$P_{pUC19} = \alpha_{fast} e^{-k_{fast}t} + \alpha_{slow} e^{-k_{slow}t} + (1 - \alpha_{fast} - \alpha_{slow}) \quad (\text{Equation 2.2b})$$

Where  $k_{fast}$  and  $k_{slow}$  are the observed DNA cleavage rate constants (units of  $\text{min}^{-1}$ ) for the biphasic reaction corresponding to the fast and slow phases of the cleavage reaction respectively and  $\alpha_{fast}$  and  $\alpha_{slow}$  are active fractions (the fraction of DNA that has been cleaved at each phase).

#### 2.4.5.3 Off-target analysis

The results of the off-target DNA cleavage assays were graphed using the average DNA linearization for three independent replications. Linear fraction was

obtained by dividing the intensity of the linear band (L) by the sum of the intensities of the nicked (N), L, and supercoiled (S) bands as follows:

$$\frac{L}{N + L + S} \quad (\text{Equation 2.3})$$

The no protein control linear fraction values are reported along with the experimental values to account for any spontaneous DNA linearization differences among the various mismatched substrates. In all DNA cleavage assays, standard error of the mean (SEM) of the DNA band intensity values was calculated to describe the variation among the replications (Babu et al., 2019).

#### 2.4.6 Setup of SpyCas9-ntDNA complex models for molecular dynamics (MD) simulations

##### 2.4.6.1 Initial structure of SpyCas9-ntDNA complex

Molecular dynamics (MD) simulations can be used to analyze the dynamic interactions of molecules to develop molecular mechanisms of complex systems (Palermo, 2019; Palermo et al., 2018b; Palermo et al., 2016, 2017a; Ricci et al., 2019; Zuo and Liu, 2016). SpyCas9+ntDNA models were used in MD simulations to achieve the following two goals: 1) decipher the molecular mechanism of SpyCas9's gRNA-free DNA cleavage and 2) unveil the possible mechanism of gRNA-free DNA cleavage resistance of SpyCas9<sup>H982A</sup>. Since the SpyCas9 RuvC active site is responsible for gRNA-free DNA cleavage (Sundaresan et al., 2017), the ideal initial structure for MD simulation of SpyCas9 poised for gRNA-free DNA cleavage is SpyCas9 complexed with ssDNA close to the RuvC catalytic center. We obtained such gRNA-free models (SpyCas9<sup>WT</sup>-ntDNA-Mg<sup>2+</sup>, SpyCas9<sup>WT</sup>-ntDNA-Mn<sup>2+</sup>, and SpyCas9<sup>H982A</sup>-ntDNA-Mn<sup>2+</sup>) for all-atom molecular simulations using a crystal structure tailoring strategy that has previously been applied for MD simulations of CRISPR-Cas9 (Palermo et al., 2016). This involved stripping sgRNA and tDNA from the structural ensemble of a catalytically active SpyCas9 complex that was developed for our previous study (full active state of

model B (Babu et al., 2021), which contains SpyCas9-sgRNA-tDNA-ntDNA derived from the experimental structure PDB code 5F9R (Jiang et al., 2016) and catalytic  $Mg^{2+}$  placed in a DNA cleavage competent state based on the positions of  $Mn^{2+}(A)$  and  $Mn^{2+}(B)$  in experimental structure PDB code: 4CMQ (Jinek et al., 2014). This stripping and catalytic ion placement tailoring strategy yielded SpyCas9<sup>WT</sup>-ntDNA- $Mg^{2+}$ , which models SpyCas9 in a gRNA-free environment.  $Mn^{2+}$  replacement was then used on SpyCas9<sup>WT</sup>-ntDNA- $Mg^{2+}$  to obtain SpyCas9<sup>WT</sup>-ntDNA- $Mn^{2+}$ , from which SpyCas9<sup>H982A</sup>-ntDNA- $Mn^{2+}$  was derived by substituting His982 to Ala.

#### 2.4.6.2 Setup of ion-perturbation model

Ion perturbation models were used to investigate the role of residue H982 in the recruitment of different metal ions to the catalytic center. In the RuvC center of SpyCas9<sup>WT</sup>-ntDNA- $Mg^{2+}$ , SpyCas9<sup>WT</sup>-ntDNA- $Mn^{2+}$  and SpyCas9<sup>H982A</sup>-ntDNA- $Mn^{2+}$ , there are two catalytic  $Mg^{2+}/Mn^{2+}$  ions. The one closer to H982 is  $Mg^{2+}_A/Mn^{2+}_A$ , and the other one is  $Mg^{2+}_B/Mn^{2+}_B$  based on the two-metal-ion model for SpyCas9's RuvC cleavage (Zuo and Liu, 2016). To investigate the role of residue 982 in the recruitment of  $Mg^{2+}_A/Mn^{2+}_A$  to the catalytic center, the original  $Mg^{2+}_A/Mn^{2+}_A$  was removed. Ten  $Mg^{2+}/Mn^{2+}$  ions were randomly placed around the RuvC enzymatic center by the CPPTRAJ program in Amber Tools (Roe and Cheatham, 2013). Among these 10  $Mg^{2+}/Mn^{2+}$  ions, the ion with the shortest distance to the originally removed  $Mg^{2+}_A/Mn^{2+}_A$  location was named as "perturbed  $Mg^{2+}_A/Mn^{2+}_A$ " for analysis in Figure 2.14.

#### 2.4.7 Procedure of MD simulations

##### 2.4.7.1 Preparations for all MD simulations

The above structures were solvated in truncated octahedral boxes of TIP3P waters with a minimal thickness of 13 Å from each edge using tLEaP of Ambergtools 16 (Case, 2016). The system was neutralized by  $Cl^-$ , and additional  $K^+$  and  $Cl^-$  were added to achieve the physiological ionic condition of 150 mM. Additional  $Mg^{2+}/Mn^{2+}$  ions were added to reach 5 mM of  $Mg^{2+}/Mn^{2+}$  concentration.

Adding the additional 10  $\text{Mg}^{2+}/\text{Mn}^{2+}$  ions in the ion-perturbation model increased the  $\text{Mg}^{2+}/\text{Mn}^{2+}$  concentration to 8.6 mM.

The system applied the amber force fields ff14SB (Maier et al., 2015) for protein, OL15 (Galindo-Murillo et al., 2016) for DNA, TIP3P (Jorgensen, 1983) for waters, the Joung-Cheatham parameter sets (Joung and Cheatham, 2008) for monovalent ions ( $\text{K}^+$ ,  $\text{Cl}^-$ ) and the optimized IOD-LJ parameter set (Li et al., 2013) for  $\text{Mg}^{2+}/\text{Mn}^{2+}$ . The CPPTRAJ program in Amber Tools (Roe and Cheatham, 2013) was utilized to randomize the positions of  $\text{K}^+$  and  $\text{Cl}^-$  such that the randomized ions were at least 4 Å away from any protein or DNA atom, eliminating the bias from monovalent ion-protein and monovalent ion-DNA interactions during the MD simulations.

#### 2.4.7.2 Protocols for all Molecular Dynamics simulations

All systems underwent a sequence of steps to reach equilibration as described in our previous work (Babu et al., 2021). First, the system was subjected to energy minimization, followed by slow heating from 0 K to the target of 310.15 K in an isothermal isochoric (NVT) ensemble, while the heavy atoms were restrained with the same force constant. 1 ns equilibration was performed in the isothermal-isobaric (NPT) condition with the weaker heavy atoms constrained (force constant for restraint: 1.0). Further,  $\text{Mg}^{2+}/\text{Mn}^{2+}$  ions in both HNH and RuvC domain received additional constraint in the equilibration process. Finally, the simulation was performed under the NPT ensemble without restraints, extending up to 500 ns. The pressure was controlled at 1.013 bar *via* the Monte Carlo barostat, and the temperature was maintained at 310.15 K through the Langevin thermostat implemented in AMBER16. The bonds involving hydrogens were constrained by the SHAKE algorithm (Kollman, 1992) with 2 fs integration steps during the heating, equilibration and production stages. Hydrogen mass repartitioning (Feenstra, 1999; Hopkins et al., 2015) was used to facilitate a 4 fs-time step in MD production. For each setting, three individual 500 ns MD simulations were conducted. The first 100 ns of each trajectory of all simulations

was considered as the equilibration and discarded based on RMSD analysis. All simulations were performed by the CUDA-accelerated version of AMBER16 pmemd engine (Salomon-Ferrer R, 2013).

#### 2.4.8 Protein sequence conservation analysis of SpyCas9's H982-H985 region

The search term “Cas9” was used in the NCBI nucleotide RefSeq database to get 33210 genomes with an annotated Cas9. These genome accession numbers were downloaded in default order, which is reverse chronological order of the date the genomes were deposited in the NCBI database (Romiti, 2006). One genome from each species was downloaded as a representative by submitting the first genome accession number for each species in the list to Batch Entrez (Romiti, 2006). The resulting 1467 representative genomes were submitted to CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci which allowed the genome accession numbers to be sorted by CRISPR-Cas subtype (Russel et al., 2020). The CDS of proteins was then collected for each genome accession number and Cas9 protein fastas were compiled resulting in 558 sequences for type II-A, 16 sequences for type II-B, and 461 sequences for type II-C for a total of 1035 Cas9s. To obtain comparable representation from each subtype, the number of subtype A and C sequences were reduced using cd-hit to cluster the proteins with at least 40% sequence identity for the resulting number of sequences: II-A: 119 clusters; II-C: 90 clusters (Li and Godzik, 2006). To further reduce the number of subtype A and C sequences, the representatives (the longest sequences) from the 14 largest clusters were taken. Clustal omega was used for multiple sequence alignment of the subtyped Cas9s (Sievers and Higgins, 2014), and iqtree was used to make molecular phylogeny using Maximum Likelihood (Kalyaanamoorthy et al., 2017; Minh et al., 2020). The tree was visualized using Iroki: automatic customization and visualization of phylogenetic trees (Moore et al., 2020). The MSA was visualized using the NCBI tool (<https://www.ncbi.nlm.nih.gov/projects/msaviewer/>).

#### 2.4.9 Statistical analyses

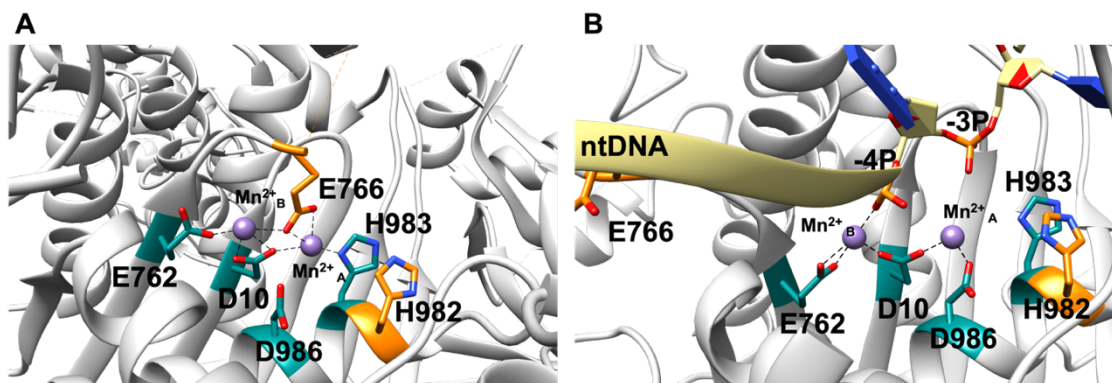
Three replications using two independent protein preparations were done for most assays comparing the effects of SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup>. For gRNA-dependent DNA cleavage time-course assays, four replications were included. Only one protein preparation each was used for assays with SpyCas9<sup>E766</sup> and SpyCas9<sup>E766/H982A</sup> since these variants were screened to identify a variant devoid of gRNA-free DNA cleavage (Figure 2.2). Equations and other kinetic fitting parameters used in this study are stated along with each experimental section. We used t-test to identify which conditions yielded statistically significant differences between averages shown in the gRNA-free all-metals assay and gRNA-dependent mismatch assay bar graphs (Figures 2.5B and 2.12B). Significance is indicated using one, two, or three asterisks meaning  $p \leq 0.05$ , 0.01, or 0.001 respectively.

#### 2.5.0 Results

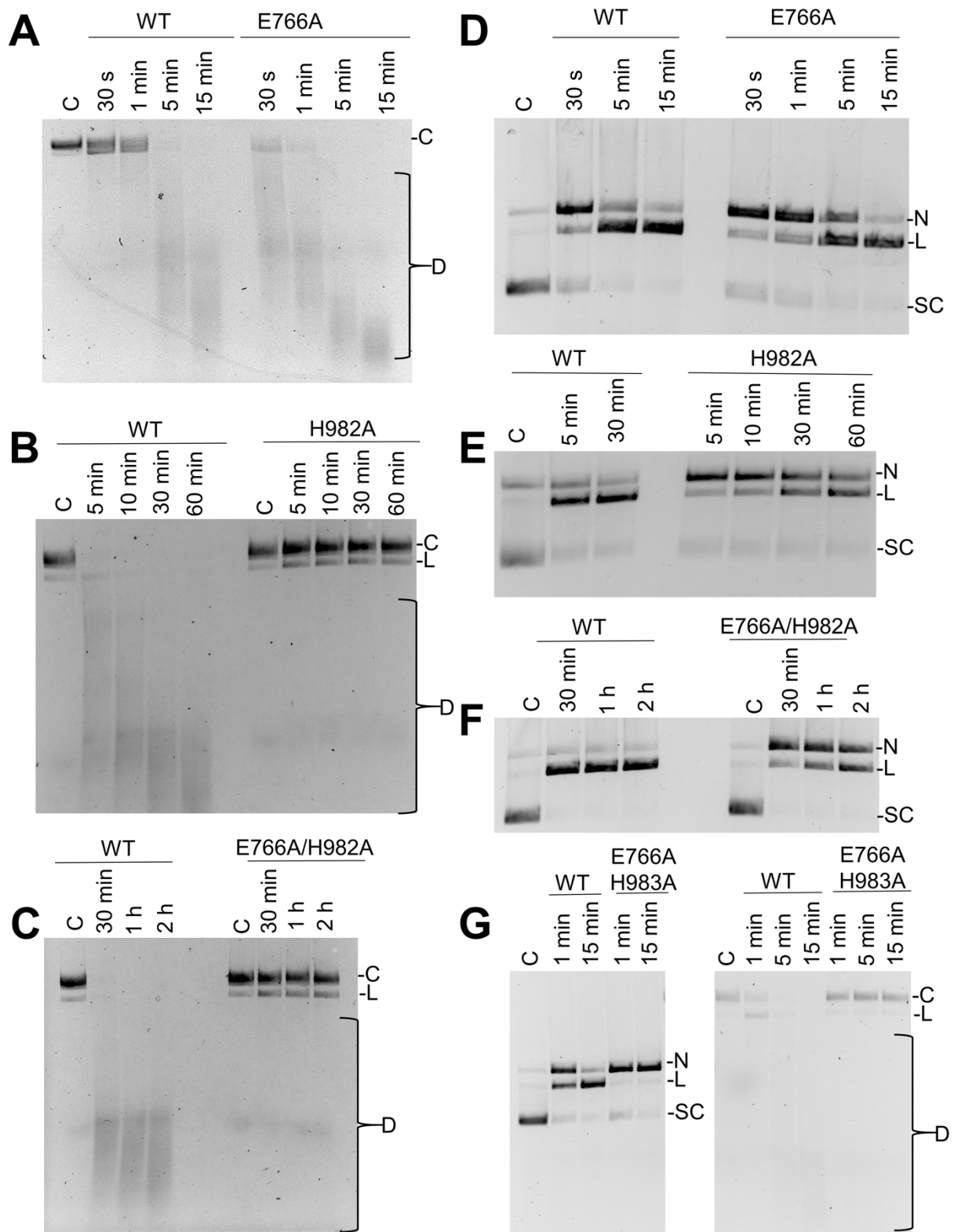
##### 2.5.1 Residues in the RuvC active site of SpyCas9 contributes differently towards guide RNA-free DNA cleavage

Previously we had shown that gRNA-free DNA cleavage is performed by the RuvC active in SpyCas9 (Sundaresan et al., 2017). We analyzed the orientation of amino acids, divalent metals and nucleic acids in SpyCas9's RuvC endonuclease domain to select residues that may contribute uniquely to gRNA-free DNA cleavage. Analysis of the RuvC active site in apoCas9 (PDB ID: 4CMQ (Jinek et al., 2014)) shows that essential residues for gRNA-dependent DNA cleavage (D10, E762, H983, and D986) and a non-essential residue E766 are involved in the octahedral coordination of the  $Mn^{2+}$  ions (Figure 2.1A). Analyzing the RuvC active site in the molecular dynamics structure of gRNA-free DNA cleavage, H982 is identified as the second shell coordinating residue for  $Mn^{2+}$ , with it being closer to  $Mn^{2+}(A)$  (Figure 2.1B). Our hypothesis is that even though E766 and H982 are not essential for gRNA-dependent DNA cleavage possibly due to favorable interactions from gRNA enabling DNA cleavage, active site structuring by these residues is necessary to facilitate gRNA-free DNA cleavage.

We created E766A and/or H982A variants (SpyCas9<sup>E766A</sup>, SpyCas9<sup>H982A</sup>, and SpyCas9<sup>E766/H982A</sup>) to test our hypothesis that these residues contribute to gRNA-free DNA cleavage without impacting gRNA-dependent DNA cleavage (Figure 2.2).



**Figure 2.1. RuvC nuclease active site.** (A) RuvC catalytic site in apo SpyCas9 (PDB ID: 4CMQ) (Jinek et al., 2014). Catalytic residues essential for gRNA-dependent DNA cleavage (D10, E762, H983, and D986 colored as dark cyan) are shown among non-essential residue E766 and H982 colored as orange. (B)  $Mn^{2+}$  coordination by the non-target DNA strand (ntDNA colored as khaki) and catalytic residues in the absence of gRNA. The catalytic residues are shown among non-essential residue E766 and H982. This figure is generated from MD simulation created structure after 500 ns simulation. The non-essential E766 is away, and  $Mn^{2+}$  coordinates to -4P ntDNA.

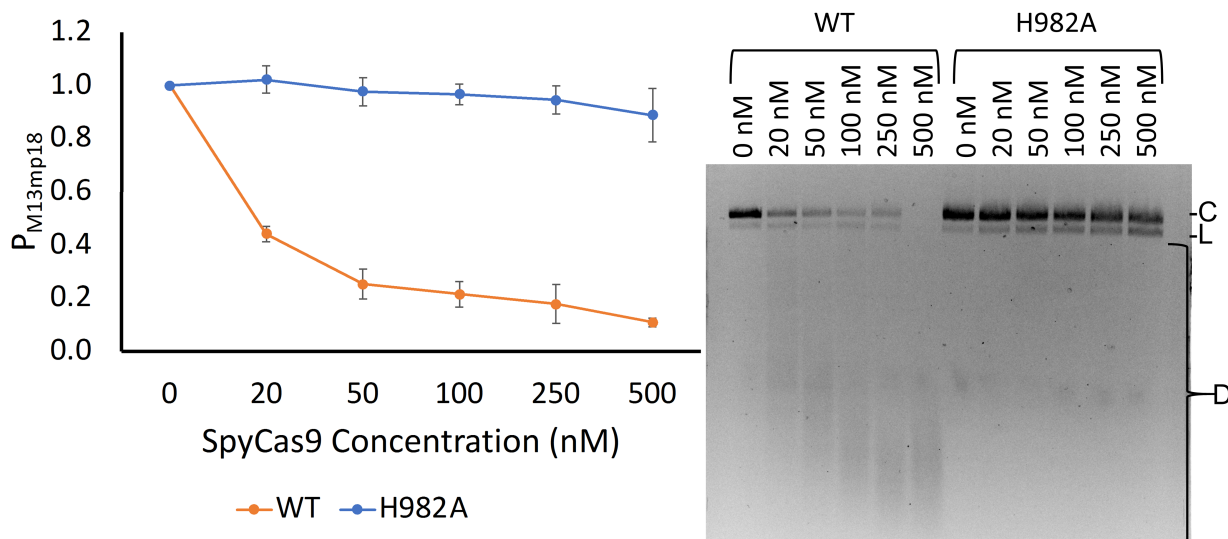


**Figure 2.2. DNA cleavage time-course assays for (A) SpyCas9<sup>WT</sup> and**



SpyCas9<sup>E766A</sup> gRNA-free cleavage of single stranded DNA plasmid (M13mp18), **(B)** SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> gRNA-free cleavage of M13mp18, **(C)** SpyCas9<sup>WT</sup> and SpyCas9<sup>E766A/H982A</sup> double variant gRNA-free cleavage of M13mp18, **(D)** SpyCas9<sup>WT</sup> and SpyCas9<sup>E766A</sup> gRNA-dependent cleavage of double stranded DNA plasmid (pUC19), **(E)** SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> gRNA-dependent cleavage of pUC19, **(F)** SpyCas9<sup>WT</sup> and SpyCas9<sup>E766A/H982A</sup> double variant gRNA-dependent cleavage of pUC19, and **(G)** SpyCas9<sup>E766/H983A</sup> gRNA-dependent cleavage of pUC19 (left) and gRNA-free cleavage of M13mp18 (right). Only one protein preparation each was done for SpyCas9<sup>E766/H983A</sup>, SpyCas9<sup>E766</sup> and SpyCas9<sup>E766/H982A</sup>. Protein concentration is 100 nM. gRNA-free DNA cleavage (panels A, B C, and G) was done with 10 mM Mn<sup>2+</sup>, and gRNA-dependent DNA cleavage (panels D, E, F, and G) was done with 10 mM Mg<sup>2+</sup>. [C – circular, D – degraded, N – nicked, L – Linear, SC – supercoiled.]

SpyCas9<sup>E766A</sup>, SpyCas9<sup>H982A</sup>, SpyCas9<sup>E766A/H982A</sup>, and SpyCas9<sup>E766A/H983A</sup> were assayed *in vitro* for gRNA-free and gRNA-dependent DNA cleavage activities (Figure 2.2). The data showed that SpyCas9<sup>H982A</sup> has negligible gRNA-free DNA cleavage activity even at 500 nM protein concentration (Figure 2.3); however, it is active for the desired gRNA-dependent DNA cleavage (Figure 2.2E). SpyCas9<sup>E766A</sup> does not remove the undesired gRNA-free cleavage activity (Figure 2.2A), and the double substitution variant SpyCas9<sup>E766A/H982</sup> does not noticeably decrease this activity relative to the single substitution variant SpyCas9<sup>H982A</sup> (Figures 2.2B and C), so the SpyCas9<sup>H982A</sup> variant was chosen for further in-depth analyses towards its role in gRNA-free DNA cleavage.

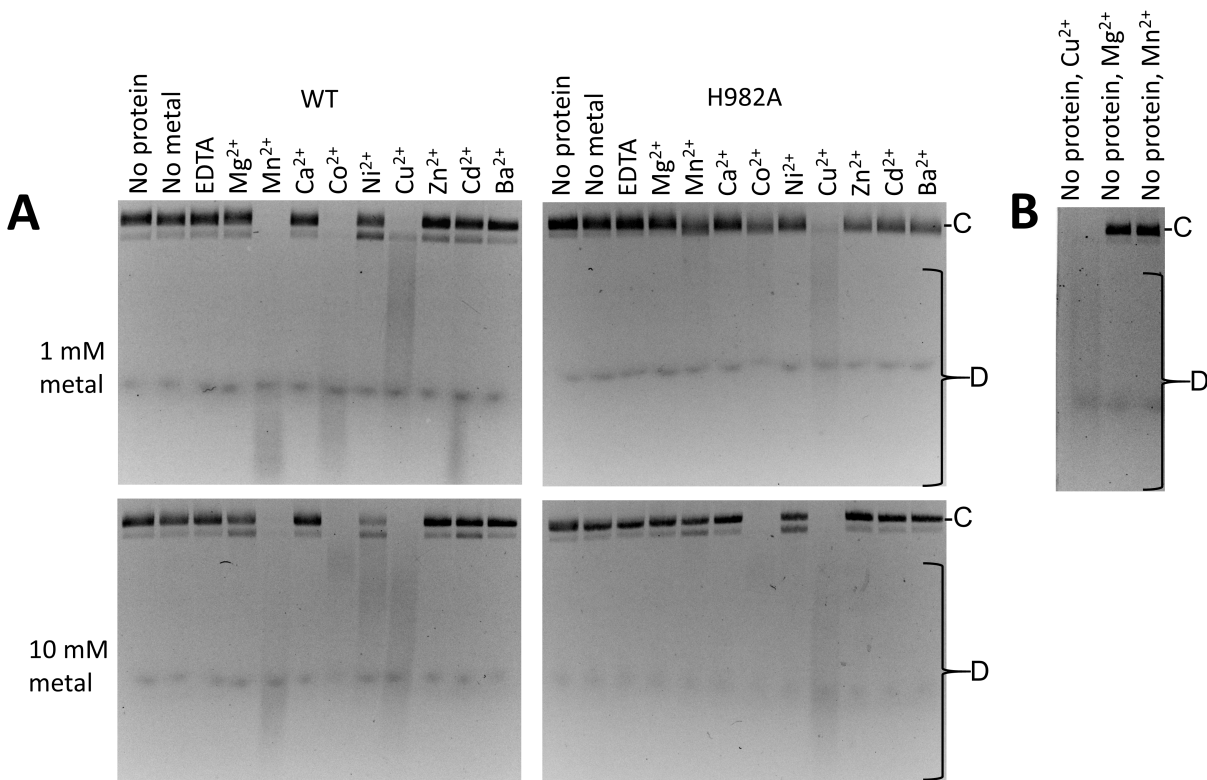


**Figure 2.3. SpyCas9 protein concentration titration assay of gRNA-free ssDNA cleavage with  $Mn^{2+}$ .** Reaction conditions were 10 mM  $Mn^{2+}$  and 100 ng M13mp18 ssDNA plasmid incubated for 30 min at 37°C.  $P_{M13mp18}$  is the fraction of uncleaved DNA (Equation 2.1a). Error bars are the standard error of the mean (SEM) of three replications. A representative gel is shown to the right. [C – circular, L – linear, D – degraded.]

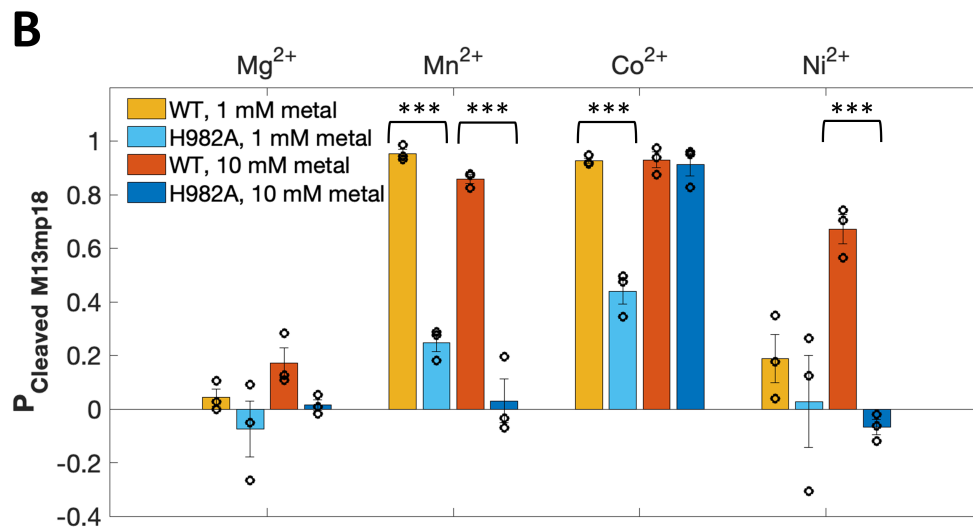
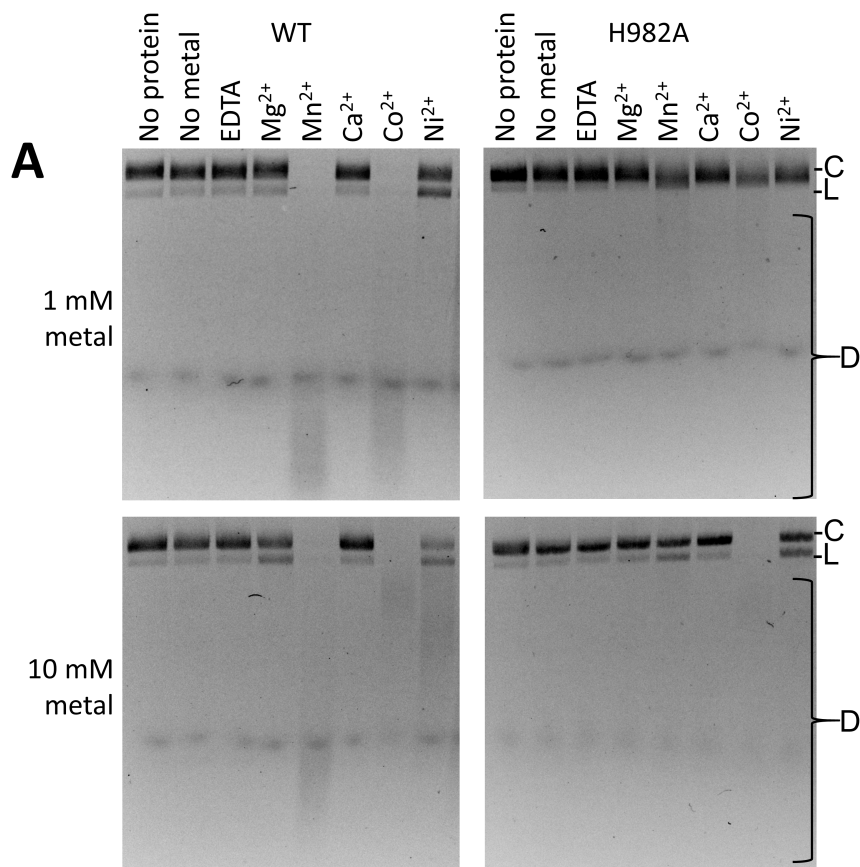
### 2.5.2 His982 of SpyCas9's RuvC active site is a major enhancer of gRNA-free DNA cleavage

To analyze the effect of H982 on gRNA-free DNA cleavage, we tested gRNA-free DNA cleavage with different divalent metal ions at two different concentrations (Figures 2.4 and 2.5). Results from a two-sided t-test showed that the H982A substitution significantly impairs the ability of SpyCas9 to catalyze gRNA-free DNA cleavage with 1 mM and 10 mM  $Mn^{2+}$ , 1 mM  $Co^{2+}$ , and 10 mM  $Ni^{2+}$ . In SpyCas9<sup>H982A</sup>, gRNA-free DNA cleavage is reduced by 74% and 96% for 1 mM and 10 mM  $Mn^{2+}$  respectively, 53% for 1 mM  $Co^{2+}$ , and 100% for 10 mM  $Ni^{2+}$  (Figure 2.5B). Divalent metal titration assays with  $Mn^{2+}$  and  $Co^{2+}$  showed that

gRNA-free DNA cleavage is concentration dependent and that the H982A substitution inhibits gRNA-free DNA cleavage at higher metal concentrations compared to SpyCas9<sup>WT</sup>, where lower metal concentrations can promote efficient DNA cleavage (Figure 2.6). This effect is more pronounced with Mn<sup>2+</sup> than with Co<sup>2+</sup>. SpyCas9<sup>WT</sup> can completely degrade all the substrate with 0.025 mM Mn<sup>2+</sup>, while a significant amount of the substrate is left uncleaved at 10 mM Mn<sup>2+</sup> in the case of SpyCas9<sup>H982A</sup>. For Co<sup>2+</sup>, SpyCas9<sup>WT</sup> completely degraded all the substrate at a concentration of 0.1 mM, while SpyCas9<sup>H982A</sup> required 0.5 mM Co<sup>2+</sup> to achieve this (Figure 2.6). The results indicate that based on the sidechain substitution and the type of divalent metal, SpyCas9 RuvC varies in its ability to catalyze gRNA-free DNA cleavage.

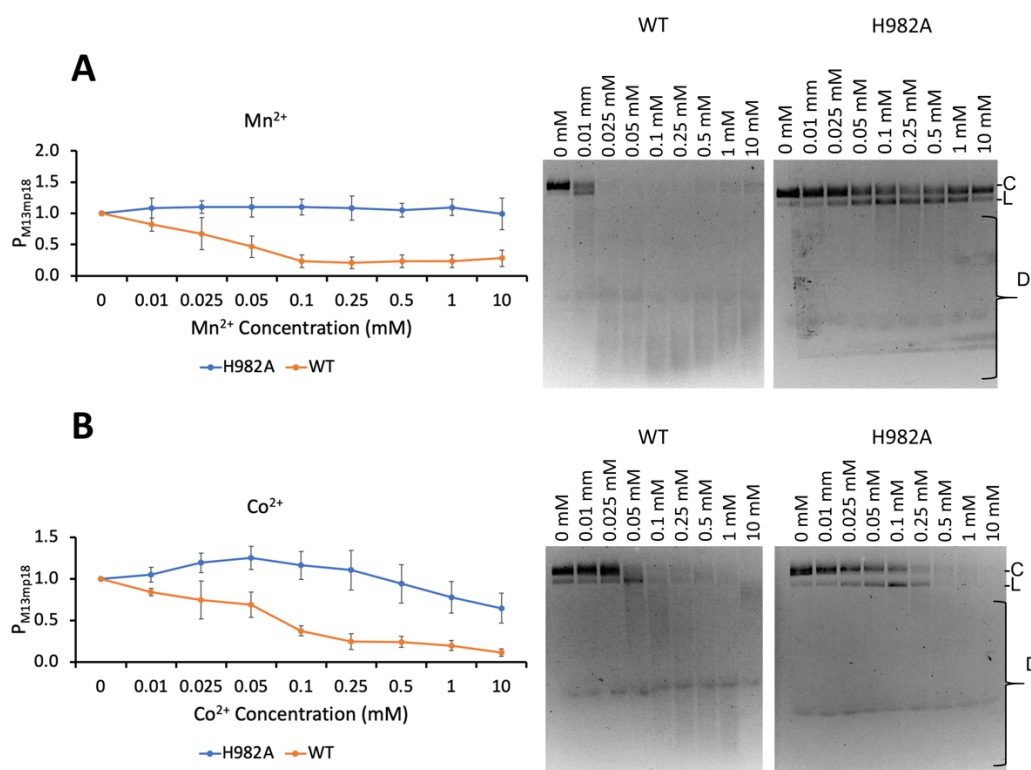


**Figure 2.4. Analysis of the effect of different divalent metals on gRNA-free DNA cleavage. (A)** Uncropped agarose gels (cropped gels are in Figure 2.5A) quantified to construct the bar graph in panel B of Figure 2.5. Only  $\text{Mg}^{2+}$ ,  $\text{Mn}^{2+}$ ,  $\text{Co}^{2+}$ ,  $\text{Ni}^{2+}$  are shown in Figure 2.5, but the gels shown in this figure include additional metals:  $\text{Cu}^{2+}$ , which was found to degrade M13mp18 ssDNA without protein (data shown in panel 2.4B), and  $\text{Ca}^{2+}$ ,  $\text{Zn}^{2+}$ ,  $\text{Cd}^{2+}$ , and  $\text{Ba}^{2+}$ , which were not found to facilitate gRNA-free DNA cleavage. **(B).** Protein-independent degradation of M13mp18 in the presence of  $\text{Cu}^{2+}$ . [C – circular, L – linear, D – degraded.]



**Figure 2.5. Divalent metal dependence of gRNA-free DNA cleavage. (A)** Representative agarose gels that were quantified to construct the bar graph in panel B. Figure 2.4 has the uncropped gels and shows a wider variety of metals.

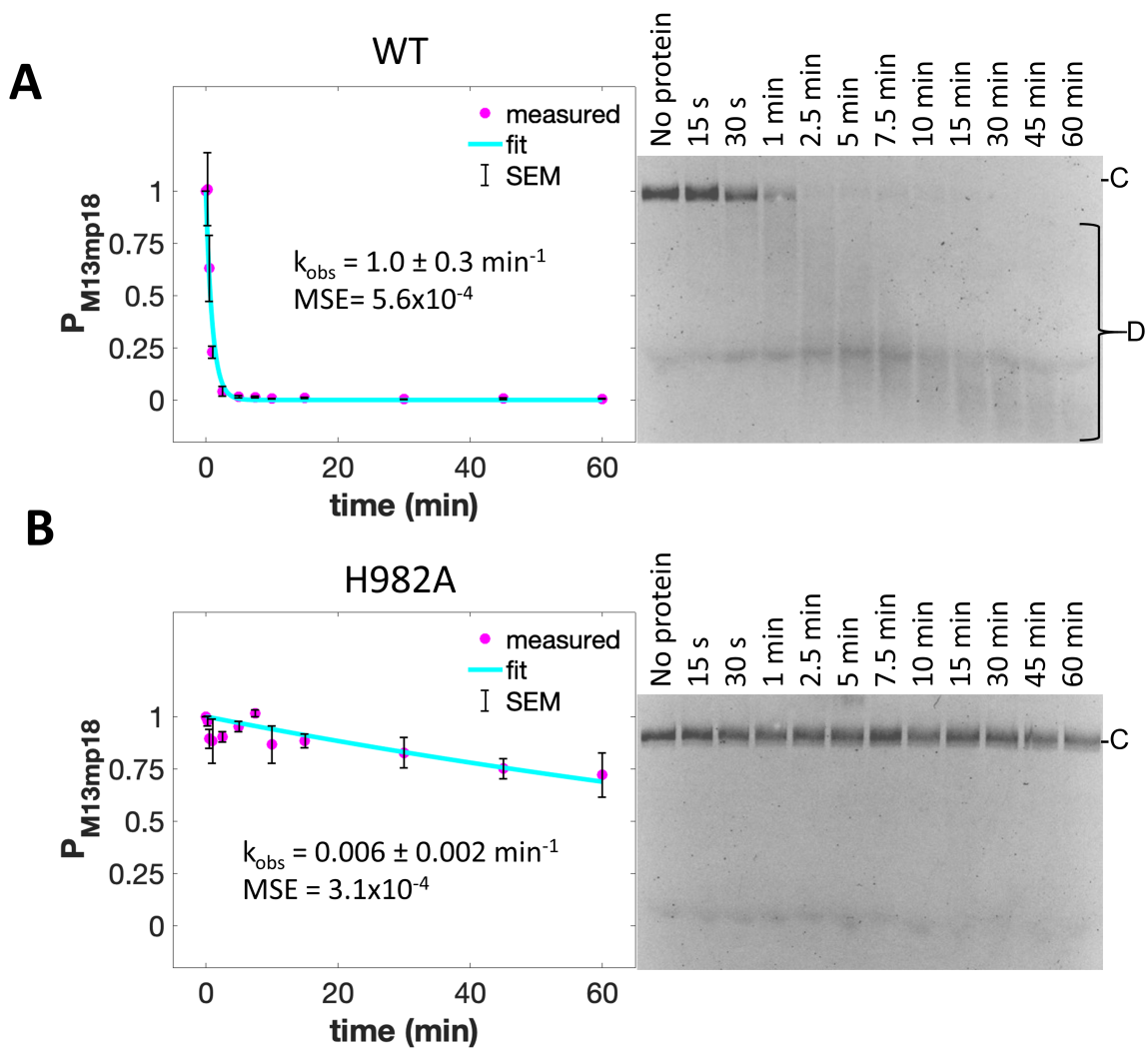
**(B)** The fraction of cleaved single stranded circular M13mp18 DNA ( $1-P_{M13mp18}$  from Equation 2.1a,  $P_{\text{cleaved M13mp18}}$ ) is plotted for four conditions: SpyCas9<sup>WT</sup> (WT) and SpyCas9<sup>H982A</sup> (H982A) protein, and 1 mM and 10 mM divalent metal. As observed, Mg<sup>2+</sup> was expected to have very little gRNA-free DNA cleavage with SpyCas9<sup>WT</sup>. Also as observed, Mn<sup>2+</sup>, Co<sup>2+</sup>, and Ni<sup>2+</sup> were expected to facilitate significant gRNA-free DNA cleavage in SpyCas9<sup>WT</sup>. The error bars plotted are standard error of the mean (SEM) of three replications shown as black circles. According to the two-sample t-test, there is a significant difference between the SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> for 1 mM and 10 mM Mn<sup>2+</sup>, 1 mM Co<sup>2+</sup>, and 10 mM Ni<sup>2+</sup>. One, two, or three asterisks means  $p \leq 0.05$ , 0.01, or 0.001 respectively. [C – circular, L – linear, D – degraded.]



**Figure 2.6. Divalent metal titration assay of SpyCas9 gRNA-free ssDNA cleavage. (A) Mn<sup>2+</sup> and (B) Co<sup>2+</sup>.** Reaction conditions were 100 nM SpyCas9

and 100 ng M13mp18 ssDNA plasmid incubated for 30 min at 37°C.  $P_{M13mp18}$  is the fraction of uncleaved DNA (Equation 2.1a). Error bars are the standard error of the mean (SEM) of three replications. Representative gels are shown to the right. [C – circular, L – linear, D – degraded.]

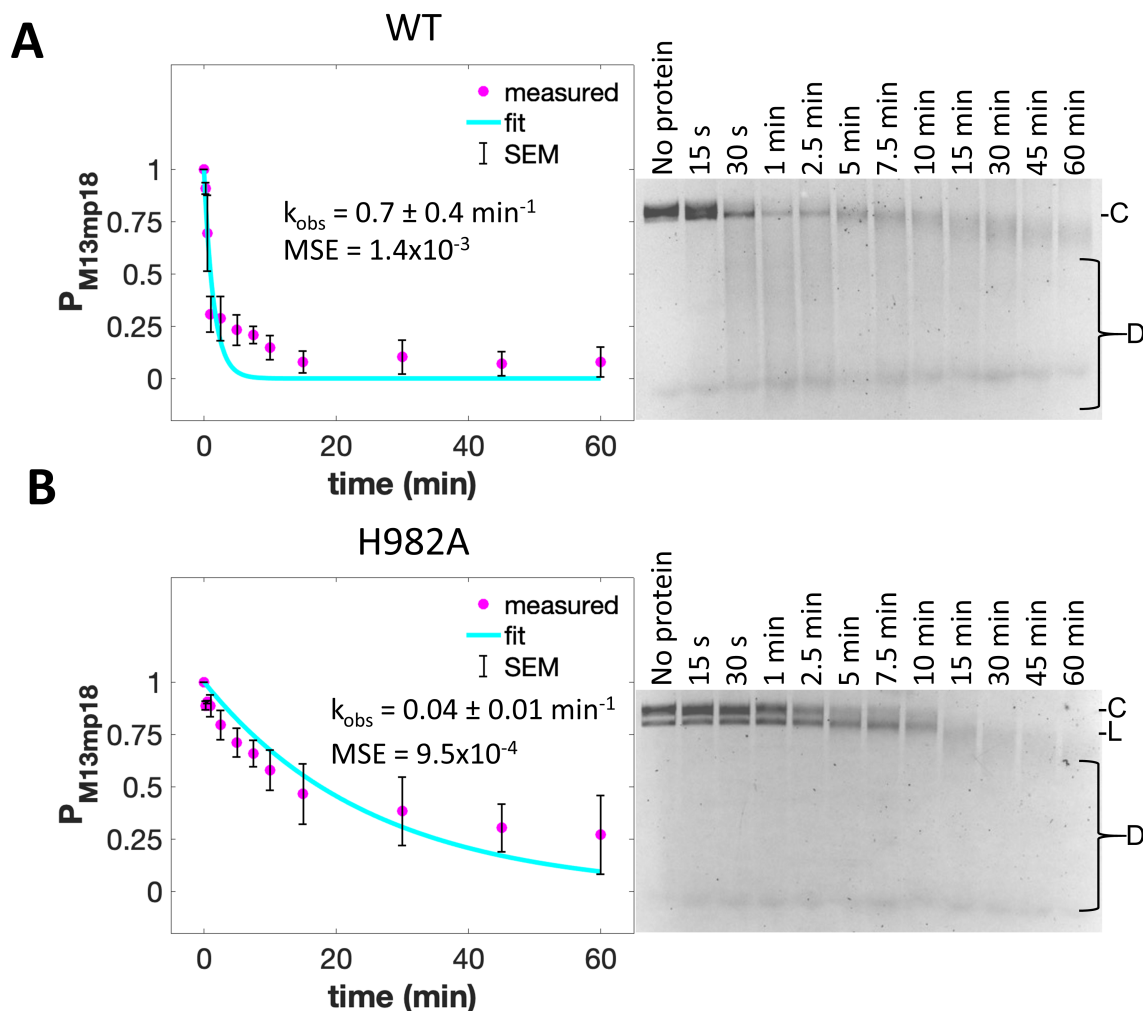
To quantify the inhibition of gRNA-free DNA cleavage by the H982A substitution, we calculated rate constants from single exponential fitting ( $k_{obs}$ , Equation 2.1b, Figures 2.7 and 2.8) of data (Equation 2.1a) obtained from time course assays. Briefly, we calculated the disappearance of the circular single stranded M13mp18 DNA substrate at each time point compared to the control reaction where no protein was added. The rate constants for the different conditions are summarized in Figure 2.9A and Table 2.2.



**Figure 2.7. Determination of rate constants ( $k_{obs}$ ,  $\text{min}^{-1}$ ) and representative gels for gRNA-free cleavage of circular, single stranded M13mp18 DNA with  $\text{Mn}^{2+}$ . Single exponential fit (cyan line, Equation 2.1b) of average fraction of uncleaved DNA (Equation 2.1a,  $P_{M13mp18}$ ) plotted against time ( $t$ , min) for (A) SpyCas9<sup>WT</sup>- $\text{Mn}^{2+}$  and (B) SpyCas9<sup>H982A</sup>- $\text{Mn}^{2+}$ . The data points (magenta) are the average of three replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from**



the single exponential fit line. The error reported for  $k_{\text{obs}}$  is the 95% confidence interval determined with the Matlab fitting program. [C – circular, D – degraded.]

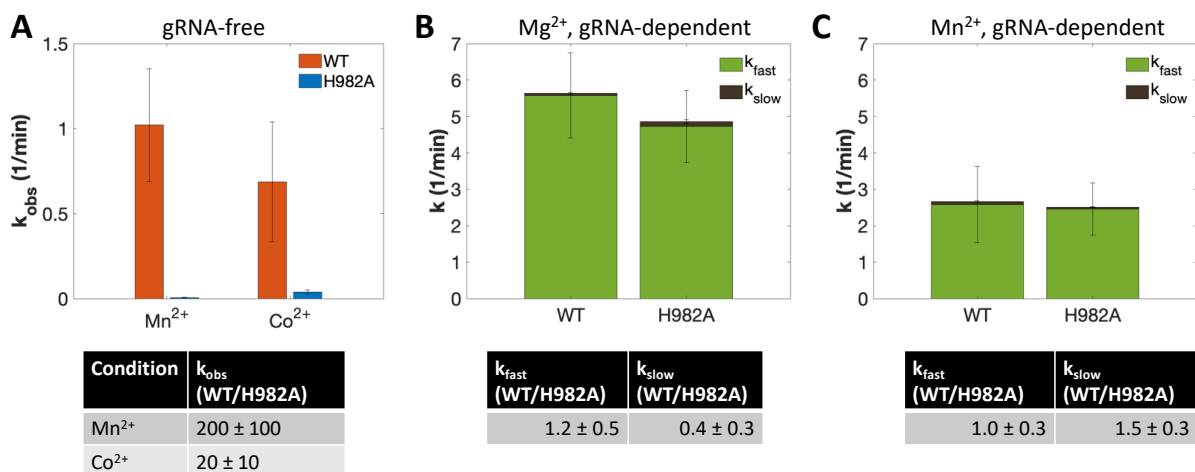


**Figure 2.8. Determination of rate constants ( $k_{\text{obs}}$ ,  $\text{min}^{-1}$ ) and representative gels for gRNA-free cleavage of circular, single stranded M13mp18 DNA with  $\text{Co}^{2+}$ . Single exponential fit (cyan line, Equation 2.1b) of average fraction of uncleaved DNA (Equation 2.1a,  $P_{\text{M13mp18}}$ ) plotted against time ( $t$ , min) for (A) SpyCas9<sup>WT</sup>- $\text{Co}^{2+}$  and (B) SpyCas9<sup>H982A</sup>- $\text{Co}^{2+}$ . The data points (magenta) are the average of three replications with standard error of the mean (SEM) shown as**

error bars. Mean square error (MSE) represents deviation of data points from the single exponential fit line. The error reported for  $k_{\text{obs}}$  is the 95% confidence interval determined with the Matlab fitting program. [C – circular, L – linear, D – degraded.]

|                                            | $k_{\text{obs}}$ ( $\text{min}^{-1}$ ) |
|--------------------------------------------|----------------------------------------|
| SpyCas9 <sup>WT</sup> -Mn <sup>2+</sup>    | $1.0 \pm 0.3$                          |
| SpyCas9 <sup>H982A</sup> -Mn <sup>2+</sup> | $0.006 \pm 0.002$                      |
| SpyCas9 <sup>WT</sup> -Co <sup>2+</sup>    | $0.7 \pm 0.4$                          |
| SpyCas9 <sup>H982A</sup> -Co <sup>2+</sup> | $0.04 \pm 0.01$                        |

**Table 2.2.** Rate constants for gRNA-free DNA cleavage from single exponential fitting ( $k_{\text{obs}}$ , Equation 2.1b) as shown in Figures 2.7 and 2.8.



**Figure 2.9 Comparison of SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> DNA cleavage rates.** (A) Bar graph of gRNA-free DNA degradation rate constants ( $k_{\text{obs}}$ ) and table of fold change (WT/H982A) in  $k_{\text{obs}}$ . The gRNA-free DNA cleavage assays were done using 10 mM metal (Mn<sup>2+</sup> or Co<sup>2+</sup>). (B) Bar graph of Mg<sup>2+</sup>-mediated gRNA-dependent DNA degradation rates from both fast ( $k_{\text{fast}}$ ) and slow ( $k_{\text{slow}}$ )

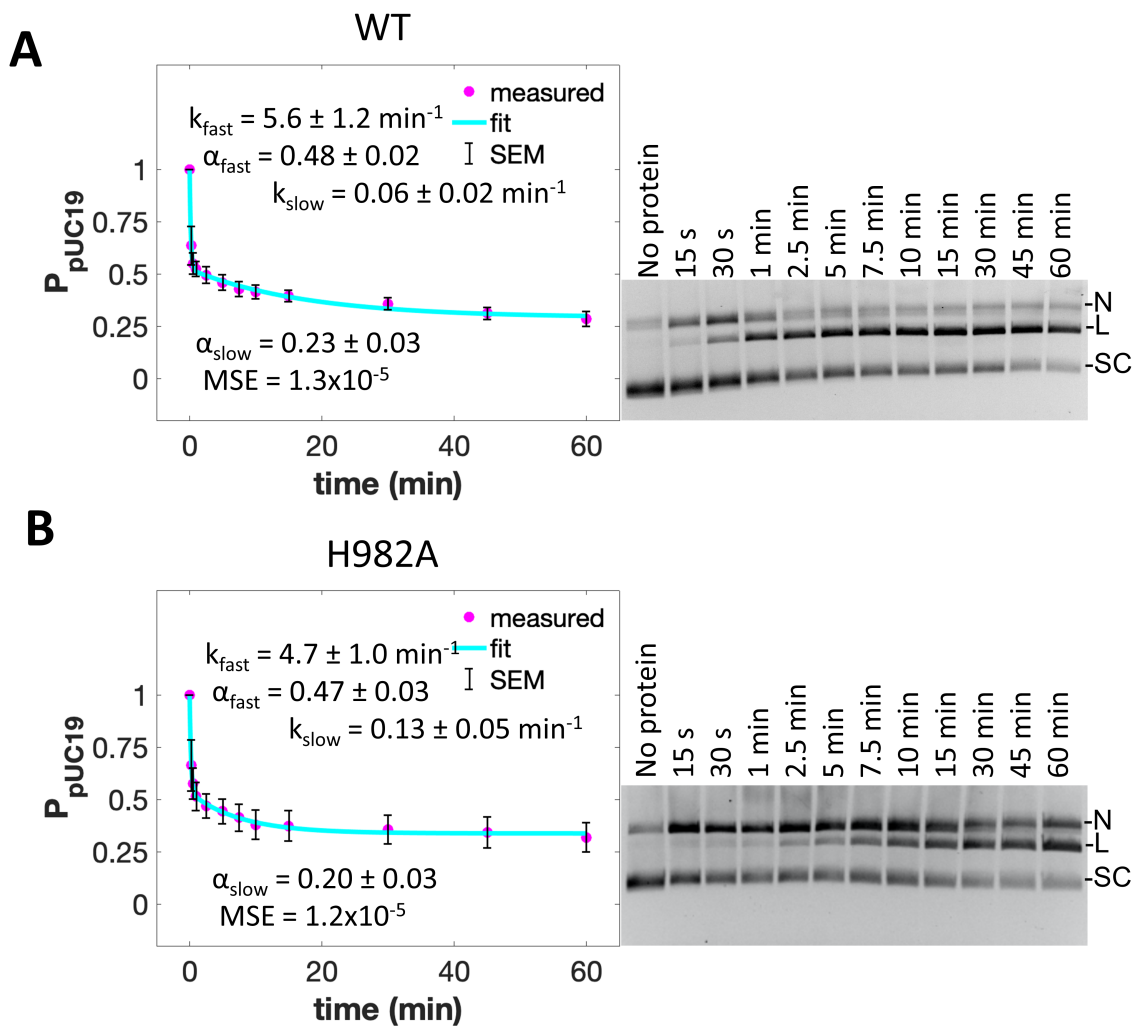
phases of the reaction. Below is a table of fold change (WT/H982A) in  $k_{\text{fast}}$  and  $k_{\text{slow}}$ . **(C)** Bar graph of  $\text{Mn}^{2+}$ -mediated gRNA-dependent DNA degradation rates from both fast ( $k_{\text{fast}}$ ) and slow ( $k_{\text{slow}}$ ) phases of the reaction. Below is a table of fold change (WT/H982A) in  $k_{\text{fast}}$  and  $k_{\text{slow}}$ . The gRNA-dependent DNA cleavage reactions were done using 10 mM metal. Error bars represent the 95% confidence interval determined with the Matlab fitting program.

The H982A substitution reduced gRNA-free DNA cleavage activity ~167-fold ( $200 \pm 100$ -fold with  $\pm 100$  being the propagated error from the 95% confidence intervals calculated while determining the rate constants) and ~17-fold ( $20 \pm 10$ -fold) compared to SpyCas9<sup>WT</sup> with  $\text{Mn}^{2+}$  and  $\text{Co}^{2+}$  respectively (Figures 2.7, 2.8, 2.9A and Table 2.2). While both  $\text{Mn}^{2+}$  and  $\text{Co}^{2+}$  can promote gRNA-free DNA cleavage at a similar efficiency in SpyCas9<sup>WT</sup> ( $\sim 1 \text{ min}^{-1}$  and  $0.7 \text{ min}^{-1}$  respectively), there is an ~7-fold ( $\sim 0.006 \text{ min}^{-1}$  *versus*  $\sim 0.04 \text{ min}^{-1}$ ) difference in how  $\text{Mn}^{2+}$  and  $\text{Co}^{2+}$  can organize the DNA cleavage competent state in SpyCas9<sup>H982A</sup> (Table 2.2). These strongly support our hypothesis that RuvC residues and their interactions with divalent metal ions play a significant role in catalyzing gRNA-free DNA cleavage.

### 2.5.3 Substitution of His982 does not severely impair guide gRNA-dependent DNA cleavage

As mentioned previously, we proposed that while H982 plays a significant role in organizing the DNA cleavage competent state of RuvC in the absence of gRNA, its role in DNA cleavage in the presence of gRNA should not be critical. A SpyCas9 variant possessing these properties is essential for gene editing applications. We proceeded to testing the effect of H982A substitution in gRNA-dependent DNA cleavage (Figures 2.10 and 2.11). We performed time course gRNA-dependent DNA cleavage assays using 10 mM  $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$ , ~5 nM substrate, 100 nM SpyCas9 and 120 nM sgRNA. The results show a comparable use of substrate (supercoiled DNA) by both SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup>, even though there is a slightly higher accumulation of nicked DNA in the  $\text{Mg}^{2+}$

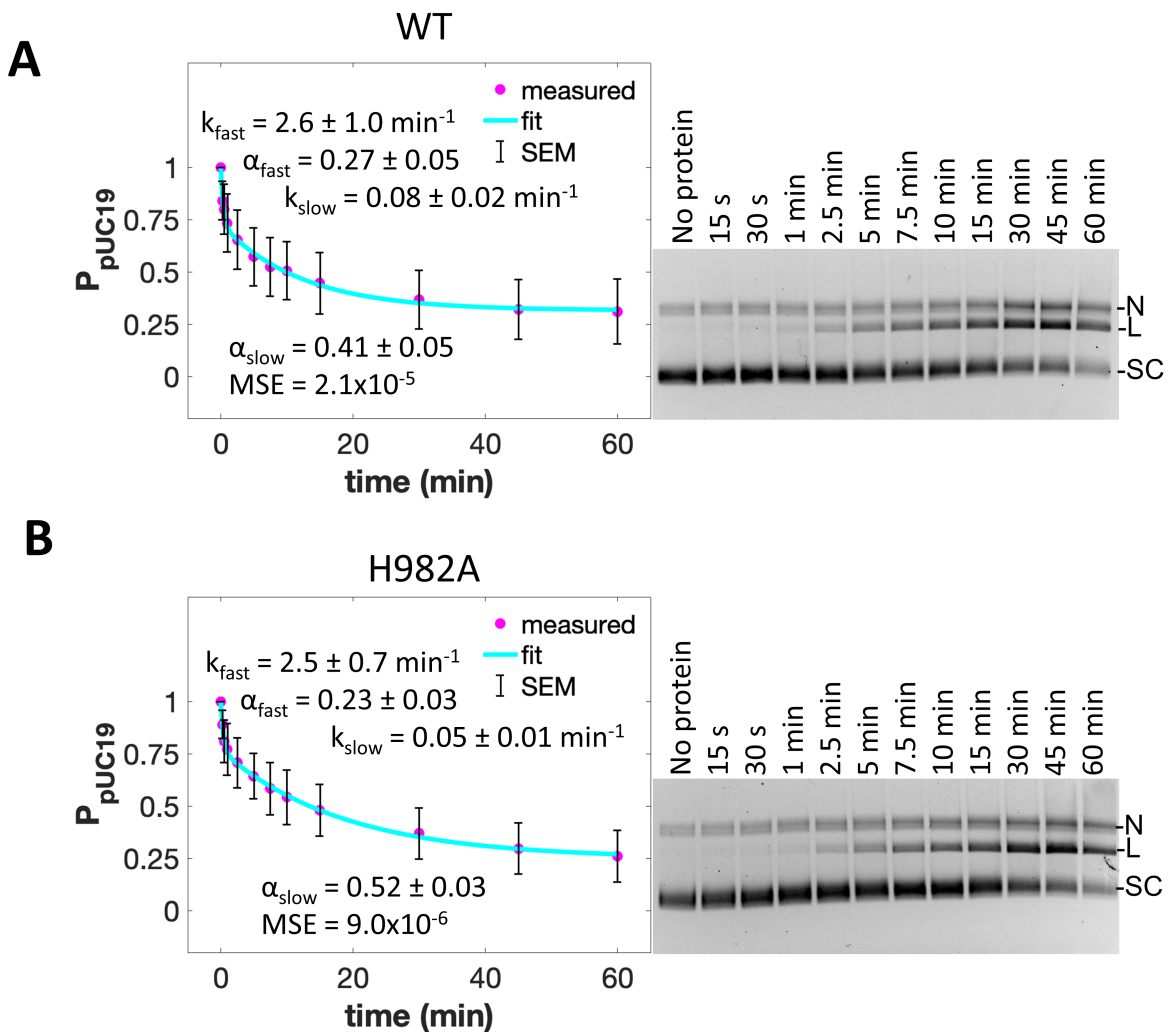
condition for SpyCas9<sup>H982A</sup> relative to SpyCas9<sup>WT</sup> (Figure 2.10). The data from the time-course assays (Equation 2.2a) were used to calculate rate constants for gRNA-dependent DNA cleavage using a double exponential fitting (Equation 2.2b). This gave two rate constants ( $k_{\text{fast}}$  and  $k_{\text{slow}}$  for the early and late phases of the reaction respectively) and two active fractions ( $\alpha_{\text{fast}}$  and  $\alpha_{\text{slow}}$ ) representing the fraction of DNA substrate being cleaved during the fast and slow phases. We see that the  $k_{\text{fast}}$  and  $\alpha_{\text{fast}}$  for both proteins are similar, with SpyCas9<sup>H982A</sup> having a slightly faster  $k_{\text{slow}}$  value (Table 2.3). There is a  $1.2 \pm 0.5$ -fold reduction in fast reaction rate constant in SpyCas9<sup>H982A</sup> compared to SpyCas9<sup>WT</sup> (Figure 2.9B), which is an insignificant reduction compared to the  $\sim 167$ -fold reduction in gRNA-free DNA cleavage. Fitting of the cleavage data using a single exponential equation also revealed a similar trend where both SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> demonstrated similar efficiencies in gRNA-dependent DNA cleavage (data not shown).



**Figure 2.10. Determination of rate constants for SpyCas9 gRNA-dependent cleavage of supercoiled, double stranded DNA in the presence of  $Mg^{2+}$ .** The graph shows the double exponential fit (cyan line, Equation 2.2.3b) of average fraction of uncleaved DNA (Equation 3a,  $P_{pUC19}$ ) plotted against time ( $t$ , min) for (A) SpyCas9<sup>WT</sup>- $Mg^{2+}$  and (B) SpyCas9<sup>H982A</sup>- $Mg^{2+}$ . Active fraction ( $\alpha_{fast}$  and  $\alpha_{slow}$ ) is the fraction of DNA that has been cleaved. The data points (magenta) are the average of four replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from the double exponential fit line. The error reported for  $k_{fast}$ ,  $k_{slow}$ ,  $\alpha_{fast}$ , and  $\alpha_{slow}$  is the

95% confidence interval determined with the Matlab fitting program. [N – nicked, L – Linear, SC – supercoiled.]

Since we saw differential effect for  $Mg^{2+}$  and  $Mn^{2+}$  in gRNA-free DNA cleavage, we tested the effect of  $Mn^{2+}$  on catalyzing gRNA-dependent DNA cleavage (Figure 2.11). Specifically, we wanted to analyze if the enhancement of gRNA-free DNA cleavage by  $Mn^{2+}$  translates in some fashion to gRNA-dependent DNA cleavage. We observed similar efficiencies in cleaving supercoiled DNA for both the proteins, with SpyCas9<sup>WT</sup> accumulating nicked products with  $Mn^{2+}$  compared to that with  $Mg^{2+}$ , indicating that  $Mg^{2+}$  is a preferred catalyst for the gRNA-dependent DNA cleavage (Figures 2.10 and 2.11). Interestingly, we also saw that the rate constants for the reaction in the presence of  $Mn^{2+}$  reduced for both the proteins ( $k_{fast}$  of  $\sim 5\text{-}6\text{ min}^{-1}$  for  $Mg^{2+}$  *versus*  $\sim 3\text{ min}^{-1}$  for  $Mn^{2+}$ ; Tables 2) supporting that  $Mg^{2+}$  is more favorable for gRNA-dependent DNA cleavage.



**Figure 2.11. Determination of rate constants for SpyCas9 gRNA-dependent cleavage of supercoiled, double stranded DNA in the presence of  $Mn^{2+}$ .**

Figure 2.10 shows  $Mg^{2+}$ -facilitated RNA guided DNA cleavage, while this figure shows results for  $Mn^{2+}$ -facilitated gRNA-dependent DNA cleavage. The graph shows the double exponential fit (cyan line, Equation 2.2b) of average fraction of uncleaved DNA (Equation 2.2a,  $P_{pUC19}$ ) plotted against time ( $t$ , min) for (A) SpyCas9<sup>WT</sup>- $Mn^{2+}$  and (B) SpyCas9<sup>H982A</sup>- $Mn^{2+}$ . Active fraction ( $\alpha_{fast}$  and  $\alpha_{slow}$ ) is the fraction of DNA that has been cleaved. The data points (magenta) are the average of four replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from the

double exponential fit line. The error reported for  $k_{\text{fast}}$ ,  $k_{\text{slow}}$ ,  $\alpha_{\text{fast}}$ , and  $\alpha_{\text{slow}}$  is the 95% confidence interval determined with the Matlab fitting program. [N – nicked, L – Linear, SC – supercoiled.]

|                                            | $k_{\text{fast}}$ ( $\text{min}^{-1}$ ) | $\alpha_{\text{fast}}$ | $k_{\text{slow}}$ ( $\text{min}^{-1}$ ) | $\alpha_{\text{slow}}$ |
|--------------------------------------------|-----------------------------------------|------------------------|-----------------------------------------|------------------------|
| SpyCas9 <sup>WT</sup> -Mg <sup>2+</sup>    | $5.6 \pm 1.2$                           | $0.48 \pm 0.02$        | $0.06 \pm 0.02$                         | $0.23 \pm 0.03$        |
| SpyCas9 <sup>H982A</sup> -Mg <sup>2+</sup> | $4.7 \pm 1.0$                           | $0.47 \pm 0.03$        | $0.13 \pm 0.05$                         | $0.20 \pm 0.03$        |
| SpyCas9 <sup>WT</sup> -Mn <sup>2+</sup>    | $2.6 \pm 1.0$                           | $0.27 \pm 0.05$        | $0.08 \pm 0.02$                         | $0.41 \pm 0.05$        |
| SpyCas9 <sup>H982A</sup> -Mn <sup>2+</sup> | $2.5 \pm 0.7$                           | $0.23 \pm 0.03$        | $0.05 \pm 0.01$                         | $0.52 \pm 0.03$        |

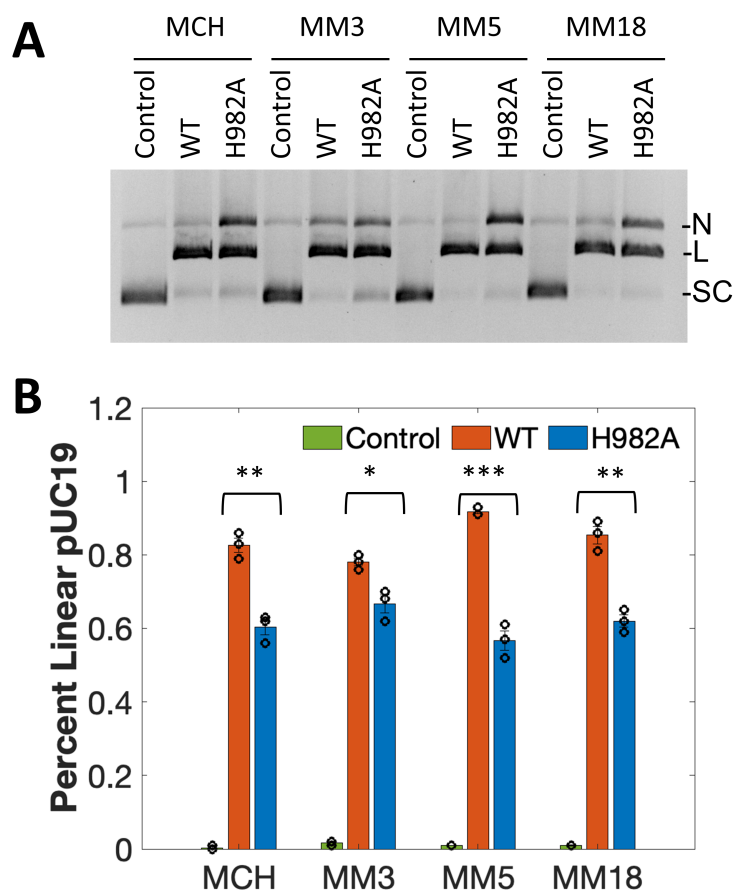
**Table 2.3.** Rate constants and active fraction values for Mg<sup>2+</sup> and Mn<sup>2+</sup>-mediated gRNA-dependent DNA cleavage from the fast ( $k_{\text{fast}}$  and  $\alpha_{\text{fast}}$  respectively) and slow ( $k_{\text{slow}}$  and  $\alpha_{\text{slow}}$  respectively) phases of the reaction based on the double exponential fitting (Equation 2.2b) shown in Figures 2.10 and 2.11.

Another observation from our results is that the rate constants for SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup>, with both with Mg<sup>2+</sup> and Mn<sup>2+</sup>, are comparable indicating that H982 coordination of the divalent metal ion does not contribute significantly towards the gRNA-dependent DNA cleavage process (Figures 2.9B and C, 2.10, 2.11, and Table 2.3). This supports our initial hypothesis that active site residues contribute differentially towards gRNA-free and gRNA-dependent DNA cleavages. The variations in the observed  $k$  and  $\alpha$  values when comparing Mg<sup>2+</sup> and Mn<sup>2+</sup> for the same protein sets (Figures 2.9B and C, 2.10, 2.11, and Table 2.3) indicate that the mechanisms by which these ions promote gRNA-dependent DNA cleavage differ. Delineating these mechanisms will require further in-depth studies. Nevertheless, in relation to the current study, the H982A substitution exhibits similar rate constants ( $k_{\text{fast}}$  and  $k_{\text{slow}}$ ) and active fractions ( $\alpha_{\text{fast}}$  and  $\alpha_{\text{slow}}$ ) for gRNA-dependent DNA cleavage in the presence of both Mg<sup>2+</sup> and Mn<sup>2+</sup> indicating that H982A substitution reduces gRNA-free DNA cleavage without negatively impacting gRNA-dependent DNA cleavage.



#### 2.5.4 Substitution of His982 enhances DNA mismatch sensitivity of SpyCas9

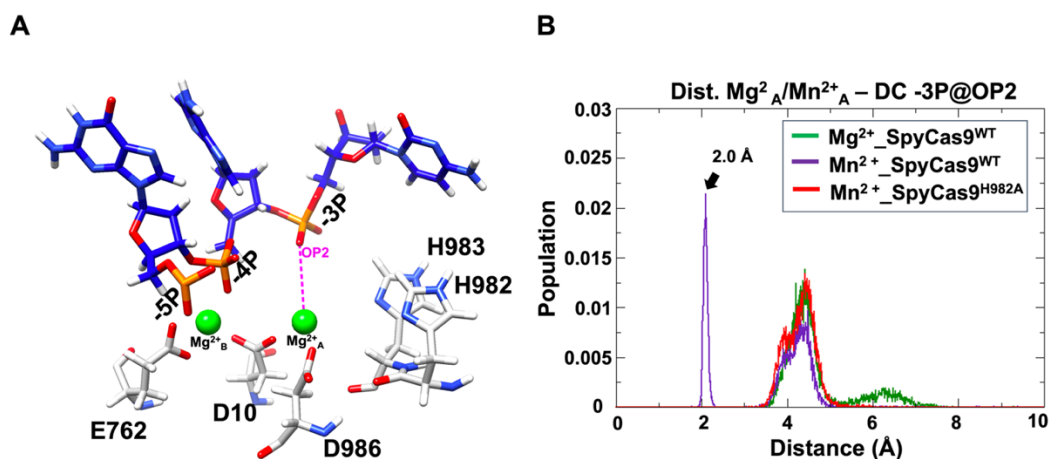
Although the H982A mutation was originally intended to remove undesirable gRNA-free DNA cleavage activity, it is also possible that it increases the selectivity of the on-target DNA cleavage activity. To test the hypothesis that H982A changes the RuvC nuclease domain such that the likelihood of cleavage decreases when there is an imperfect complementarity between the guide region of the gRNA and the DNA substrate, DNA substrates with a single mismatch at 3<sup>rd</sup>, 5<sup>th</sup>, or 18<sup>th</sup> nt (MM3, MM5, and MM18 respectively) position downstream of the PAM were used for gRNA-dependent off-target DNA cleavage assays. The total reaction time for this assay is 15 minutes using Mg<sup>2+</sup> as the divalent metal ion. Previous studies have shown that PAM-proximal mismatches are cleaved at varying efficiencies by SpyCas9 based on the mismatch position whereas PAM-distal mismatches are cleaved at similar efficiencies as the completely matched DNA (Jinek et al., 2012; Kuscu et al., 2014). The results from our mismatch cleavage assay showed that in general SpyCas9<sup>H982A</sup> has a lower efficiency in linearizing DNA substrates with a corresponding accumulation of nicked DNA (Figure 2.12). Comparing the different mismatches, SpyCas9<sup>H982A</sup> was better than SpyCas9<sup>WT</sup> at discriminating a mismatch at position 5 followed by MM18 (Figure 2.12B). Direct comparison of SpyCas9<sup>H982A</sup> and SpyCas9<sup>WT</sup> shows higher amounts of leftover supercoiled substrate and nicked intermediate in the case of SpyCas9<sup>H982A</sup> implicating better discrimination of DNA mismatches by SpyCas9<sup>H982A</sup>, albeit its slightly reduced on-target cleavage efficiency (Figure 2.12A).



**Figure 2.12. Mismatched DNA substrate cleavage assay to assess on-target selectivity (off-target discrimination).** The activity being tested is SpyCas9's gRNA-dependent DNA cleavage of supercoiled, double stranded pUC19 DNA in the presence of  $Mg^{2+}$ , and the DNA substrates vary based on the presence of mismatches between the gRNA and the target DNA at positions 3 (MM3), 5 (MM5), and 18 (MM18) nucleotides away from the PAM. **(A)** Representative gel and **(B)** quantification of fraction of DNA linearized for each lane (Equation 2.1). Control reactions have no protein. The error bars plotted are standard error of the mean (SEM) of three replications shown as black circles. According to the two-sample t-test with a p value of 0.05, there is a significant difference between WT and H982A for each substrate type. One, two, or three asterisks means  $p \leq 0.05$ , 0.01, or 0.001 respectively.

### 2.5.5 H982A proposedly modulates gRNA-free DNA cleavage by preventing proper positioning of single-stranded DNA and by weakening the interaction between $\text{Mn}^{2+}(\text{A})$ and D10

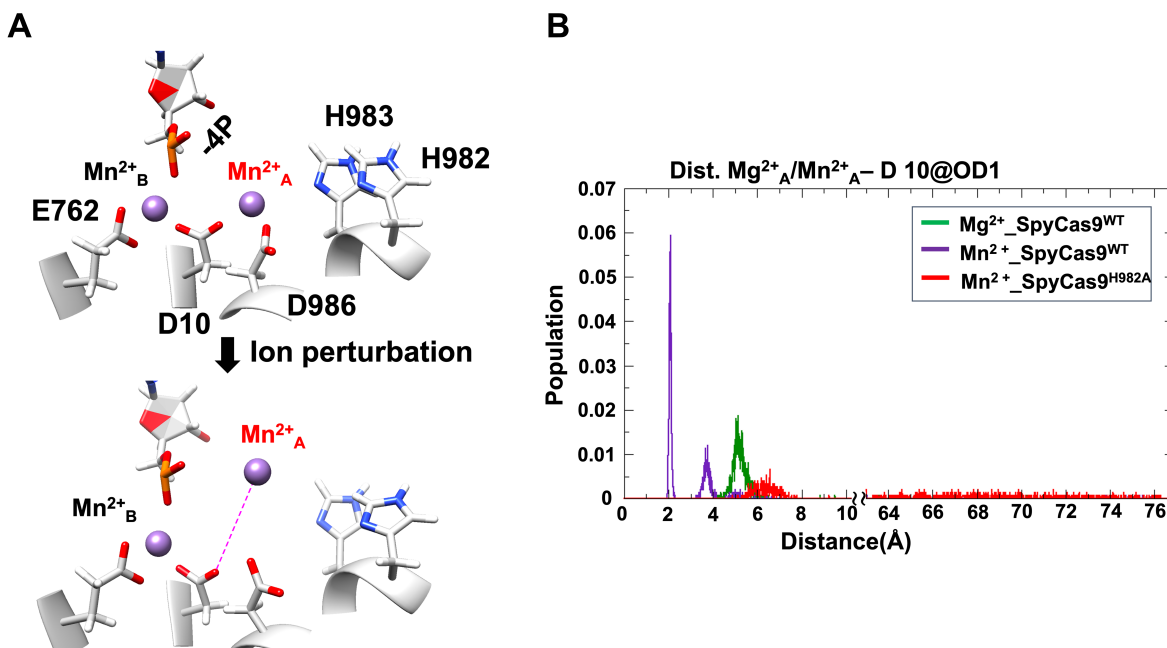
Molecular dynamics simulations (500 ns) were done with SpyCas9 in a gRNA-free state with either  $\text{Mg}^{2+}$  or  $\text{Mn}^{2+}$  in the RuvC catalytic center to investigate the mechanism of H982A's modulation of gRNA-free DNA cleavage activity. As exemplified by the SpyCas9<sup>WT</sup>- $\text{Mg}^{2+}$  condition that is biochemically shown to have little gRNA-free DNA cleavage activity (Figure 2.5), metal and DNA are improperly positioned in the RuvC catalytic center for gRNA-free DNA cleavage *via* the two-metal-ion catalysis mechanism when the distance between metal (A) and the phosphate backbone of DNA is  $\sim 4$  Å (Figure 2.13, green line in Figure 2.13B; (Zuo and Liu, 2016)). As exemplified by the SpyCas9<sup>WT</sup>- $\text{Mn}^{2+}$  condition biochemically shown to facilitate gRNA-free DNA cleavage (Figure 2.5), the RuvC is organized for gRNA-free DNA cleavage when the distance between metal (A) and the phosphate backbone is  $\sim 2$  Å (purple line in Figure 2.13B). Figure 2.13B shows that, as in the SpyCas9<sup>WT</sup>- $\text{Mg}^{2+}$  condition, the SpyCas9<sup>H982A</sup>- $\text{Mn}^{2+}$  condition makes the distance between metal (A) and the phosphate backbone  $\sim 4$  Å (red line in Figure 2.13B), so the H982A substitution disrupts the positioning of metal (A) and the DNA phosphate for gRNA-free DNA cleavage.



**Figure 2.13. The distance between  $\text{Mg}^{2+}_\text{A}/\text{Mn}^{2+}_\text{A}$  and non-target DNA (ntDNA) DC -3P (@OP2) in simulations. (A)** Scheme of the distance between  $\text{Mg}^{2+}_\text{A}$  and ntDNA (-3P) in our structural model. The dashed magenta line denotes the

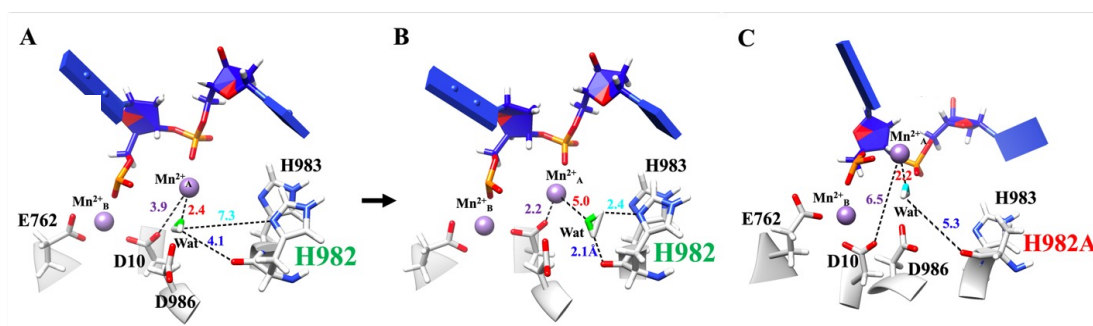
distance. **(B)** Distance distribution of the interactions between  $\text{Mg}^{2+}_A$  /  $\text{Mn}^{2+}_A$  and ntDNA DC -3P (@OP2). [100-400 ns; accumulated 12000 frames].

In addition to the ssDNA stabilization effect of  $\text{Mn}^{2+}(A)$  in a gRNA-free SpyCas9<sup>WT</sup> RuvC, which effect is disrupted by the H982A substitution, there may also be a  $\text{Mn}^{2+}$ -recruitment phenomenon in the WT, but not in the H982A variant, that brings  $\text{Mn}^{2+}(A)$  into the RuvC for two-metal-ion catalysis (Zuo and Liu, 2016). Ion perturbation experiments in which  $\text{Mn}^{2+}(A)$  was removed and followed by the addition of new divalent ions was allowed to re-equilibrate for 200-500 ns post-ion-perturbation (Figure 2.14A). The results were used to investigate  $\text{Mn}^{2+}$  recruitment. Results show that  $\text{Mn}^{2+}(A)$  is recruited to the RuvC catalytic center and positioned  $\sim 2 \text{ \AA}$  from catalytic residue D10 in the SpyCas9<sup>WT</sup>- $\text{Mn}^{2+}$  condition biochemically shown to facilitate gRNA-free DNA cleavage (purple line in Figure 2.14B; Figure 2.5). As exemplified by the SpyCas9<sup>WT</sup>- $\text{Mg}^{2+}$  and SpyCas9<sup>H982A</sup>- $\text{Mn}^{2+}$  conditions that are biochemically shown to not facilitate gRNA-free DNA cleavage (Figure 2.5), metal (A) is recruited to  $>4 \text{ \AA}$  from catalytic residue D10 in the simulations, showing that it is not ideally positioned for the RuvC two-metal-ion catalysis mechanism (green and red lines in Figure 2.14B; (Zuo and Liu, 2016)). Altogether, this shows that the H982 is essential for recruitment of  $\text{Mn}^{2+}(A)$  to the RuvC to facilitate gRNA-free DNA cleavage (Figure 2.14).



**Figure 2.14. Ion perturbation results comparing the ability of SpyCas9<sup>WT</sup> versus SpyCas9<sup>H982A</sup> RuvC to recruit Mg<sup>2+</sup>/Mn<sup>2+</sup>. (A)** Scheme of ion perturbation. **(B)** Distance distribution of the interactions between D10 (@OD1) and perturbed Mg<sup>2+</sup><sub>A</sub>/Mn<sup>2+</sup><sub>A</sub> in simulations.

In addition to the DNA backbone, metals, and catalytic residues, water molecules must also be positioned correctly in the RuvC active site to facilitate gRNA-free DNA cleavage. To observe how this positioning occurs, the trajectories of the ion perturbation experiments were analyzed. Results show that Mn<sup>2+</sup>(A) is coordinated by both D10 and a water molecule at the start of the simulation. During the simulation, H982 pulls that water molecule towards itself away from D10 (Figure 2.15A and B). This enables closer coordination of Mn<sup>2+</sup>(A) and D10. The H982A substitution causes the water molecule and Mn<sup>2+</sup>(A) to drift away from D10 (Figure 2.15C). Based on these observations, we propose that recruitment of the Mn<sup>2+</sup>(A) molecule to SpyCas9's RuvC domain for gRNA-free DNA cleavage involves proper positioning of water by H982.



**Figure 2.15. The role of a water molecule in the  $\text{Mn}^{2+}$  recruitment. (A and B)** The representative structure of SpyCas9<sup>WT</sup>-ntDNA- $\text{Mn}^{2+}$  before (A) and after (B) the formation of  $\text{Mn}^{2+}_A$  – D10 coordination. (C) The representative structure of SpyCas9<sup>H982A</sup>-ntDNA- $\text{Mn}^{2+}$  shown after  $\text{Mn}^{2+}_A$  moves away from D10.

## 2.6.0 Discussion

### 2.6.1 $\text{Mn}^{2+}$ binding in RuvC active site allosterically improves the interaction between SpyCas9 and ntDNA

Our activity assays showed that RuvC promoted gRNA-free ssDNA cleavage in the presence of  $\text{Mn}^{2+}$ , but not with  $\text{Mg}^{2+}$  (Figure 2.5). Specifically, a significant reduction of gRNA-free DNA cleavage in SpyCas9<sup>H982A</sup> in the presence of  $\text{Mn}^{2+}$  was observed (Figures 2.7 and 2.8, Table 2.2). While SpyCas9<sup>H982A</sup> shows almost complete absence of gRNA-free DNA cleavage,  $\text{Co}^{2+}$  can still promote DNA cleavage. The higher electronegativity of  $\text{Co}^{2+}$  may enable enhanced metal coordination even in the absence of H982. Nevertheless, a ~17-fold reduction of rate constant in SpyCas9<sup>H982A</sup> compared to that in SpyCas9<sup>WT</sup> indicates that H982 plays a role in stabilizing  $\text{Co}^{2+}$  under gRNA-free conditions. Interestingly, comparing the effects of  $\text{Mg}^{2+}$  and  $\text{Mn}^{2+}$  for gRNA-dependent DNA cleavage demonstrates that  $\text{Mn}^{2+}$  slightly reduces the rate of this DNA cleavage. It is possible that there are differences in the organization of the RuvC active site in the presence of gRNA with  $\text{Mg}^{2+}$  *versus*  $\text{Mn}^{2+}$  that causes differences in DNA cleavage rates. The hypothesized mechanism for  $\text{Mn}^{2+}$ -mediated gRNA-free DNA cleavage in SpyCas9<sup>WT</sup> is that, relative to  $\text{Mg}^{2+}$ (A) under gRNA-free conditions,  $\text{Mn}^{2+}$ (A) under gRNA-free conditions is better recruited and positioned (along with ssDNA, RuvC catalytic residue D10, and water) for two-metal-ion

catalysis of ssDNA cleavage. The proposed mechanism of H982A's modulation of  $Mn^{2+}$ -mediated gRNA-free DNA cleavage is that H982 helps recruit  $Mn^{2+}$ (A) to the RuvC catalytic center and position a water molecule such that  $Mn^{2+}$ (A) is ideally coordinated by RuvC catalytic residue D10 to facilitate ssDNA cleavage.

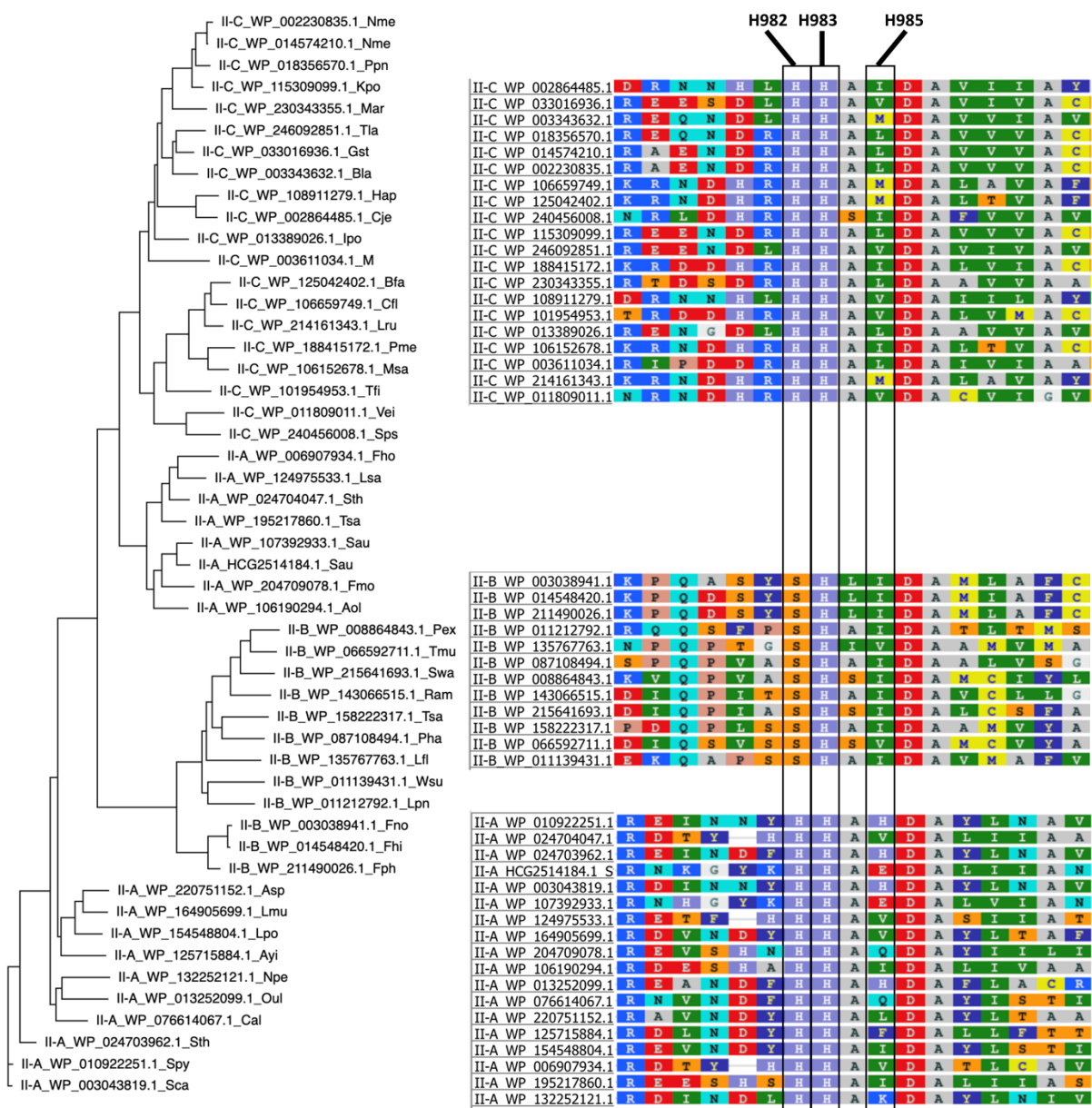
#### 2.6.2 Phylogenetic conservation of histidine residues in Cas9 active sites

There are three histidine residues (H982, H983, and H985) in SpyCas9's RuvC active site and their prevalence indicates that these residues may partially compensate for each other as general (Lewis) bases as indicated by the ability of imidazole to rescue the activity of the SpyCas9<sup>H982A/H983D/H985A</sup> variant (Tang et al., 2021). Studies have shown that the main catalytic mediator for gRNA-dependent DNA cleavage is H983 (Casalino et al., 2020), with a complete inactivity of ntDNA cleavage with a H983A substitution in the presence of  $Mg^{2+}$  (Furuhata and Kato, 2021) (Figure 2.2G). Interestingly,  $Mn^{2+}$  can partially rescue the gRNA-dependent DNA cleavage activity in H983A or H983G substitutions, but its rescue effect is significantly reduced in H983G/H982A double substitution (Furuhata and Kato, 2021). Unlike other RuvC-like enzymes, SpyCas9 contains only a DED (Asp10, Glu762 and Asp986) feature instead of the typical DEDD motif in its RuvC center. It is possible that histidine (H983) compensates for the last D of DEDD motif (Dürr SL, 2019), and that extra histidine residues in the RuvC site may enable the enzyme to adapt to different types of DNA cleavage conditions.

Our data shows that H982 plays a very prominent role in SpyCas9's gRNA-free DNA cleavage, with H982A substitution causing ~167-fold reduction in  $Mn^{2+}$ -mediated gRNA-free DNA cleavage. The fact that the H982A variant is unable to perform gRNA-free DNA cleavage (Figure 2.2B) indicates that for SpyCas9's gRNA-free DNA cleavage, H983 acts as the base while H982 is involved in metal coordination and DNA binding.

To further probe the essentiality of multiple histidine residues in Cas9 sequences, we performed multiple sequence alignment with Cas9 protein sequences. The results show that the three His in the SpyCas9 RuvC active site are conserved to varying degrees among CRISPR-Cas subtypes II-A, II-B, and II-C (2.). This molecular phylogeny analysis included the most variable and frequently occurring Cas9s in the [NCBI database](#) (see Methods section 9 for details). Cas9s tested for gene editing (see the Figure 2.16 caption) were individually added to this list of Cas9s (complete list of Cas9 sequences in supplementary file Cas9-seq.csv). As expected, H983 is conserved among all subtypes since it is crucial to the two metal-ion catalysis mechanism (Figure 2.2G). The multiple sequence alignment provides evidence that H985 may not be crucial to the two metal-ion catalysis mechanism since it is only present in a selected few type II-A Cas9s (Figure 2.16). Cas9 subtype II-B has a conserved Serine where the H982 equivalent is located in subtypes A and C (Figure 2.16). Interestingly, the RuvC domain of *Francisella novicida* Cas9 (FnoCas9) has severely reduced/absent Mn<sup>2+</sup>-dependent, gRNA-free ssDNA cleavage (Sundaresan et al., 2017), while *Campylobacter jejuni* Cas9 (CjeCas9) has been shown to possess gRNA-free ssDNA degradation using a combination of HNH and RuvC domains for the DNA cleavage mechanism (Saha et al., 2020c). The RuvC active sites at the H982 position are conserved among type II-C and type II-A Cas9s, implicating their potential to promote gRNA-free DNA cleavage. Further experimental characterization of multiple Cas9 sequences with differences in the prevalence of histidine residues in the RuvC active site and different structural features (for example, different domain compositions due to protein size differences between the different type II subtypes) may enable such conclusions.





**Figure 2.16. Conservation of H982A in Cas9 orthologs.** Molecular phylogeny of Cas9 and multiple sequence alignment (MSA) of the H982 among the different type II CRISPR-Cas subtypes. Both the molecular phylogeny and MSA have the same proteins, but the labels differ in that the three-letter abbreviations of the bacterial species names are appended to the molecular phylogeny branch labels. The Cas9 proteins were obtained by downloading one genome for each bacterial species containing an annotated *cas9* gene in NCBI. Subtypes were identified

using a program called CRISPRCasTyper (Russel et al., 2020). *Streptococcus pyogenes* Cas9 (SpyCas9 (Jinek et al., 2012), type II-A), *Streptococcus thermophilus* Cas9s (Sth1Cas9 from strain DGCC7710 (Müller et al., 2016) and Sth3Cas9 from strain LMD-9 between open reading frames (ORFs) STER\_1474 and STER\_1473 (Horvath et al., 2008; Müller et al., 2016; Xu et al., 2015), type II-A), *Staphylococcus aureus* subspecies *aureus* strain 1276-3 Cas9 (SauCas9 (Ran et al., 2015), type II-A), *Streptococcus canis* strain FSL Z3-227 Cas9 (ScaCas9 (Chatterjee et al.), type II-A), *Staphylococcus auricularis* (SauriCas9 (Hu et al., 2020; Xu et al., 2022), type II-A), *Francisella tularensis* subspecies *novicida* strain U112 Cas9 (FnoCas9 (Oh et al., 2022), type II-B), *Campylobacter jejuni* strain NCTC 11168 Cas9 (CjeCas9 (Kim et al., 2017), type II-C), *Geobacillus stearothermophilus* (GoeCas9 (Harrington et al., 2017), type II-C), *Brevibacillus laterosporus* strain SSP360D4 Cas9 (BlatCas9 (Gao et al., 2020; Karvelis et al., 2015), type II-C), *Pasteurella pneumotropica* (also known as *Rodentibacter pneumotropicus*) Cas9 (PpCas9 (Fedorova et al., 2020), type II-C), and *Neisseria meningitidis* Cas9s (Nme1Cas9 from strain 8013 (Hou et al., 2013; Zhang et al., 2013) and Nme2Cas9 from strain De11444 (Edraki et al., 2019), type II-C) were added to the Cas9 dataset because they are proteins of interest for gene editing (Tadić et al., 2019; Wei et al., 2022). The Cas9s of interest for gene editing have the abbreviation of the organism and protein name added to the end of their labels in the MSA. This comparison of SpyCas9s of interest for gene editing biotechnology (including those with known gRNA-free DNA cleavage activity: SpyCas9 (Sundaresan et al., 2017), FnoCas9 (Sundaresan et al., 2017), and CjeCas9 (Saha et al., 2020c)) enables assessment of amino acid distributions in the RuvC active site and analyze the propensity to catalyze gRNA-free DNA cleavage. The Cas9 sequences used in this analysis are tabulated in Cas9-seq.csv in the supplementary information. SpyCas9 residue positions H982, H983, and H985 are labeled in the MSA.

### 2.6.3 RuvC is organized as an autonomous endonuclease that can access and cleave DNA without gRNA-dependent conformational changes

The conformational changes upon gRNA binding and organization of HNH and RuvC endonucleases are essential for concerted cleavage needed to make a dsDNA break through gRNA-dependent DNA cleavage. Upon R-loop formation past the seed region of the gRNA, conformational changes of the REC domains of Cas9 enables docking of HNH in a cleavage competent form. HNH docking in turn triggers activation of RuvC for DNA cleavage through conformational changes of a hinge connecting HNH and RuvC domains. For gRNA-free DNA cleavage, these conformational changes seem to be dispensable for RuvC to cleave ssDNA. The fact that SpyCas9 preferentially degrades ssDNA by gRNA-free DNA cleavage points to RuvC's ability to accommodate only ssDNA in its catalytic center. It is possible that the improved SpyCas9-ntDNA interactions in the presence of  $Mn^{2+}$ , and being mediated by H982, are important for by-passing the extensive conformational changes needed for RuvC to access and cleave DNA by gRNA-dependent DNA cleavage (Jinek et al., 2014). In this sense, gRNA-free DNA cleavage is much like the *trans*-DNA cleavage that is an aftermath of *cis*-cleavage in Cas12a where the RuvC active site is exposed for progressive ssDNA degradation. For *trans*-cleavage by Cas12a, RuvC is activated by removing occlusions to its active site through crRNA-target DNA binding or after the cleaved DNA product at the PAM-distal site diffuses away creating free access of ssDNA into RuvC's active site pocket. It can be envisioned that interactions of H982,  $Mn^{2+}$  and ssDNA are similar to activation needed for Cas12a's *trans*-cleavage by keeping RuvC active site accessible for progressive degradation of ssDNA.

#### 2.6.4 RuvC active site modifications can potentially remove gRNA-free and off-target DNA cleavages

One of the undesirable aspects of SpyCas9 based genome applications is its tolerance to gRNA-DNA mismatches, which causes off-target DNA cleavages during genome applications. SpyCas9<sup>WT</sup> cleaves PAM-proximal mismatches at positions 3-7 at variable efficiencies, with MM5 being the most sensitive (Jinek et al., 2012). Our results show that there is further increase in discrimination of

MM5 in SpyCas9<sup>H982A</sup> compared to that in SpyCas9<sup>WT</sup> (Figure 2.12). A recent study analyzing DNA cleavage efficiencies of SpyCas9 showed that SpyCas9<sup>WT</sup> discriminated against 2 consecutive nt mismatches anywhere from nucleotide position 1-12 from the PAM (Tang et al., 2021). The H982A and H983D variants were able to discriminate mismatches in positions 1-15 and additionally 18, and the double mutant (H982A/H983D) was able to discriminate mismatches in positions 1-15 and additionally 18-19 (Tang et al., 2021). These results indicate that SpyCas9<sup>H982A</sup> can reduce both gRNA-free and off-target DNA cleavages, two of the undesirable effects of SpyCas9-based genome applications.

#### 2.6.5 Fine-tuning the SpyCas9 active site divalent metal ion interactions for a safer gene editing tool

Since SpyCas9 is a powerful gene editing tool for clinical gene therapy, it will raise critical concerns if its specificity is impaired by certain divalent metal cations. The gRNA-free DNA cleavage phenomena we addressed serves as a good example of how the specificity of SpyCas9 is compromised by different transition metals such as  $\text{Mn}^{2+}$ . Since it is unlikely that SpyCas9 is applied solely to  $\text{Mn}^{2+}$  free systems, optimizing SpyCas9 becomes a potential strategy to tackle the issue of non-specific DNA cleavage. Our study of gRNA-free DNA cleavage offers perspectives on how various metals modulate biological function of the same protein, and it might shed light on off-target issues of CRISPR-SpyCas9.

### 3 Chapter 3: Distinguishing the roles of helix-integrity *versus* protein-nucleic acid interactions of the SpyCas9 bridge helix in modulating RNA-guided DNA cleavage target selectivity

#### 3.1.0 Authorship

The SpyCas9 cloning, purification, and *in vitro* DNA cleavage assays were done by Dr. Kesavan Babu and me at the University of Oklahoma Department of Chemistry and Biochemistry. Protein variants Spy<sup>E60C/2Pro/C80S/C574S</sup>, SpyCas9<sup>E60C/2Ala/C574S</sup>, SpyCas9<sup>2Pro/C574S</sup> and SpyCas9<sup>E60C/2Pro/C80S/C574S</sup> were cloned and/or purified by Tamara Flusche at the University of Oklahoma Department of Chemistry and Biochemistry. Ongoing work with electron paramagnetic resonance measurements of SpyCas9 is being done by Difei Wu, a member of Dr. Peter Qin's laboratory at the University of Southern California Department of Chemistry. Figure 3.2 was made by Venkatesan Kathiresan from Dr. Peter Qin's laboratory.

#### 3.2.0 Overview

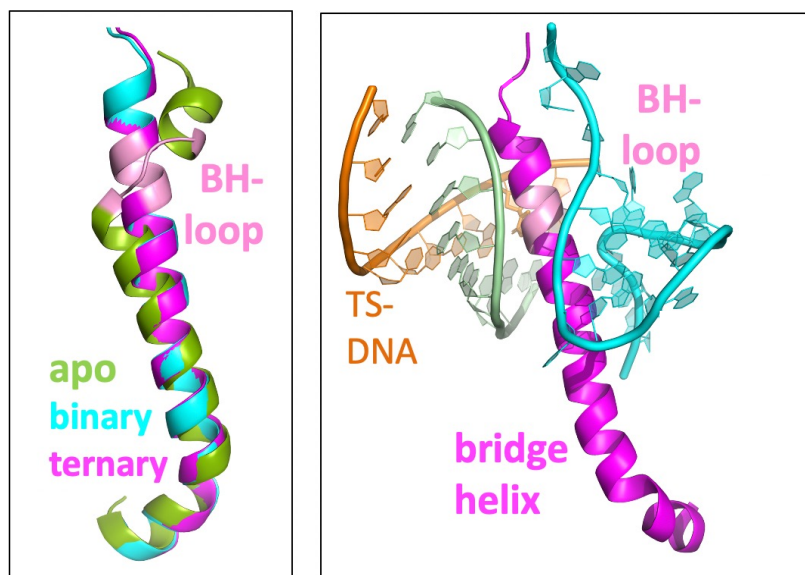
*Streptococcus pyogenes* Cas9 (SpyCas9) is the Cas protein commonly used in gene editing biotechnologies, but there are concerns with its safety in gene therapy applications because it can cleave DNA in regions similar to, but not perfectly complementary to the target site. A previously established SpyCas9 variant (SpyCas9<sup>2Pro</sup>), which has an L64P/K65P substitution in a region called the bridge helix (BH) connecting the nucleic acid-recognition and nuclease lobes of the protein, has increased on-target DNA cleavage selectivity. The proposed mechanism is that the two-proline substitution compromises a loop-to-helix transition that occurs in the BH when SpyCas9 binds to guide RNA, and that the compromised integrity of BH in the SpyCas9<sup>2Pro</sup>-gRNA ribonuclear particle impairs the concerted conformational changes essential for target DNA cleavage. Another possibility is that the loss of intra-protein/nucleic acid interactions of the residues being substituted can contribute either solely or in combination with the integrity of the BH to impair off-target DNA cleavage. To test these scenarios, a

two-alanine substitution variant (L64A/K65A, SpyCas9<sup>2Ala</sup>), which is expected to have a fully helical BH but lack sidechain interactions, was assayed *in vitro* for DNA cleavage efficiencies with different target DNAs. SpyCas9<sup>2Ala</sup> proved to be faster than SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> at DNA cleavage. In terms of DNA cleavage efficiency relative to SpyCas9<sup>WT</sup>, SpyCas9<sup>2Ala</sup> displayed reduced selectivity in cleaving DNA with mismatches at positions closer to the protospacer adjacent motif (PAM); SpyCas9<sup>2Pro</sup>, however, showed increased selectivity. Our results indicate a prominent role of BH integrity, as opposed to sidechain interactions, in mismatch selectivity regarding target DNA linearization efficiency. Ongoing studies are focused on experimentally measuring the helical conformation of the BH in SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup>, and SpyCas9<sup>2Ala</sup> with different types of DNA substrates. Electron paramagnetic resonance (EPR) studies are being used to detect the loop-to-helix transition, and several SpyCas9 variants with ideally placed cysteine residues have been prepared to probe the conformational changes. The results of these studies will shed light on mechanisms that determine the on-target DNA cleavage selectivity of SpyCas9 to guide the development of safer gene therapy biotechnologies.

### 3.3.0 Introduction

The two-proline (L65P/K65P) substitution in SpyCas9 (SpyCas9<sup>2Pro</sup>) improves its RNA-guided DNA cleavage specificity (Babu et al., 2019). In other words, the two-proline substitution makes SpyCas9 less likely to fully linearize dsDNA that has an imperfect complementarity with the gRNA. Specifically, SpyCas9<sup>2Pro</sup> is most sensitive to mismatches in the substrate dsDNA that are 7 nt upstream of the PAM and in general displays reduced efficiency in cleaving DNA harboring PAM-proximal mismatches (Babu et al., 2019). There could be many aspects of the mechanism for this increase in selectivity for on-target DNA cleavage. One potential explanation is that the two-proline substitution disrupts the loop-to-helix transition that occurs upon sgRNA binding, and this disruption affects the integrity of the BH and its role in positioning the HNH and RuvC nuclease domains for complete cleavage of the target and non-target strands, respectively,

of the dsDNA (Figure 3.1) (Babu et al., 2019). Another possibility is that the proline substitutions eliminate intra-protein and protein-nucleic acid interactions that are essential for conformational changes needed for target DNA cleavage. Of importance is the interaction between the K56 sidechain and the phosphate lock loop residue E1108 that has been shown to be important for strand switching of the DNA and the accompanying R-loop formation that is a prerequisite for Cas9 conformational changes and target DNA cleavage (Anders et al., 2014; Babu et al., 2019). We propose that both BH integrity and interactions with protein and nucleic acids contribute to target DNA cleavage selectivity, with a more prominent contribution from the BH integrity. Previous studies from our lab have shown altered coordination between the HNH and RuvC endonuclease domains in SpyCas9<sup>2Pro</sup> with the coordination pattern changing with respect to the DNA mismatch position (Babu et al., 2021). *In vitro* DNA cleavage assays and kinetic analyses showed that the relative use of two possible DNA cleavage pathways, which differ based on whether the initial nick occurs in the target strand by HNH or in the non-target strand by RuvC, is affected by the position of the mismatch relative to the PAM in the substrate DNA (Babu et al., 2021).



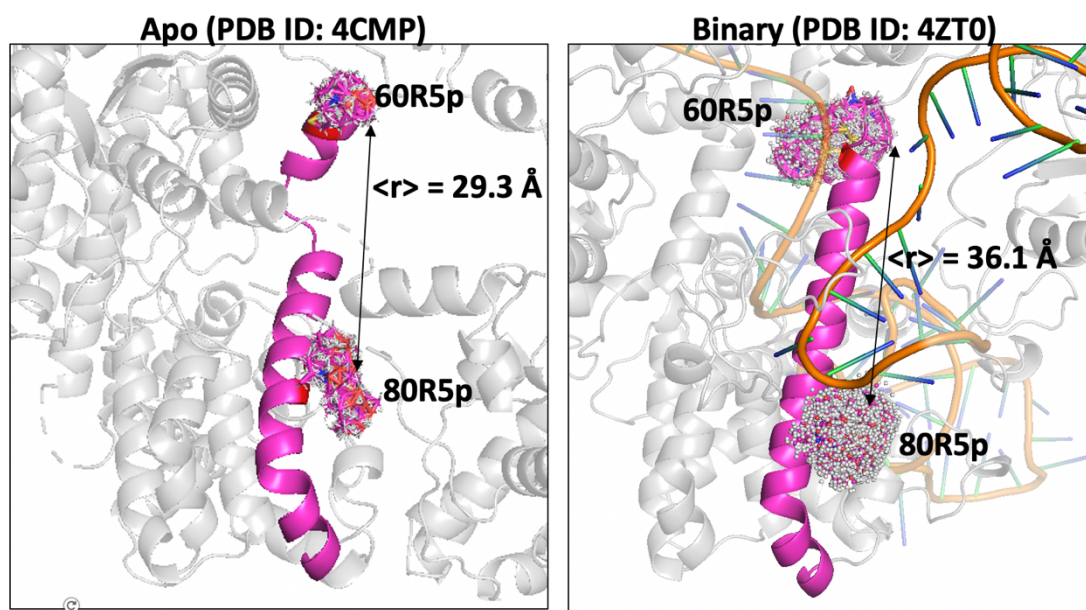


**Figure 3.1 Loop to helix conversion that occurs in the region Leu64-Thr67 in the bridge helix (BH) of SpyCas9 upon gRNA-binding.** The loop form of Leu64-Thr67 (labeled BH-loop) is shown in light pink and the helical regions are shown in magenta. Apo is when SpyCas9 is not bound to any nucleic acid, binary is when it is bound to gRNA, and ternary is when it is bound to gRNA and target DNA. The target strand (TS) of the target DNA is the strand that hybridizes to the gRNA.

To identify whether compromised BH integrity or loss of intra-protein and protein-nucleic acid interactions has the greater effect on on-target DNA cleavage selectivity, *in vitro* DNA cleavage assays with a SpyCas9 variant that both lacks the sidechain interaction and has a strong propensity to undergo the loop-to-helix transition is needed. Toward this end, the 2Ala (L64A/K65A) variant (SpyCas9<sup>2Ala</sup>) was constructed since alanine has a strong propensity toward helix formation in proteins and does not have a positive charge to facilitate the intra-protein interactions observed for the K65 sidechain (Blaber et al., 1993).

To confirm that the two-proline substitution impairs the loop-to-helix transition, electron paramagnetic resonance (EPR) can be used to measure changes in distance between BH residues located on either side of the loop since the distance between these residues will change during the loop-to-helix transition (or lack thereof) upon SpyCas9-sgRNA-binding (Figure 3.2) (Vazquez Reyes et al., 2017). The more efficient this transition is, the more robust (and less selective) the DNA cleavage is expected to be when the target DNA has a mismatch. In collaboration with Dr. Peter Qin at the University of Southern California, we plan to use site-directed spin labeling (SDSL) EPR to detect the change in distance between residues on opposite ends of the loop region that transitions to a helix make distance measurements. SDSL-EPR requires specific cysteine residues to be introduced into the protein where nitroxide labels will be attached for distance measurements between them. Following the structures of BH in the crystal structures of apo-SpyCas9 (PDB ID: 4CMP, (Jinek et al., 2014))

and SpyCas9-sgRNA structure (PDB ID: 4ZT0, (Jiang et al., 2015)), we identified Cys80 and Glu60 as the ideal candidates for labeling. Dr. Qin's lab used the program ALLNOX (Ding et al., 2019) to measure the distance between nitroxide labels at 60<sup>th</sup> and 80<sup>th</sup> positions and calculated a distance change of 29.3 Å to 36.1 Å in response to loop-to-helix transition, which is a measurable distance by SDSL EPR. Glu60 was substituted with a Cys (E60C) to enable labeling at C80 and E60C (Babu et al., 2019).



**Figure 3.2. Distance comparison between SpyCas9 residue positions 60 and 80 before (Apo, PDB ID: 4CMP) and after (Binary, PDB ID: 4ZT0) gRNA-binding.** The loop-to-helix transition upon gRNA-binding occurs at residue positions 64-67 to increase the distance between positions 60 and 80 from 29.3 Å to 36.1 Å. The nitroxide label is designated as R5p. gRNA is shown in orange in the binary structure.

In this research, *in vitro* DNA cleavage assays with SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> are used to elucidate whether a robust loop-to-helix transition or K65 sidechain interactions have the most prominent role in the observed increase in the target DNA cleavage selectivity of SpyCas9<sup>2Pro</sup>. Furthermore, SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> variants that facilitate detection of the loop-to-helix transition using

EPR measurements were prepared to confirm that the loop-to-helix transition is impacted by proline substitution but not with alanine substitution.

#### 3.4.0 Methods

##### 3.4.1 Construction of BH loop variants to measure contributions of bridge helix integrity and sidechain interaction in target DNA cleavage sensitivity

Previous studies in the lab created SpyCas9<sup>2Pro</sup> that possessed L64P and K65P substitutions in SpyCas9 protein (Babu et al., 2019; Babu et al., 2021). To maintain BH integrity while changing sidechain interactions, we constructed a SpyCas9 variant with L64A and K65A substitutions termed SpyCas9<sup>2Ala</sup>.

##### 3.4.2 Designing SpyCas9 constructs for nitroxide labeling for SDSL EPR studies

To tag SpyCas9 residue positions 60 and 80 with nitroxide labels that enables the measurement of distances between these positions using EPR, the residues must be either native or substituted cysteines which form thio-ether bonds with the nitroxide label precursor 3-iodomethyl-1-oxyl-2,2,5,5-tetramethylpyrroline (Vazquez Reyes et al., 2017). SpyCas9 natively possesses two cysteines, C80 and C574. Previous research in the lab had constructed a SpyCas9<sup>C574S</sup> variant for EPR studies that retains one cysteine at 80<sup>th</sup> position (C80) (Vazquez Reyes et al., 2017). This backbone was used to construct a E60C variant, followed by using that to create the BH variants needed in this study. The EPR experiments use the following variants: SpyCas9<sup>WT-E60C</sup> (contains E60C and C574S substitutions, named SpyCas9<sup>WT-BH-epr</sup>), SpyCas9<sup>2Pro-E60C</sup> (contains E60C, L64P, K65P, and C574S substitutions, named SpyCas9<sup>2Pro-BH-epr</sup>), and SpyCas9<sup>2Ala-E60C</sup> (contains E60C, L64A, K65A, and C574S substitutions, named SpyCas9<sup>2Ala-BH-epr</sup>).

To ensure that Cys substitutions do not impact protein's activity, different variants were purified to and assayed *in vitro* for RNA-guided DNA cleavage to confirm the proteins are properly folded (Figure 3.9 lanes 5-12). The EPR experimental SpyCas9 variants are SpyCas9<sup>WT-BH-epr</sup>, SpyCas9<sup>2Pro-BH-epr</sup>, and SpyCas9<sup>2Ala-BH-</sup>

<sup>epr</sup>, (lanes 5, 6, and 12 respectively of Figure 3.9). The control variants that are needed to determine the background signal from unincorporated nitroxide label are the proteins with no cysteines (C80S/C574S, SpyCas9<sup>WT-epr-C</sup> (reported in a previous work; (Vazquez Reyes et al., 2017)), SpyCas9<sup>2Pro-epr-C</sup> (lane 8 of Figure 3.9) (Vazquez Reyes et al., 2017)).

### 3.4.3 Cloning and purifying SpyCas9 protein variants

To clone the SpyCas9 variants for nitroxide labeling (E60C/L64P/K65P/C574S (2Pro-BH-epr), E60C/C574S (WT-BH-epr), L64P/K65P/C574S, L64P/K65P/C80S/C574S (2Pro-epr-C), E60C/C80S/C574S, E60C/L64P/K65P/C80S/C574S, C574S, E60C/L64A/K65A/C574S (2Ala-BH-epr)), site-directed mutagenesis (SDM) was used to introduce the cysteine and serine mutations (see Table 3.1 for SDM template and primer sequences for the SpyCas9 variants I cloned in the Rajan lab; see the authorship section for credit given to Kesavan Babu, a previous post-doctoral fellow in the Rajan lab, and Tamara Flusche, a graduate student in the Rajan lab, who cloned the remaining variants for nitroxide labeling).

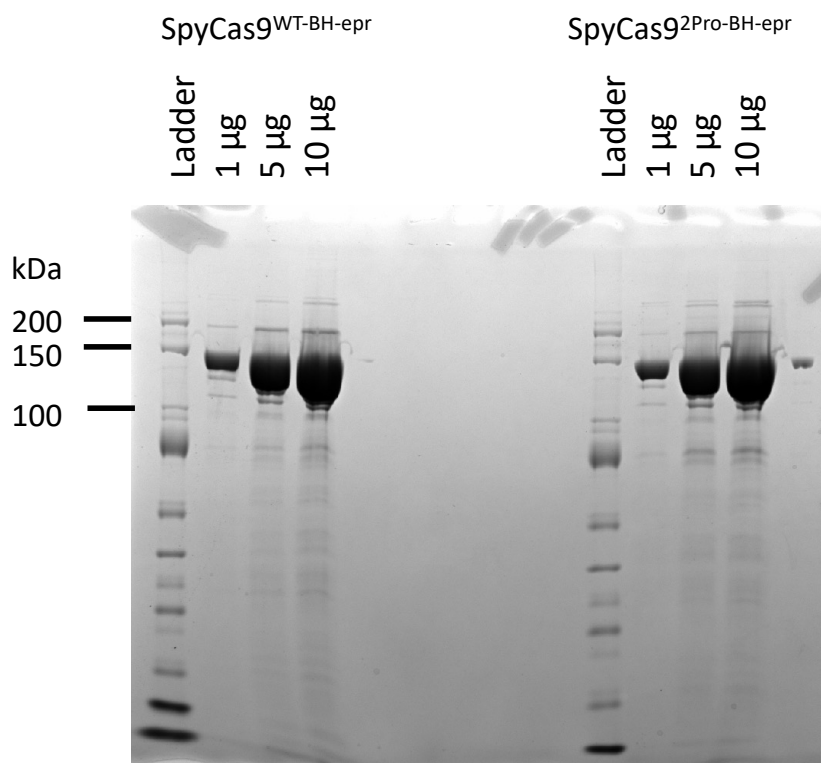
| SpyCas9 Variant        | SDM Template     | SDM Primer Name | SDM Primer Sequence                                           |
|------------------------|------------------|-----------------|---------------------------------------------------------------|
| 2Pro (L64P/K65P) C574S | 2Pro (L64P/K65P) | C574S-F         | GAAGATTATTTCAAAAAAA<br>TAGAATCCTTTGATAGTGT<br>TGAAATTTTCAGGAG |
|                        |                  | C574S-R         | CTCCTGAAATTTCAACACT<br>ATCAAAGGATTCTATTTTT<br>TTGAAATAATCTTC  |
| E60C                   | E60C             | C80S-F          | CGTCGGAAGAATCGTATT<br>TCTTATCTACAGGAGATTT<br>TTTCAAATGAG      |

|                                          |                                  |                  |                                                           |
|------------------------------------------|----------------------------------|------------------|-----------------------------------------------------------|
| 2Pro<br>(L64P/K65P<br>)<br>C80S<br>C574S | 2Pro<br>(L64P/K65P<br>)<br>C574S | C80S-R           | CTCATTTGAAAAAATCTCC<br>TG TAGATAAGAAATACGAT<br>TCTTCCGACG |
| E60C<br>2Ala<br>(L64A/K65A<br>)<br>C574S | C574S                            | E60C_L64A_K65A_F | GTGTGCGACTCGTGCCGC<br>ACGGACAGCTCGTA                      |
|                                          |                                  | E60C_L64A_K65A_R | TACGAGCTGTCCGTGCGG<br>CACGAGTCGCACAC                      |

**Table 3.1.** Site directed mutagenesis templates and primers for SpyCas9 variants.

To purify the SpyCas9 variants for nitroxide labeling (see Figure 3.3 of a sample cation exchange gel showing protein purity), nickel-NTA and cation exchange columns were used as described in previously published protocols (Jinek et al., 2012; Sundaresan et al., 2017). Briefly, the Ni-NTA column (HisTrap FF Crude, 5 mL) was washed with 15 CV buffer A (50 mM TRIS pH 8, 500 mM NaCl, 10 mM imidazole, 1 mM TCEP) and loaded with clarified cell lysate containing the Cas9 protein. Gradient elution was done using 0-100% Nickel-NTA buffer B (50 mM TRIS pH 8, 500 mM NaCl, 500 mM imidazole, 1 mM TCEP) over 100 mL. The expected size of MBP-SpyCas9 is 202 kDa. Fractions were pooled and dialyzed overnight in cation exchange buffer A (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM DTT, 10% glycerol) at 4°C with a simultaneous incubation with TEV protease to remove the MBP tag. The sample was loaded onto cation exchange column (SP HP, 5 mL) and eluted over 100 mL, 0-100% cation exchange buffer B (20 mM HEPES pH 7.5, 1 M KCl, 2 mM DTT, 10% glycerol). The SpyCas9 protein size after cleavage of MBP tag is 158 kDa. The protein was dialyzed into storage buffer (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM TCEP) and concentrated to

10 mg/mL using a Nanodrop spectrophotometer for concentration quantification. The protein was then aliquoted and flash frozen in liquid nitrogen for storage at -80°C.



**Figure 3.3. SDS-PAGE gel showing protein purity of SpyCas9 variants SpyCas9<sup>WT-BH-epr</sup> and SpyCas9<sup>2Pro-BH-epr</sup> purified for nitroxide labeling.** The ladder is Unstained Protein Standard Broad Range (10-200 kDa) from New England Biolabs. The expected protein size is 158 kDa.

While proteins used for EPR experiments had only Ni-NTA and cation exchange column purification because size exclusion would be done after nitroxide labeling, proteins used for activity assay measurements (SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup>, SpyCas9<sup>2Ala</sup>) were purified using an additional size exclusion column step. Instead of being dialyzed into storage buffer like the variants for nitroxide labeling, these proteins were concentrated using a centrifugal filter and injected onto a size exclusion column (SEC, Hi Prep 16/60 Sephacryl S-300 HR,

120 mL) equilibrated in SEC buffer (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM TCEP) and eluted isocratically with 120 mL of SEC buffer. They were then quantified using a Nanodrop spectrophotometer, aliquoted, and flash frozen in liquid nitrogen for storage at -80°C. The variants for EPR experiments were shipped to Dr. Qin's laboratory, where they incubated proteins with nitroxide/RNA/DNA and performed an SEC purification to separate unused label from the protein/protein-nucleic acid complex.

#### 3.4.4 *In vitro* gRNA-guided double stranded DNA cleavage assays

Reaction components were 1X buffer (20 mM HEPES pH 7.5, 150 mM KCl, 2 mM TCEP), 100 ng (5 nM) double stranded plasmid DNA (pUC19) containing a target site complementary to the guide region of the sgRNA (98 nt; (Babu et al., 2019; Nishimasu et al., 2014)) or with some mismatches to the guide region, 5 mM divalent metal, 50 nM sgRNA, and 50 nM SpyCas9 incubated at 37°C. The reaction products were resolved on a 1.1% agarose gel and the bands were visualized post-gel running using 0.1 µg/mL ethidium bromide.

#### 3.4.5 Quantification of *in vitro* gRNA-guided double stranded DNA cleavage assays

The fraction or percent of double stranded pUC19 substrate DNA linearized or nicked by Cas9 was calculated as follows:

$$\text{Fraction Linear pUC19} = \frac{L}{N + L + SC} = \alpha(1 - e^{-k_{\text{obs}}t}) \quad (\text{Equation 3.1})$$

$$\text{Percent linear} = \left( \frac{L}{N + L + SC} \right) \times 100\% \quad (\text{Equation 3.2})$$

$$\text{Percent nicked} = \left( \frac{N}{N + L + SC} \right) \times 100\% \quad (\text{Equation 3.3})$$

Where L is the intensity of the linear band, N is the intensity of the nicked band, SC is the intensity of the supercoiled band,  $\alpha$  is the active fraction (the fraction of

DNA linearized as time approaches 60 min),  $k_{obs}$  is the rate of linear product formation in units of  $\text{min}^{-1}$ , and  $t$  is time in units of min.

### 3.5.0 Results

#### 3.5.1 SpyCas9<sup>2Ala</sup> has a higher RNA-guided DNA linearization rate than both SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup>

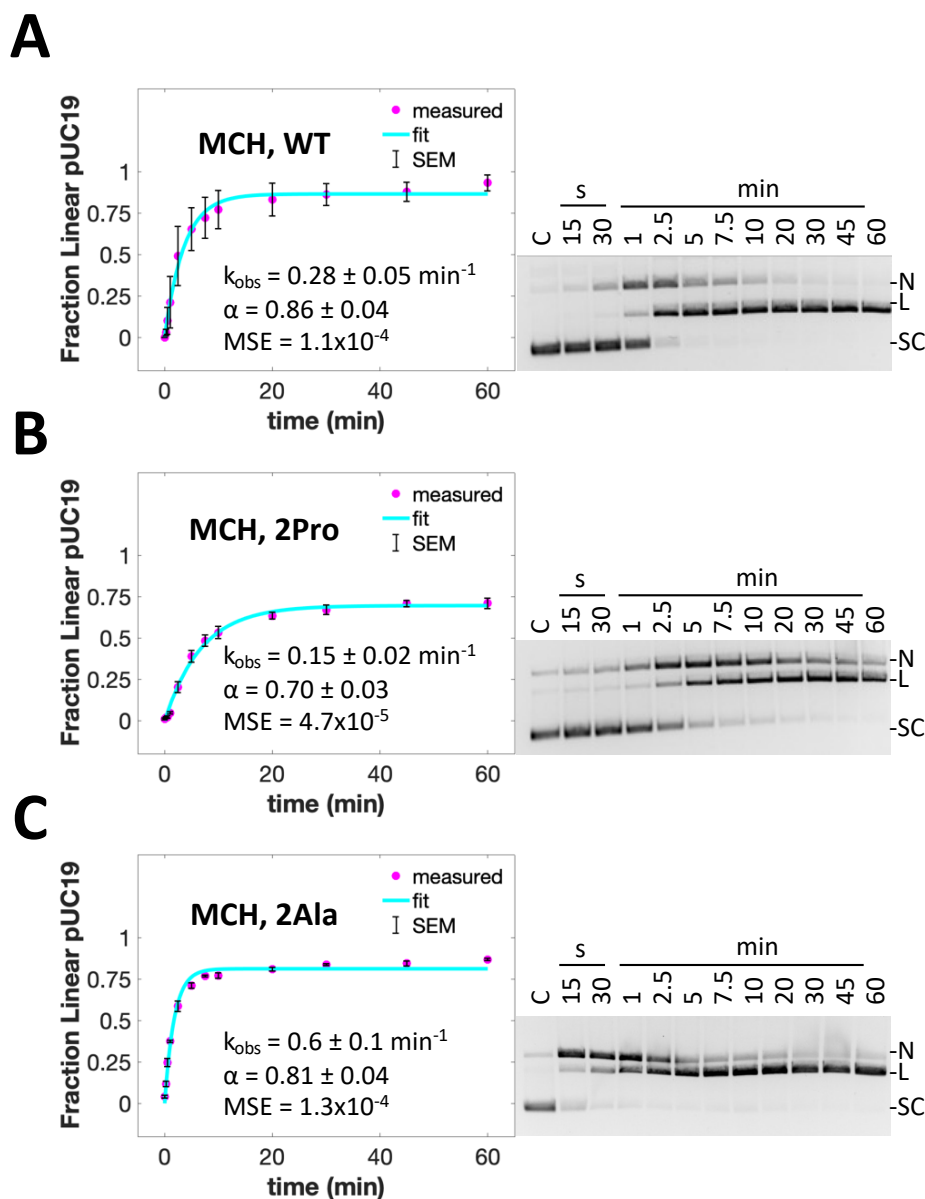
RNA-guided DNA cleavage time course assays with matched (MCH) and mismatched (MM5 and MM18) double stranded DNA substrates show that SpyCas9<sup>2Ala</sup> is faster than both SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> at linearizing each type of substrate (Figures 3.4, 3.5, 3.6 and 3.3A). In the case of MCH substrate, all three proteins displayed similar amount of efficiency of DNA linearization by the 60 min time-point, which is indicated by  $\alpha$  (70-85%), even though the rate constants ( $k_{obs}$ ) varied between the three; SpyCas9<sup>WT</sup>:  $0.28 \text{ min}^{-1}$ , SpyCas9<sup>2Pro</sup>:  $0.15 \text{ min}^{-1}$ , and SpyCas9<sup>2Ala</sup>:  $0.6 \text{ min}^{-1}$  (Figure 3.4). This indicates that a BH with a higher integrity increases the rate at which the enzyme cleaves DNA. In the case of MM5 substrate, there is a difference among both the active fractions and DNA cleavage rates of the three proteins (Figure 3.5). The  $\alpha$  and  $k_{obs}$  are as follows: SpyCas9<sup>WT</sup>:  $71\% \pm 2\%$ ,  $0.17 \pm 0.02 \text{ min}^{-1}$ ; SpyCas9<sup>2Pro</sup>:  $38\% \pm 2\%$ ,  $0.08 \pm 0.01 \text{ min}^{-1}$ ; and SpyCas9<sup>2Ala</sup>:  $77\% \pm 2\%$ ,  $0.25 \pm 0.03 \text{ min}^{-1}$ . The data indicates that SpyCas9<sup>2Pro</sup> has a compromised DNA linearization with a significant reduction in  $\alpha$  and  $k_{obs}$ . Interestingly, SpyCas9<sup>2Ala</sup> has a higher rate constant than SpyCas9<sup>WT</sup> to cleave MM5 DNA. This again shows that as the integrity of BH increases, so does the ability of Cas9 to rapidly cleave DNA.

Comparing the cleavage of MM18 substrate, all the three proteins linearize to the same extent ( $\sim 70\text{-}75\%$ ), but there is difference in their  $k_{obs}$  as follows:

SpyCas9<sup>WT</sup>:  $0.5 \text{ min}^{-1}$ , SpyCas9<sup>2Pro</sup>:  $0.6 \text{ min}^{-1}$ , and SpyCas9<sup>2Ala</sup>:  $1.5 \text{ min}^{-1}$  (Figure 3.6). While the data shows a similar trend for SpyCas9<sup>2Ala</sup> with the highest rate of linearizing MM18 DNA among the three proteins, the interesting difference is in the rate of SpyCas9<sup>2Pro</sup> to linearize MM18 DNA since SpyCas9<sup>2Pro</sup> linearizes MM18 faster than does SpyCas9<sup>WT</sup>. MM18 DNA linearization is the only instance

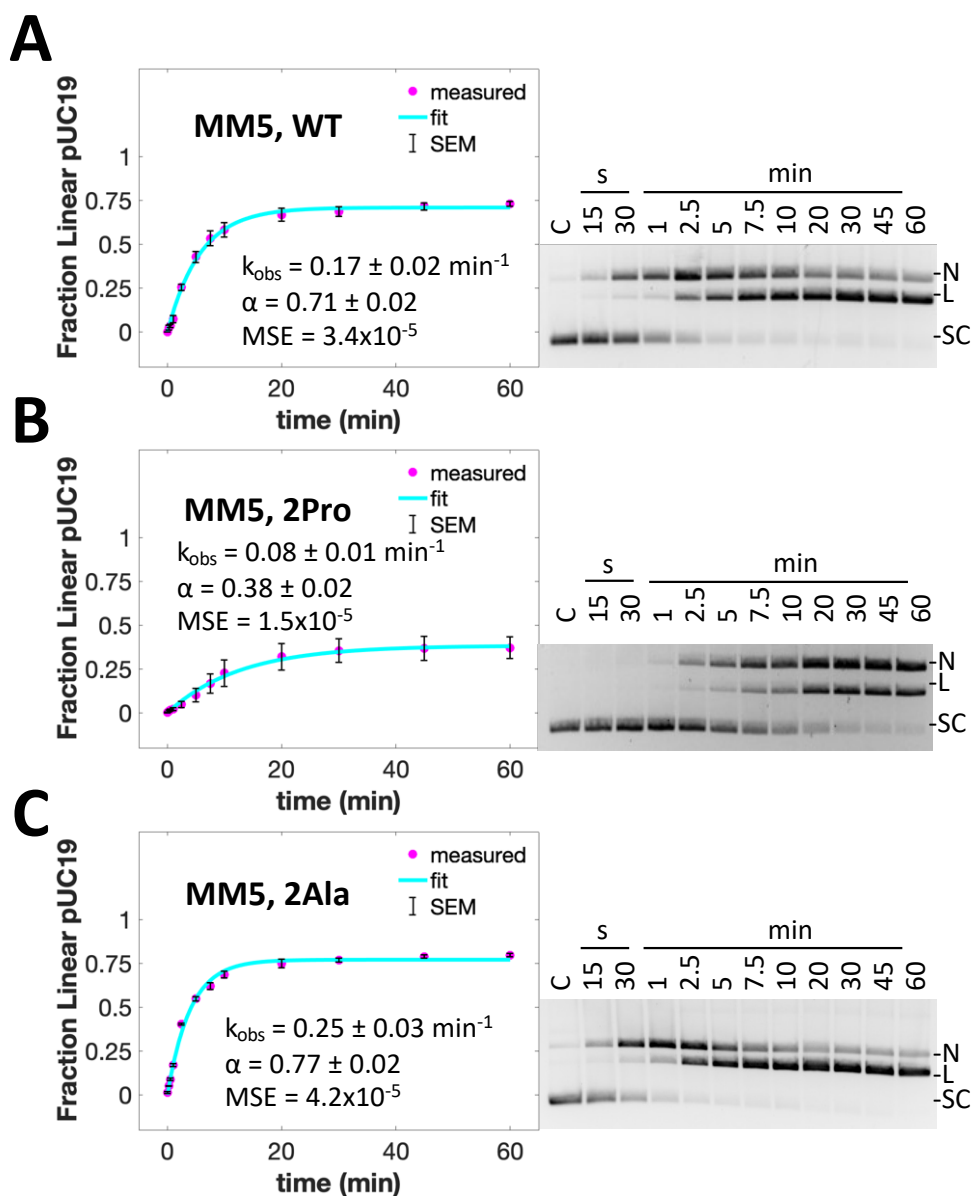


where SpyCas9<sup>2Pro</sup> displayed faster kinetics than SpyCas9<sup>WT</sup> in the different experiments that we conducted. This implies that while BH integrity is important in sensitivity towards PAM-proximal mismatches, it does not contribute towards recognizing DNA mismatches towards the PAM-distal end.

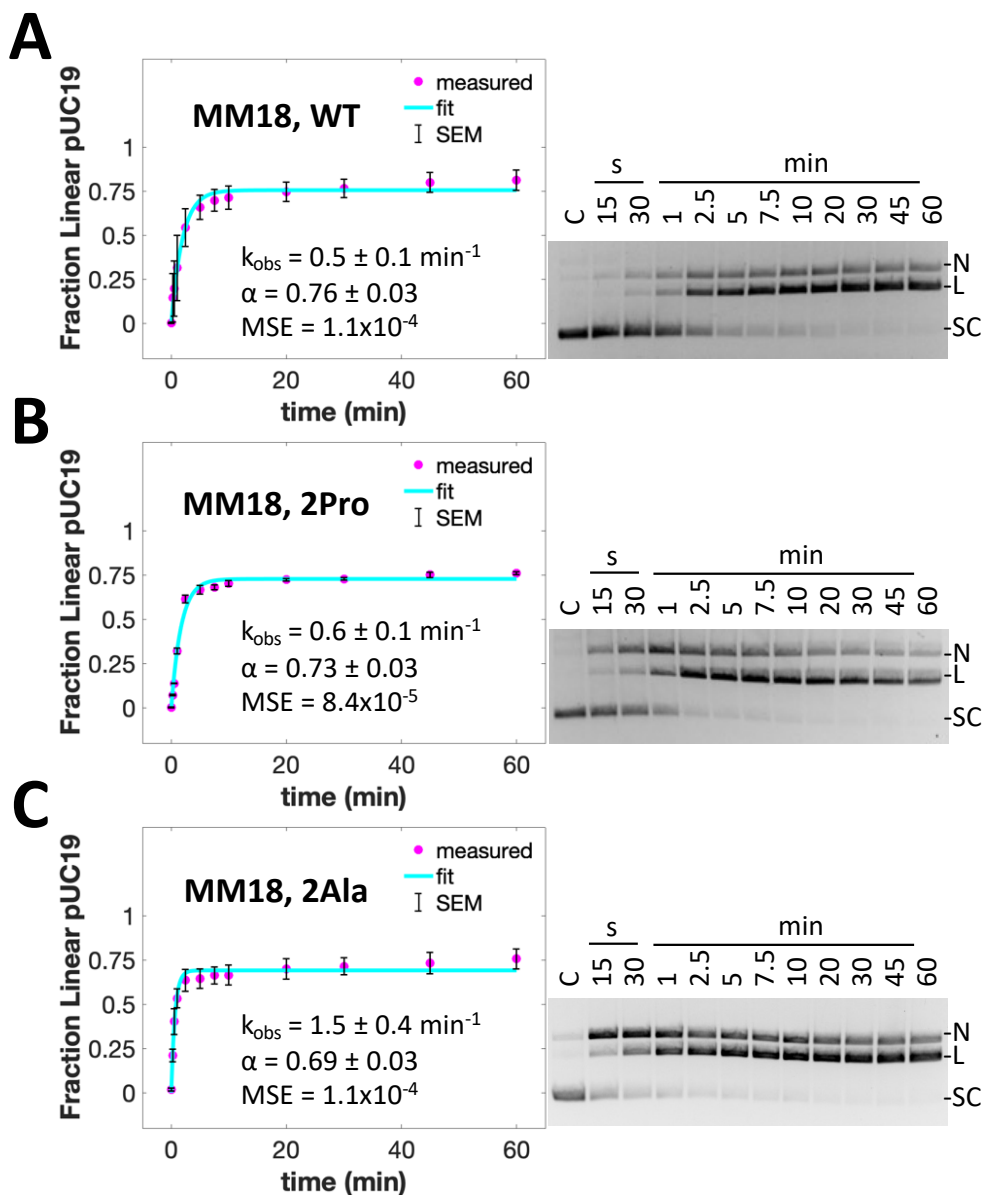


**Figure 3.4. Rate constants ( $k_{obs}$ ,  $\text{min}^{-1}$ ) and representative gels for RNA-guided cleavage of double stranded plasmid DNA substrate matched (MCH) to the guide RNA. Single exponential fit (cyan line, Equation 3.1) of average**

fraction of linearized double-stranded pUC19 DNA (Equation 3.1) plotted against time (t, min) for SpyCas9 **(A)** WT (SpyCas9<sup>WT</sup>), **(B)** 2Pro (SpyCas9<sup>2Pro</sup>), and **(C)** 2Ala (SpyCas9<sup>2Ala</sup>). Active fraction ( $\alpha$ ) is the fraction of DNA linearized as time approaches 60 min. The data points (magenta) are the average of three replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from the single exponential fit line. The error reported for  $k_{obs}$  is the 95% confidence interval. [N – Nicked, L – linear, SC – Supercoiled.]



**Figure 3.5. Rate constants ( $k_{\text{obs}}$ , min<sup>-1</sup>) and representative gels for RNA-guided cleavage of double stranded plasmid DNA substrate mismatched at position 5 (MM5) to the guide RNA.** Single exponential fit (cyan line, Equation 3.1) of average fraction of linearized double-stranded pUC19 DNA (Equation 3.1) plotted against time (t, min) for SpyCas9 **(A)** WT (SpyCas9<sup>WT</sup>), **(B)** 2Pro (SpyCas9<sup>2Pro</sup>), and **(C)** 2Ala (SpyCas9<sup>2Ala</sup>). Active fraction ( $\alpha$ ) is the fraction of DNA linearized as time approaches 60 min. The data points (magenta) are the average of three replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from the single exponential fit line. The error reported for  $k_{\text{obs}}$  is the 95% confidence interval. [N – Nicked, L – linear, SC – Supercoiled.]



**Figure 3.6. Rate constants ( $k_{obs}$ ,  $\text{min}^{-1}$ ) and representative gels for RNA-guided cleavage of double stranded plasmid DNA substrate mismatched at position 18 (MM18) to the guide RNA. Single exponential fit (cyan line, Equation 3.1) of average fraction of linearized double-stranded pUC19 DNA (Equation 3.1) plotted against time ( $t$ , min) for SpyCas9 **(A)** WT (SpyCas9<sup>WT</sup>), **(B)** 2Pro (SpyCas9<sup>2Pro</sup>), and **(C)** 2Ala (SpyCas9<sup>2Ala</sup>). Active fraction ( $\alpha$ ) is the fraction**

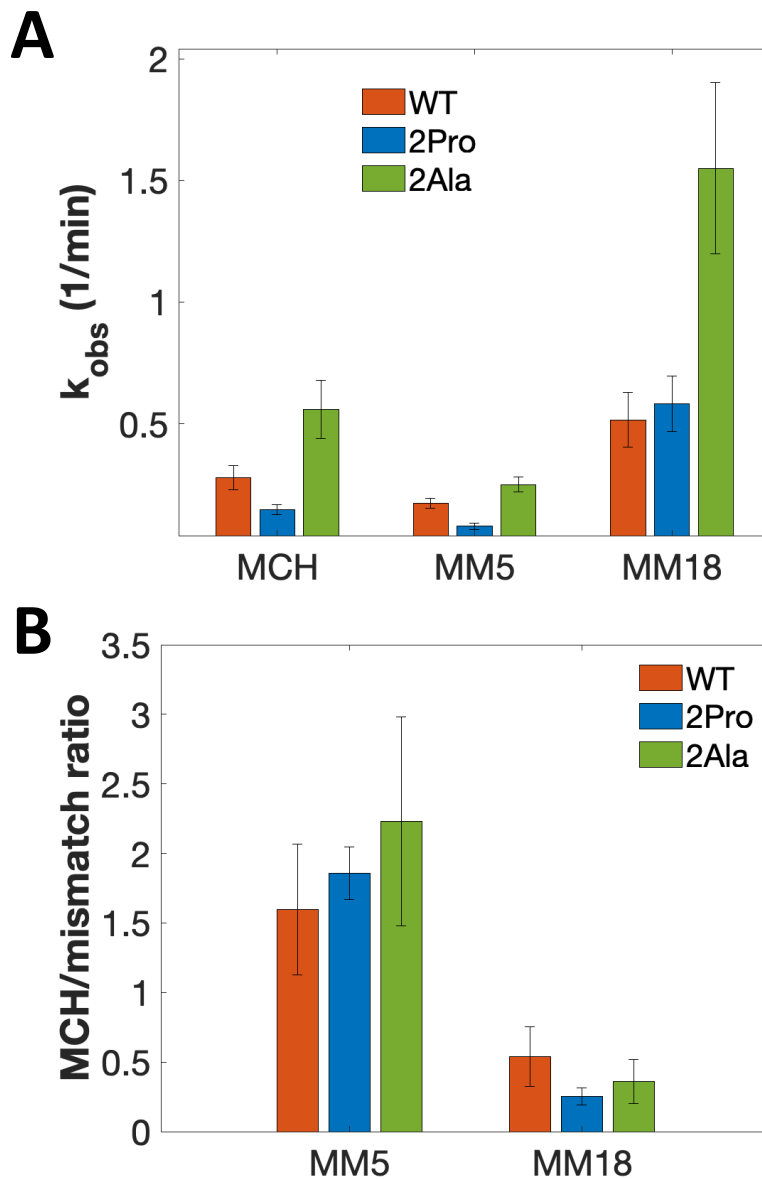
of DNA linearized as time approaches 60 min. The data points (magenta) are the average of three replications with standard error of the mean (SEM) shown as error bars. Mean square error (MSE) represents deviation of data points from the single exponential fit line. The error reported for  $k_{obs}$  is the 95% confidence interval. [N – Nicked, L – linear, SC – Supercoiled.]

### 3.5.2 Integrity of BH is essential for Cas9 to cleave PAM-proximal mismatch containing DNA

While a target DNA containing a PAM proximal mismatch (MM5) is cleaved more slowly by SpyCas9<sup>2Pro</sup> compared to that by SpyCas9<sup>WT</sup> and SpyCas9<sup>2Ala</sup> (Figure 3.7A), the total amount of DNA linearization is compromised only in SpyCas9<sup>2Pro</sup> (active fractions ( $\alpha$ ) for MM5: SpyCas9<sup>WT</sup>:  $0.71 \pm 0.02$ ; SpyCas9<sup>2Pro</sup>:  $0.38 \pm 0.02$ ; and SpyCas9<sup>2Ala</sup>:  $0.77 \pm 0.02$ ; Figure 3.5). To visualize how rapidly SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup>, and SpyCas9<sup>2Ala</sup> cleave MCH substrate relative to MM5, a ratio of the  $k_{obs}$  for MCH over that of MM5 was calculated for each protein (Figure 3.7B). A ratio greater than 1 indicates the protein cleaves MCH substrate more rapidly than it cleaves the mismatch substrate. Relative to SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup>, SpyCas9<sup>2Ala</sup> has a higher rate constant for cleavage of MM5 (Figure 3.7A); however, SpyCas9<sup>2Ala</sup> also cleaves MCH substrate most rapidly (Figure 3.7A), which shows that SpyCas9<sup>2Ala</sup> is a faster enzyme regardless of DNA mismatches (Figure 3.7A), but is still affected by PAM-proximal mismatches (Figure 3.7B).

Comparing similar features while cleaving target DNA containing a PAM distal mismatch (MM18) shows differences in the role of BH in discriminating PAM-distal mismatches. SpyCas9<sup>2Ala</sup> displays the fastest kinetics in cleaving MM18 compared to SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> (Figure 3.7A). Furthermore, SpyCas9<sup>2Ala</sup> is faster at cleaving MM18 than at cleaving MCH (Figure 3.7A). The total amount of DNA linearization is comparable for all the three proteins (active fractions ( $\alpha$ ) for MM18: SpyCas9<sup>WT</sup>:  $0.76 \pm 0.03$ ; SpyCas9<sup>2Pro</sup>:  $0.73 \pm 0.03$ ; and SpyCas9<sup>2Ala</sup>:  $0.69 \pm 0.03$ ; Figure 3.6). SpyCas9<sup>WT</sup> has a MCH/mismatch ratio

slightly higher than that of either SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> for MM18 (Figure 3.7B), so SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> prefer to cleave MM18 more rapidly and MCH more slowly (Figure 3.7B). The lack of a significant decrease in MM18 cleavage for SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> indicates that both these proteins do not sense PAM-distal mismatch through the BH. Together, these results indicate that the integrity of BH is important in enabling SpyCas9 to cleave PAM-proximal mismatch-containing DNA, but not PAM-distal mismatches. The proposed reason for this is that SpyCas9 first identifies the PAM and then the gRNA invades and begins hybridizing to the DNA substrate near the PAM, so the presence of a mismatch is most likely to disrupt the DNA cleavage mechanism if it is located near the PAM (Jinek et al., 2012). A BH that is intact provides extra stability to initiate gRNA hybridization with target DNA that has PAM-proximal mismatches. Since SpyCas9<sup>WT</sup> demonstrates the ability to slow down cleavage of both MM5 and MM18 substrates compared to the SpyCas9<sup>2Ala</sup> cleavage rates (Figure 3.7A), nature appears to have selected amino acids in the BH-loop such that there is some tolerance of DNA mismatches, such that they cannot be cleaved so quickly as was observed in SpyCas9<sup>2Ala</sup> which cleaved all DNA substrates at a faster rate than did SpyCas9<sup>WT</sup>.



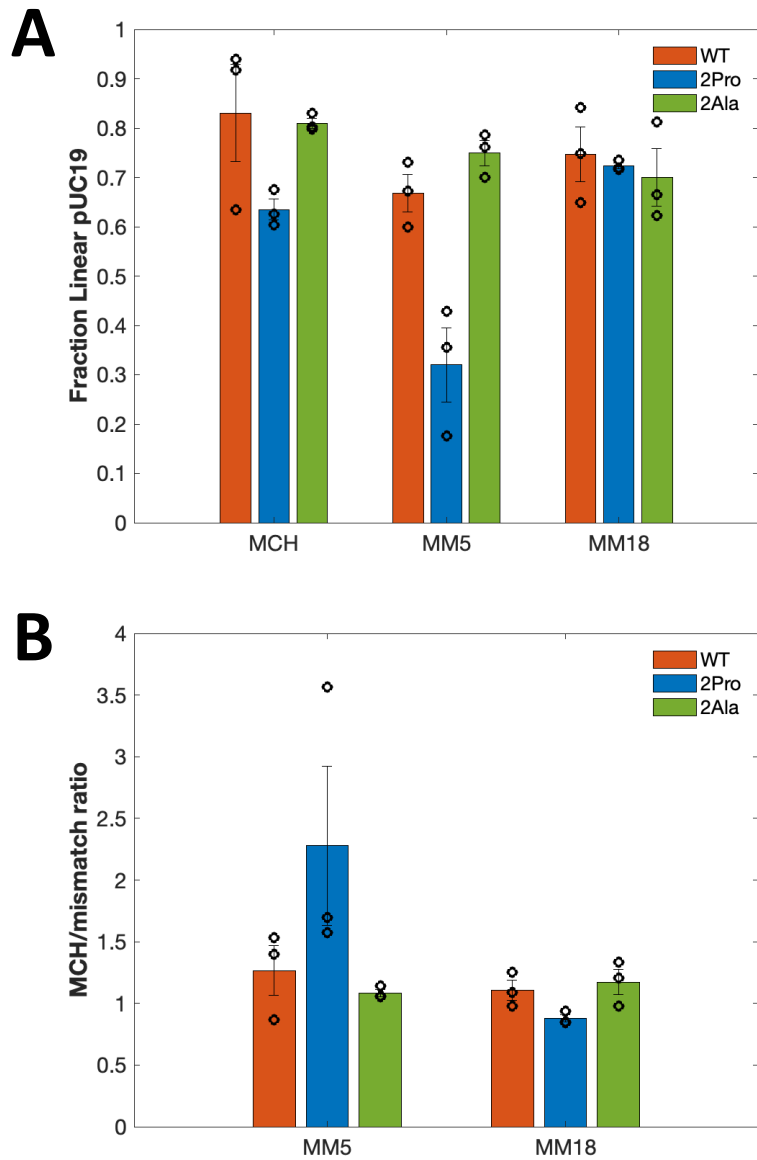
**Figure 3.7 Rate constant comparison for RNA-guided SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup>, and SpyCas9<sup>2Ala</sup> DNA cleavage of double stranded DNA plasmid matched (MCH) or mismatched (MM5 and MM18) to the guide RNA. (A) Bar graph of gRNA-dependent DNA degradation rate constants ( $k_{obs}$ ). The RNA-guided DNA cleavage assays were done using 5 mM metal ( $Mg^{2+}$ ). Error bars represent the 95% confidence interval. (B) Graph of the ratio of the  $k_{obs}$  of matched DNA linearization over the  $k_{obs}$  of mismatched DNA linearization (MCH/mismatch Ratio) such that a high ratio corresponds to greater selectivity for cleavage of the MCH substrate relative to the mismatch substrate. The error**

bars plotted are propagated error from the 95% confidence intervals of the rate constants.

### 3.5.3 DNA linearization is reduced in SpyCas9<sup>2Pro</sup> with PAM-proximal mismatches indicating better performance for gene editing

While rate constants provide a mechanistic implication of how BH affects the DNA cleavage process in the presence of DNA mismatches in different positions (PAM-proximal and PAM-distal), the most important aspect of a protein variant's efficient use as a gene editing tool is its ability to linearize mismatch containing DNA. We analyzed the total DNA linearization at the 20 min time-point (Figure 3.8) and made a table of the percent nicked *versus* percent linear values at the 20 min time-point as well (Table 3.2, Equations 3.2 and 3.3).





**Figure 3.8. Efficiency of RNA-guided SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup>, and SpyCas9<sup>2Ala</sup> DNA linearization for double stranded DNA plasmid (pUC19) matched (MCH) or mismatched (MM5 and MM18) to the guide RNA. (A)** Fraction Linear pUC19 is calculated according to Equation 3.1 using the 20 min time point from the time-course assays. **(B)** Graph of the ratio of Fraction Linear pUC19 (Equation 3.1) for matched DNA over that of mismatched DNA (MCH/mismatch Ratio) such that a high ratio corresponds to greater selectivity

for linearization of the MCH substrate relative to the mismatch substrate. Error bars represent the standard error of the mean (SEM) for three replications.

| Protein                 | DNA  | Percent linear | Percent nicked |
|-------------------------|------|----------------|----------------|
| SpyCas9 <sup>WT</sup>   | MCH  | 83% ± 10%      | 13% ± 7%       |
|                         | MM5  | 67% ± 4%       | 28% ± 2%       |
|                         | MM18 | 75% ± 6%       | 19% ± 6%       |
| SpyCas9 <sup>2Pro</sup> | MCH  | 63% ± 2%       | 32% ± 1%       |
|                         | MM5  | 32% ± 8%       | 48% ± 2%       |
|                         | MM18 | 72% ± 1%       | 25% ± 1%       |
| SpyCas9 <sup>2Ala</sup> | MCH  | 81% ± 1%       | 16% ± 1%       |
|                         | MM5  | 75% ± 3%       | 22% ± 2%       |
|                         | MM18 | 70% ± 6%       | 28% ± 6%       |

**Table 3.2. Comparison of amount of DNA linearization (final cleavage product in gene editing applications) and DNA nicking (intermediate) for different proteins and DNA substrates.** Percentages are calculated at the 20 min DNA cleavage time-point according to Equations 2.2 and 2.3.

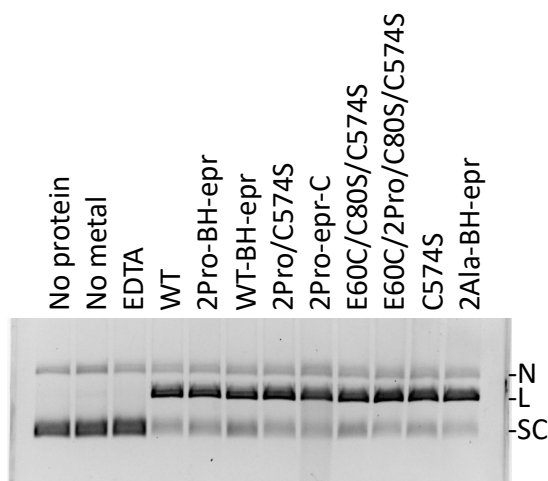
SpyCas9<sup>2Pro</sup> has a severe reduction in linearizing MM5 DNA compared to both SpyCas9<sup>WT</sup> and SpyCas9<sup>2Ala</sup> indicating its ability to limit cleavage of PAM-proximal mismatch containing DNA (Figure 3.8). To further characterize this, we analyzed the amount of nicked DNA produced by all these variants with different DNA substrates (Table 3.2). Nicked DNA is an intermediate in the DNA cleavage reaction by SpyCas9 where only one of the two DNA strands is cleaved by one of the endonucleases (HNH or RuvC). Nicked product is acted upon further by one of the endonucleases to produce a linear DNA. As expected based on a previous study (Babu et al., 2021), the percent of nicked product produced by SpyCas9<sup>2Pro</sup> is higher than what is produced by SpyCas9<sup>WT</sup> and SyCas9<sup>2Ala</sup> with MCH and especially with MM5 DNA (Table 3.2). This shows that SpyCas9<sup>2Pro</sup> stalls DNA

linearization especially when there is a PAM-proximal mismatch, while SpyCas9<sup>WT</sup> and SpyCas9<sup>2Ala</sup> are not able to control DNA cleavage similarly. This indicates the role of BH integrity in the coordinated DNA cleavage by HNH and RuvC. We previously showed that BH is needed for coordinated actions of HNH and RuvC, and this work provides a mechanistic role of BH integrity in the coordinated actions of HNH and RuvC. The proposed mechanism is that a sturdy BH, which has all intra-helical H-bonds within the helix being capable of coordinating cleavages by HNH and RuvC, will facilitate efficient DNA cleavage despite DNA mismatches. It should be noted that under gene editing conditions, only linearization at an undesired site contributes to off-target cleavage since nicked products are faithfully repaired by cellular machinery to the wild-type genomic content. Thus, SpyCas9<sup>2Pro</sup> is effective in reducing unwanted off-target effects when the mismatches are present at the PAM-proximal side compared to SpyCas9<sup>WT</sup> and SpyCa9<sup>2Ala</sup>.

#### 3.5.4 EPR variants are active for RNA-guided DNA cleavage

While our DNA cleavage assays provides mechanistic roles of BH integrity being an important contributor for DNA mismatch sensitivity, this is not direct measure of integrity of BH in the different Cas9 proteins. Towards this, we are using SDSL EPR to measure the distance between two nitroxide labels placed at 60<sup>th</sup> and 80<sup>th</sup> amino acids of SpyCas9 that flanks either side of the loop-to-helix transition of the BH. We have purified different variants that are amenable for EPR studies: SpyCas9<sup>WT-E60C</sup> (contains E60C and C574S substitutions, named SpyCas9<sup>WT-BH-epr</sup>), SpyCas9<sup>2Pro-E60C</sup> (contains E60C, L64P, K65P, and C574S substitutions, named SpyCas9<sup>2Pro-BH-epr</sup>), and SpyCas9<sup>2Ala-E60C</sup> (contains E60C, L64A, K65A, and C574S substitutions, named SpyCas9<sup>2Ala-BH-epr</sup>). A cysteine-free variant of SpyCas9<sup>2Pro</sup> (contains C80S and C574S substitutions, named SpyCas9<sup>2Pro-epr-C</sup>) and single cysteine control variants (SpyCas9<sup>2Pro/C574S</sup>, SpyCas9<sup>E60C/C80S/C574S</sup>, SpyCas9<sup>E60C/2Pro/C80S/C574S</sup>, and SpyCas9<sup>C574S</sup>) have also been constructed. To verify that the SpyCas9 variants are properly folded, *in vitro* DNA cleavage assays were used to see if their structure is correct to facilitate RNA-guided DNA

cleavage. All eight SpyCas9 variants linearized the DNA, indicating that they are properly folded and fit for studies using EPR (Figure 3.9).



**Figure 3.9.** RNA-guided DNA cleavage assay with SpyCas9 variants to be carried forward for further study using electron paramagnetic resonance. [N – nicked, L – Linear, SC – supercoiled.]

The EPR variants are being used by the Qin group at University of Southern California to conduct SDSL-EPR measurements. Current results show that in solution state, sgRNA binding to SpyCas9<sup>WT</sup> converts BH from a helix-loop-helix conformation to a stable contiguous helix as reported in the crystal structures (apo, PDB ID: 4CMP; (Jinek et al., 2014) and binary, PDB ID: 4XT0; (Jiang et al., 2015)). With SpyCas9<sup>2Pro</sup>, sgRNA binding also drives towards BH toward the full-helix conformation. However, the data reveal heterogeneous conformations of BH. This indicates that the integrity of the full-helix is compromised, consistent with the role of proline in destabilizing a helix. These initial studies indicate that the differences in BH integrity likely contributes to the mismatch discrimination that we reported previously (Babu et al., 2021). On-going studies are focused on performing similar experiments with SpyCas9<sup>2Ala</sup> and with different DNA.

### 3.6.0 Discussion and future directions

#### 3.6.1 BH integrity plays a more important role than sidechain interaction in DNA mismatch sensitivity

The purpose of this work was to distinguish between two possible mechanisms by which BH contributes to DNA mismatch sensitivity: integrity of the helix after loop-to-helix conversion of the BH *versus* sidechain interactions of the residues within the BH loop with other domains of SpyCas9 and nucleic acids. To dissect these roles, we compared the effect of two-proline and two-alanine substitutions within the loop region that converts into a contiguous helix in response to sgRNA binding. Our results indicate that two-alanine substitution rapidly cleaves DNA (Figure 3.7A), with reduced ability to refrain from completely linearizing PAM-proximal DNA mismatches like MM5 (Figure 3.8B). On the contrary, two-proline substitution reduces the rate of target DNA cleavage with MCH and MM5, but not with MM18 DNA (relative to the wild-type) (Figure 3.7A). SpyCas9<sup>2Pro</sup> is closer to reaching similar levels of linearization efficiency in MCH than it is in MM5 (Figure 3.8). This indicates that the disruption in helical integrity after the loop-to-helix transition in SpyCas9<sup>2Pro</sup> is partially restored by the complete complementarity between the RNA and DNA regions in MCH substrate. When the complementarity is disrupted in MM5, at a position very crucial for stabilizing R-loop, the impact of reduced BH integrity is not reparable. Our data show that integrity of BH, and not the loss of intra-protein and protein-nucleic acid electrostatic interactions, is the main contribution of the BH in PAM-proximal DNA mismatch sensitivity based on the increased cleavage rate of all DNA substrates by SpyCas9<sup>2Ala</sup> (Figures 3.7A) and lack of selectivity against complete linearization of MM5 (Figure 3.8B).

One of our hypotheses at the beginning of this study was that an efficient loop-to-helix transition would cause more robust (and less selective) DNA cleavage in the presence of a PAM-proximal mismatch like MM5. This can be evaluated in terms of linearization efficiency with different DNA substrates, including those with DNA mismatches. Linearization efficiency data comes from  $\alpha$  values which

are the fraction of DNA linearized as time approached 60 min (Figures 3.4-3.6, discussed in section 1.5.2 about Integrity of BH being essential for Cas9 to cleave PAM-proximal mismatch containing DNA) and from percent of DNA linearized at the 20 min time-point (which is when a significant amount of the DNA is expected to remain uncleaved, thus showing reaction progress at a point when one variant may be noticeably ahead of the other for the purpose of comparison of SpyCas9 variant linearization efficiency *in vitro*) (Figure 3.8). The 20 min time-point was analyzed in terms of MCH/mismatch ratios of DNA cleavage efficiency, and the trend expected based on helicity was observed when comparing the MCH/MM5 ratio of DNA linearization efficiency for SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> (Figure 3.8B). As expected based on the helicity trend, the MCH/MM5 DNA cleavage efficiency ratio was highest for SpyCas9<sup>2Pro</sup>, medium for SpyCas9<sup>WT</sup>, and lowest for SpyCas9<sup>2Ala</sup>. The fact that this was observed in Figure 3.8B for MM5 means that helix integrity, not the K65 intra-protein interaction, is the main contributor to increasing the MCH/MM5 linearization efficiency ratio. In other words, it is the compromised helicity that modulates the extent to which these BH substitution variants will carry MM5 cleavage to completion, which is defined as 100% linear product with no nicked intermediate remaining.

Analyzing the contribution of the side chain interaction towards DNA selectivity, in SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup>, there is a reduction in rate constant for cleaving MM5 compared to MCH DNA. SpyCas9<sup>2Ala</sup> displays a faster rate constant to cleave MM5 compared to MCH (Table 3.4). These results show that the increased BH integrity provided by the alanine substitutions, compared to the wild-type or proline substituent, enabled SpyCas9<sup>2Ala</sup> to suppress the negative impacts from losing K65 interactions towards DNA cleavage.

| Protein | DNA | k <sub>obs</sub> (min <sup>-1</sup> ) |
|---------|-----|---------------------------------------|
|---------|-----|---------------------------------------|

|                         |      |             |
|-------------------------|------|-------------|
| SpyCas9 <sup>WT</sup>   | MCH  | 0.28 ± 0.05 |
|                         | MM5  | 0.15 ± 0.02 |
|                         | MM18 | 0.6 ± 0.1   |
| SpyCas9 <sup>2Pro</sup> | MCH  | 0.17 ± 0.02 |
|                         | MM5  | 0.08 ± 0.01 |
|                         | MM18 | 0.25 ± 0.03 |
| SpyCas9 <sup>2Ala</sup> | MCH  | 0.5 ± 0.1   |
|                         | MM5  | 0.6 ± 0.1   |
|                         | MM18 | 1.5 ± 0.4   |

**Table 3.4.** Comparison of amount of rate constants ( $k_{obs}$ , min<sup>-1</sup>) for the different proteins and DNA substrates.

### 3.6.2 SpyCas9<sup>2Ala</sup>'s high DNA cleavage rate and comparable selectivity for on-target cleavage rate relative to SpyCas9<sup>WT</sup> may lend it to certain biotechnology applications

Although SpyCas9<sup>2Ala</sup> linearizes matched and mismatched DNA more rapidly than do SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> (Figure 3.7A, Table 3.4), the ratios of the matched over the MM5 and MM18 DNA cleavage rate constants show that its selectivity for rapidly cleaving matched DNA is comparable to, if not higher than, those of SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> (Figure 3.7B). Regarding Cas9-based biotechnology applications, there may be some instances where SpyCas9<sup>2Ala</sup> is preferred over SpyCas9<sup>WT</sup> because efficiency of DNA cleavage is more important than absolutely minimizing off-target DNA cleavage. For example, a diagnostic use where higher sensitivity to the presence of lower amounts of on-target DNA is a case where SpyCas9<sup>2Ala</sup> may perform more efficiently than SpyCas9<sup>WT</sup>. In the case of gene therapy applications, however, SpyCas9<sup>2Pro</sup> is probably preferred because minimizing off-target DNA cleavage is prioritized over efficiency of DNA cleavage to protect the genome from unintended damage.

3.6.3 Mechanistic insights into the effects of the two-proline and two-alanine substitutions on SpyCas9's on-target DNA linearization selectivity reveal more about how the actions of HNH and RuvC are coordinated for DNA linearization. There is evidence that the HNH domain is part of a signaling pathway within SpyCas9 that helps coordinate recognition of gRNA with RuvC activation (Chen et al., 2017; Dagdas et al., 2017; Nierzwicki et al., 2021b; Palermo et al., 2018a; Raper et al., 2018; Sternberg et al., 2015), and the loop-to-helix transition in BH may be part of this pathway (Babu et al., 2021). The fact that the two-alanine substitution, which is expected to increase the helicity of the BH, increased the DNA linearization rate indicates that the loop-to-helix-transition could be part of the signal transduced from the nucleic acid recognition lobe of SpyCas9 to the nuclease lobe for coordination of the HNH and RuvC to linearize DNA. It is possible that the ongoing EPR studies will show that the SpyCas9<sup>WT</sup>-gRNA complex has only a slightly increased population in the stable helical conformation as opposed to the loop form relative to SpyCas9<sup>2Pro</sup>-gRNA complex, and that the SpyCas9<sup>2Ala</sup>-gRNA complex has a much greater population in the helical form. If this helicity trend is observed in the EPR studies, then it will match the DNA linearization rate trend observed in this study, which is that SpyCas9<sup>2Pro</sup> linearizes DNA the slowest, SpyCas9<sup>WT</sup> being slightly faster, and SpyCas9<sup>2Ala</sup> being the fastest (Figure 3.7A). In this case, the transduction of a nuclease coordinating signal from the nucleic acid recognition lobe *via* the BH loop-to-helix transition would be a plausible mechanism describing how the two-proline substitution hinders the coordination of the HNH and RuvC for complete DNA linearization (Babu et al., 2021).

### 3.6.4 Future prospects

3.6.4.1 SDSL EPR studies with SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup> and SpyCas9<sup>2Ala</sup> and different DNA substrates to monitor the loop-to-helix transition

Our initial EPR studies with apo and binary (Cas9 bound to sgRNA) complexes of SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup> measuring the loop-to-helix transition show that the proline substitutions impact the stability of BH such that the loop within BH is



less able to form a helix upon sgRNA binding to SpyCas9 (data not shown). Further EPR studies will include testing SpyCas9<sup>2Ala</sup> and different DNA substrates to measure the stability of the BH under different conditions that will provide mechanistic insights on the DNA cleavage differences that we report in this study.

#### 3.6.4.2 Testing the effect of different amino acids in the loop region of the BH and its effect on the loop-to-helix transition

While this work tested only the effects of proline and alanine substitutions in the loop region of the BH, future studies testing different amino acids, especially assessing amino acids with different side chain properties (charge, hydrophobicity, size, aromaticity *etc.*) will not only provide mechanistic details of the loop-to-helix transition in SpyCas9's DNA cleavage and mismatch sensitivity, but will also have the potential to deliver newer variants that are more efficient for gene editing applications. A repertoire of variants with very efficient on-target DNA cleavage and severely compromised off-target DNA linearization will be valuable for genome applications.

## 4 Chapter 4: Predicting Potential Antiviral targets in Coronavirus-2 Spike Protein Using Sequence Conservation Analysis

### 4.1.0 Authorship

This is unpublished work done in collaboration with Richard Van, a member of Dr. Yihan Shao's laboratory at the University of Oklahoma Department of Chemistry and Biochemistry, and Dr. Fares Najar, former director of the University of Oklahoma Chemistry and Biochemistry Bioinformatics Core. Dr. Fares Najar obtained the spike protein sequences from NCBI, I organized the spike protein sequences from NCBI into the datasets that I then used for sequence conservation analysis, and Richard did the molecular dynamics and molecular docking analysis. Figures 4.6 and 4.7 were made by Richard.

### 4.2.0 Overview

The 2019 severe acute respiratory syndrome coronavirus 2 (CoV-2) outbreak re-opened spike protein structural characterization pipelines used to reveal the similarities and differences between CoV-2 and other coronavirus host cell entry mechanisms and re-intensified investigative efforts to elucidate the origins of novel coronaviruses. A promising antiviral strategy is to prevent contact between CoV-2 and host cell receptor proteins by targeting the spike protein. Spike protein contains conserved amino acid residues that are crucial for the host-cell entry mechanism and not amenable for frequent evolutionary changes due to their crucial functional roles. We propose that such conserved residues are potent antiviral targets and performed spike protein sequence conservation analysis to identify such residues present in spike protein. Molecular dynamics simulations were used to predict small molecule-binding pockets in the spike receptor binding domain (RBD), which is the domain of spike protein that interacts with the human cell receptor protein. Cross analysis of the conserved amino acids with the pockets in the spike RBD, which are predicted to bind potentially antiviral molecules, showed the presence of highly conserved amino

acids that are biochemically shown to be crucial for protein expression or its ability to bind to the host cell receptor protein in those pockets. To identify features of the CoV-2 spike protein that facilitated its spillover into the human population from zoonotic sources, we performed conservation analysis with zoonotic-isolate spike proteins. This analysis revealed groups of residues with similar functions that tend to be differentially conserved in human-isolate and zoonotic-isolate coronavirus spike proteins.

#### 4.3.0 Introduction

Outbreaks of novel coronaviruses (CoVs) in the human population occasionally occur as demonstrated by the appearance of severe acute respiratory syndrome coronavirus (SARS) in 2003, human coronavirus-NL63 (NL63), coronavirus-HKU1 (HKU1) in 2005, and Middle East Respiratory Syndrome coronavirus (MERS) in 2012 (Forcelloni et al., 2021). In December of 2019, the SARS-CoV-2 (CoV-2) species was detected for the first time in humans (Zhu et al., 2020). It spread world-wide, and masking and social distancing became common practices to reduce the spread of the virus. mRNA vaccines became available approximately one year after the outbreak (Hou et al., 2021), and various drugs have been repurposed to treat those infected with SARS-CoV-2 (Chakraborty et al., 2021). Research on how to prevent and treat SARS-CoV-2 is ongoing.

Despite coronaviruses causing severe respiratory outbreaks in the past, the scientific community did not have efficient models of viral evolution strategies to predict the severity of the CoV-2 pandemic. Specifically, analysis of the evolution of coronaviruses infecting humans and other animal hosts to compare sequence similarities that enable animals to serve as reservoirs in the emergence of new strains with pandemic capabilities is crucial in predicting viral strain development and also for strategic drug and antibody development plans.

We hypothesized that spike protein present on the surface of the coronavirus that directly interacts with the receptor protein ACE2 (Angiotensin-converting enzyme

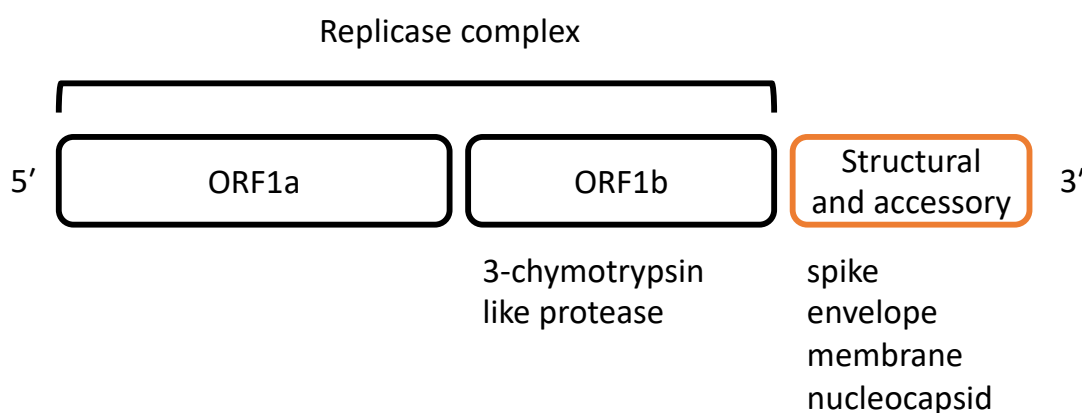
2) to gain entry into human cell will be an ideal candidate to monitor evolutionary changes to predict evolution of pandemic strains (Yan et al., 2020). Specifically, we propose that pandemic strains will possess mutations that enable tighter binding to human ACE2 protein and that some of these mutations enable binding of spike protein to ACE2 from new animal carriers (Ozono et al., 2021), potentially creating new strains and future pandemics. The emergence of the gamma, kappa, delta, and omicron CoV-2 variant has supported the hypothesis that CoV-2 spike mutations enhance human ACE2 binding (Zhang et al., 2022; Zhang et al., 2021). To observe which regions are conserved *versus* variable among spike protein sequences from SARS, MERS, and CoV-2 (in this case, the CoV-2 sequences available in Spring of 2019), we used phylogenetic analysis to map highly evolving and conserved regions of the spike protein. Further phylogenetic analysis was done with spike protein from coronaviruses isolated from both zoonotic and human hosts to identify residues that may be involved in the emergence of CoV-2 in the human population. In addition, through a collaborative work with the Shao lab, we identified compounds predicted to bind spike protein as potential antiviral drugs. We analyzed amino acid conservation at these drug binding pockets to rank the different pockets based on the level of amino acid conservation since a drug targeting a highly conserved pocket will have a long-lasting effect as an antiviral compound.

#### 4.3.1 Coronavirus taxonomy

Coronavirinae (a subfamily of Coronaviridae) contains four genera of viruses: two that infect mammals (alpha and beta) and two that primarily infect avian hosts (delta and gamma) (Zhou et al., 2021). There are currently seven coronavirus species that infect humans: human coronavirus-229E (229E), human coronavirus-OC43 (OC43), NL63, HKU1, SARS, MERS, and CoV-2 (Forcelloni et al., 2021). The alpha human-infecting coronaviruses cause common-cold like symptoms: 229E and NL63 (Zhou et al., 2021). Some beta coronaviruses cause severe symptoms in humans (SARS, MERS, and CoV-2), and others cause common cold-like symptoms (OC43 and HKU1) (Zhou et al., 2021).

#### 4.3.2 CoV-2 genes and proteins

CoV-2 is a positive sense, single-stranded RNA virus coding for the replicase complex and structural proteins (including spike, envelope, membrane, and nucleocapsid proteins), both of which resemble those of SARS, as well as a variety of accessory proteins that tend to vary among coronavirus species (Figure 4.1) (Mody et al., 2021). The host ribosomes translate the CoV-2 RNA into a polypeptide chain which then needs to be cleaved by 3-chymotrypsin like protease to produce individual, functional viral proteins (Mody et al., 2021).

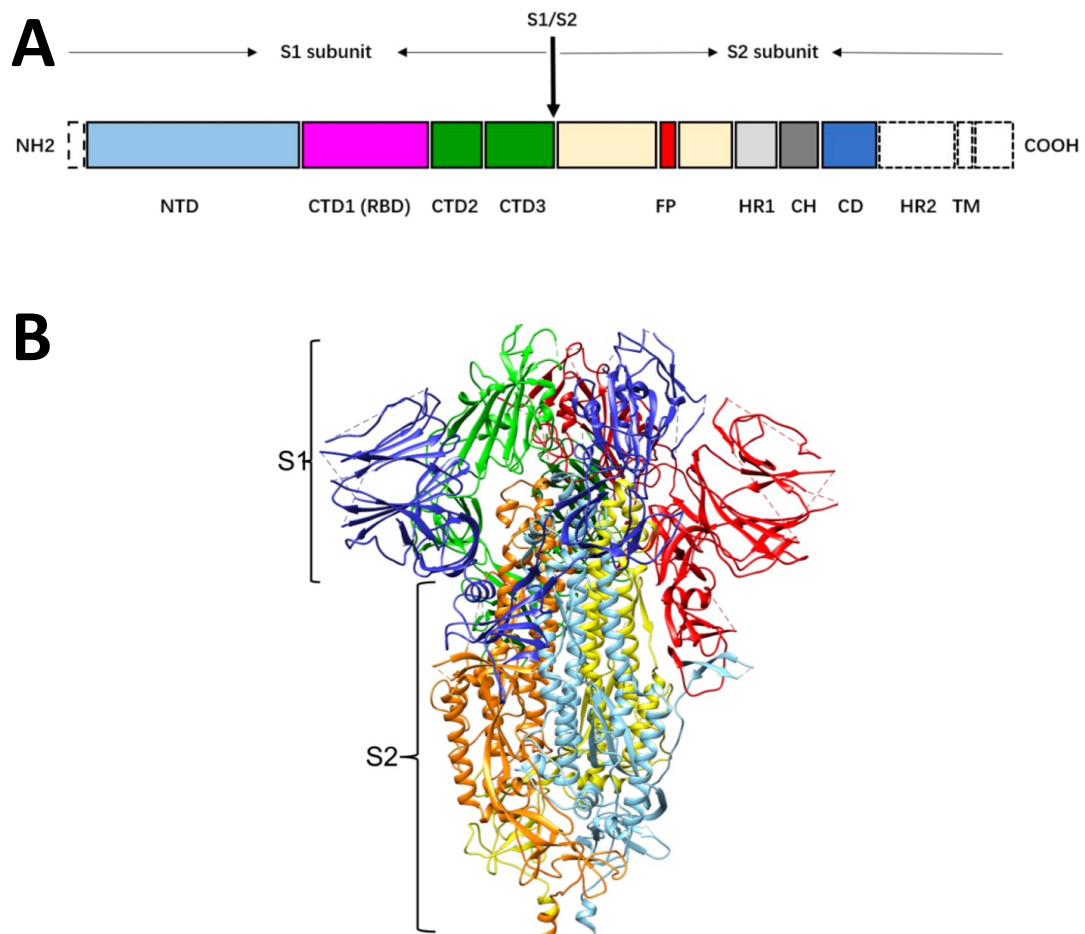


**Figure 4.1. Simplified depiction of the severe acute respiratory syndrome coronavirus 2 RNA genome** (Mody et al., 2021). 3-chymotrypsin like protease is encoded in ORF1b, which codes for part of the replicase complex. Spike is a structural protein encoded along with the other structural proteins: envelope, membrane, and nucleocapsid.

#### 4.3.3 Spike protein structure and function

Spike protein is a cell surface glycoprotein that is anchored into the membrane by a short transmembrane domain (Figure 4.2). A major part of the protein is extracellular. Spike protein occurs as a trimer on the viral envelope and appears like a three-pronged grapple hook (Figure 4.2B). Within each monomer, the S1 subunit contains the Receptor Binding Domain (RBD) the function of which is to bind the host receptor protein (ACE2), and is similar to the prongs of the grapple

hook (Wrapp et al., 2020). The S2 subunit contains the fusion components (fusion peptide and heptapeptide repeats) (Huang et al., 2020b; Wrapp et al., 2020). The fusion peptide inserts into the host cell, and the heptapeptide repeats help draw the viral particle near to the host cell membrane for fusion ultimately resulting in introduction of the coronavirus RNA into the host cell (Ashour et al., 2020; Huang et al., 2020b; White et al., 2008). Spike monomers can be in the closed conformation where the RBD in each of the three monomers in a trimer stays closed, which enhances spike protein stability and evasion of the host-immune system. During binding to ACE2, one of the RBD in the trimer acquires an open conformation where the RBD residues are positioned to facilitate receptor binding (Shang et al., 2020; Wrapp et al., 2020), which is the initial attachment point for the virus as it begins infecting the host cell.



**Figure 4.2. Domain architecture of the CoV-2 spike protein. (A)** The domain are as follows: NH<sub>2</sub>, N-terminus; NTD, N-terminal domain; CTD1, C-terminal domain 1; RBD, receptor binding domain that interacts with the human-receptor protein ACE2 (Angiotensin-converting enzyme 2); CTD2, C-terminal domain 2; CTD3, C-terminal domain 3; S1/S2, the interface between the S1 subunit (host receptor binding subunit) and the S2 subunit (subunit that mediates fusion with the host cell); S1/S2, the protease cleavage site needed for the transition from the host receptor protein binding phase to the host cell fusion phase; FP, fusion peptide that anchors in the host cell and works alongside the heptad repeat domains to draws it near to the viral particle; HR1, heptad repeat 1; CD, connector domain; HR2, heptad repeat 2; TM, transmembrane domain that is anchored in the viral particle membrane; COOH, C-terminus. Adapted from Wang

*et al.* (2020) (Wang and Xiang, 2020). **(B)** Extracellular domain of spike protein with three monomers shown. S1 trimers: red, green, blue. S2 trimers: orange, cyan, yellow. PDB ID: 6VXX (Walls et al., 2020).

One determinant of coronavirus host specificity is spike:host receptor-binding complementarity. In the case of CoV-2, the host receptor protein is ACE2 (Zhou et al., 2020). Most vertebrates have ACE2 on the surface of their cells; however, CoV-2 spike cannot bind all ACE2s. For example, CoV-2 can bind human-ACE2, civet-ACE2, swine-ACE2, and Chinese horseshoe bat-ACE2, but not mouse-ACE2 (Zhou et al., 2020). The lack of CoV-2 spike protein binding to mouse ACE2 can be explained by residues in the mouse ACE2 that are different from those in other ACE2s: N30 (which is D30 in the human ACE2 and interacts with K417 in CoV-2 spike), F83 (which is Y83 in the human ACE2 and interacts with F486 and N487 in CoV-2 spike), and H353 (which is K353 in the human ACE2 and interacts with G496, G502, Y505, Q498, and N501 in CoV-2 spike) (Zhao et al., 2020). SARS also uses ACE2 as a host receptor protein (Li et al., 2003), but MERS uses CD26 (otherwise known as DPP4) (Lu et al., 2020; Raj et al., 2013). Since humans have both ACE2 and CD26 on the surface of their cells (Lu et al., 2013), humans can be infected by SARS, MERS, and CoV-2.

The objective of this research is to identify highly conserved antiviral target sites in spike protein that will disrupt the host receptor protein binding and membrane fusion mechanisms. This will save the host cell from ever being invaded by the viral genetic material. Our hypothesis is that conserved spike protein residues exist that are crucial for the host-cell entry mechanism, and that these residues are mutated at slower rates despite the high rate of mutation that coronaviruses undergo to be able to infect hosts. We propose that targeting drug-binding pockets in spike protein with highly conserved amino acids will maintain potency against future viral strains with variations in spike protein structure. Another goal of our research is to give insight into the key adaptations in spike protein that



contribute to spillover of coronaviruses from zoonotic sources into the human population.

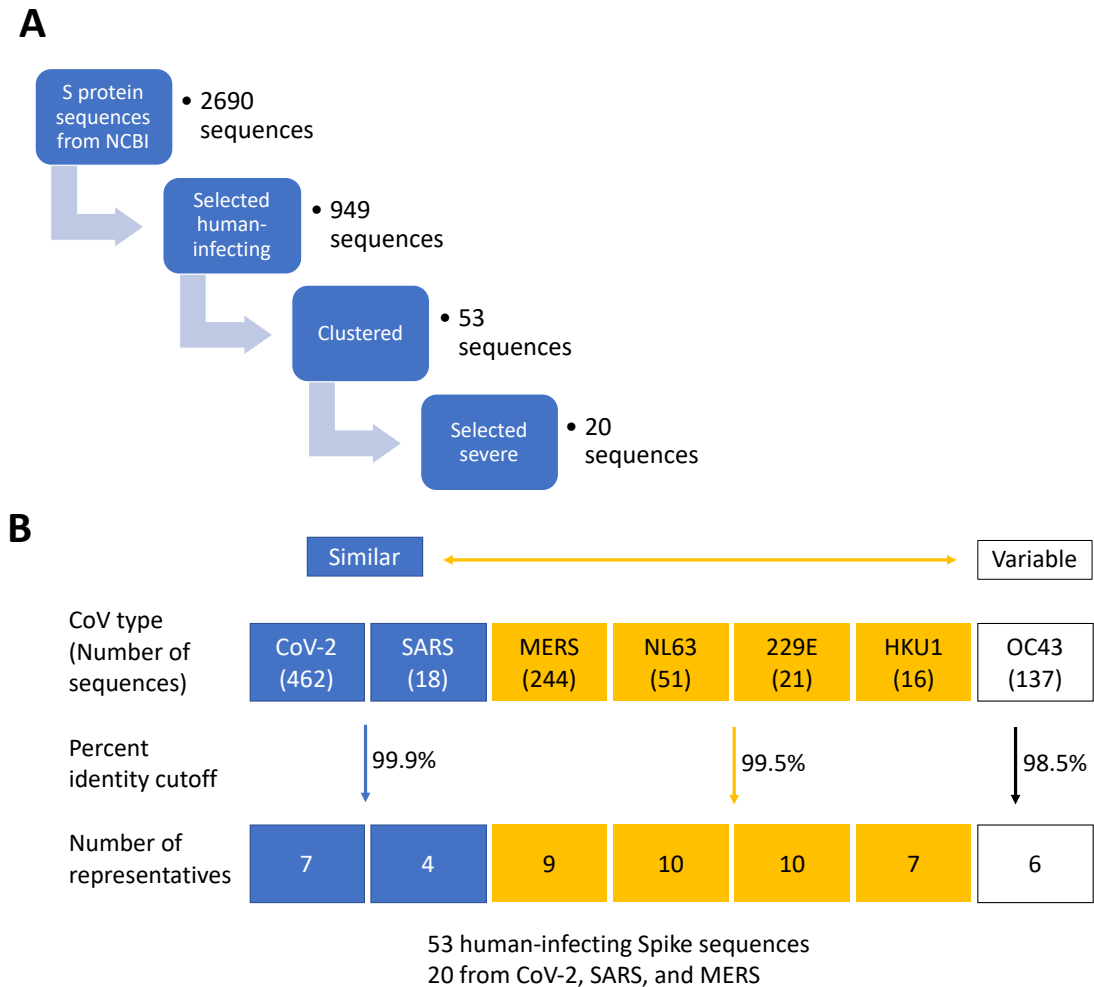
#### 4.4.0 Methods

##### 4.4.1 Curating the spike protein sequence database

Spike protein sequences were obtained from the National Center for Biotechnology Information (NCBI) in the year 2020 using the search term “S glycoprotein AND coronavirus[organism]” to download genbank files from which a spike protein amino acid dataset containing 2690 spike sequences with coronavirus species and host information in the headers was constructed (Figure 4.3A). From this dataset, three spike protein sequence datasets were obtained for conservation analyses by sorting the spike sequences by coronavirus species using information in the genbank files, selecting representative spike sequences from clusters meeting certain sequence identity thresholds, and selecting spike sequences from coronaviruses infecting certain zoonotic hosts: severe, human-infecting coronavirus spikes (20 proteins with approximately equal representation from SARS, MERS and CoV-2), human-isolate only (53 proteins with approximately equal representation from SARS, MERS, CoV-2, 229E, OC43, NL63, and HKU1), and zoonotic-isolate added (the 53 proteins from the human-isolate only analysis plus 10 from zoonotic isolates). Clustering was used to obtain representative spike sequences because many of the spike sequences were identical, and including multiple identical protein sequences in the conservation analysis would artificially increase its weight in the conservation analysis.

Cd-hit (Li and Godzik, 2006) was used to cluster and obtain 7 representatives from 462 CoV-2 spikes and 4 representatives from 18 SARS spikes at a 99.9% sequence identity threshold, 9 representatives from 244 MERS spikes and 10 representatives from 51 NL63 spikes and 10 representatives from 21 229E spikes and 7 representatives from 16 HKU1 spikes at a 99.5% sequence identity threshold, and 6 representatives from 137 OC43 spikes at a 98.5% sequence

identity threshold (Figure 4.3B). The 7 CoV-2, 4 SARS, and 9 MERS spikes makes up the severe, human infecting dataset of 20 spike sequences. That dataset of 20 sequences plus the 33 sequences from the common cold-like CoVs (NL63, 229E, HKU1, and OC43) make up the human-isolate only dataset of 53 spike sequences (Figure 4.3).



**Figure 4.3. Workflow describing how the severe, human infecting dataset of 20 spike sequences and human-isolate only dataset of 53 spike sequences were obtained from the 2960 spike sequences downloaded from NCBI. (A)** The 2690 spike sequences from NCBI are shown at the top of the process, the 949 spike sequences from human-infecting coronaviruses (SARS, MERS, CoV-2, 229E, OC43, NL63, and HKU1) are shown next, and the 53 representatives

from the human-infecting coronaviruses selected based on sequence identity clustering (shown in panel B of this figure) are shown third. Finally, the 20 spike sequences from the severe, human-infecting coronaviruses (SARS, MERS, and CoV-2) are shown fourth. **(B)** The clustering process used to select the 53 representatives from the 949 spike sequences from the seven human-infecting coronaviruses. Clustering was done using the program cd-hit to group the proteins based on the sequence identity cutoff shown next to each downward arrow (Li and Godzik, 2006). The seven human-infecting coronaviruses are named in the top row of boxes and have the number of sequences before clustering in parentheses below them. CoV-2 and SARS share a blue color because they were clustered at a 99.9% sequence identity cutoff; MERS, NL63, 229E, and HKU1 share a yellow color because they were all clustered using a 99.5% sequence identity cutoff; and OC43 is the only one colored white because it is the only one clustered at a 98.5% sequence identity cutoff. Since the spike proteins from the coronaviruses on the left had to have a higher percent identity cutoff to obtain a comparable number of clusters as the ones on the right with a lower percent identity cutoff, the spike proteins from the coronaviruses on the left (those from CoV-2 for example) were more similar to one another while those on the right (those from OC43 for example) were more different from one another. The boxes below the downward arrows show how many representative spike sequences were obtained from each human-infecting coronavirus species after clustering.

Details on the selection of spike protein sequences from zoonotic-isolates are as follows. A dataset of 82 zoonotic-isolate CoV nucleotide sequences had already been compiled by Gussow et al. (2020) (Gussow et al., 2020). To obtain the zoonotic-isolate spike sequences that are most similar to the human-isolate spike sequences used in our study, the zoonotic-infecting coronavirus spike dataset compiled by Gussow *et. al.* was submitted as a nucleotide fasta subject to tblastn (Altschul et al., 1990) along with a protein sequence query of the 53 spike sequences from the human-infecting coronaviruses. The output yielded the CoV

genome accession numbers corresponding to the top hit from each human-isolate spike sequence. The genome accession numbers were de-duplicated resulting in 10 zoonotic-isolate CoV genomes which were used to obtain the 10 spike protein sequences comprising the zoonotic-isolate spike dataset.

#### 4.4.2 Measuring spike protein conservation

Spike protein conservation was measured by constructing a multiple sequence alignment (MSA) using clustal omega (Sievers and Higgins, 2014) and then submitting that MSA to a program called [TRAnsient Pockets in Proteins with Druggability Calculation \(TRAPP\)](#) to score conservation of the MSA positions using Jensen–Shannon divergence (Capra and Singh, 2007). CoV-2 spike accession number QHO62112 (the Wuhan strain) was used as the reference sequence when reporting CoV-2 spike residues and positions.

#### 4.4.3 Identifying small-molecule-binding pockets in spike protein using molecular dynamics simulations and molecular docking

We used a molecular dynamics structure that was available freely to the scientific community from the D E Shaw Research group through a [Creative Commons Attribution 4.0 International Public License](#). This structure was modeled based on SARS CoV-2 spike structures PDB ID: 6VXX (Walls et al., 2020) and 6VW1 (Liu et al., 2021) and only contained the RBD and a short region of S1 fusion region. Two  $\mu$ s trajectories at 310 K and NPT ensemble were performed by D E Shaw Research group. Other parameters were AMBER ff99SB-ILDN force field (Lindorff-Larsen et al., 2010), TIP3P model for water (Jorgensen et al., 1983), GLYCAM06 for glycosylated asparagine (Kirschner et al., 2008), and generalized AMBER ff for small molecules (Wang et al., 2004). Frames for molecular docking analysis were taken every 0.2  $\mu$ s, which resulted in 11 snapshots representing different conformations of the protein. Molecular docking was done using each of the 11 frames with AutoDock Vina (Eberhardt et al., 2021) (Trott and Olson, 2010), which was also used to predict binding energies. 5152 compounds were

screened from the NCI diversity Set V compound library which contains FDA-investigational drugs.

#### 4.4.4 Analyzing biochemical data on spike expression and spike:ACE2 binding

Table S2 in Starr et al. 2020 reports the effect of mutating RBD residues to every other amino acid such that more negative values indicate lower expression of spike or weaker affinity of spike for ACE2 for each RBD substitution (Starr et al., 2020). For Figures 3.6-3.9, the spike expression and spike:ACE2-binding score for each residue was obtained by taking the average spike expression or spike:ACE2-binding score value of all amino acid substitutions for that residue in Table S2 in Starr et al. 2020 and then normalizing those values by dividing each by the average spike expression or average spike:ACE2-binding score value with the highest absolute value. This is shown in the steps below using the spike expression score for example:

1. Column 12 of Table S2 in Starr *et al.* 2020 contains the spike expression score for each variant (20 amino acid substitutions and one truncation variant) for each residue in the spike RBD (positions 331-531 according to Starr *et al.* 2020). For example, if substitution of an amino acid residue in the spike RBD to alanine decreased spike protein's ability to express, then that substitution for that residue would be given a positive score. If substitution of that same residue in the spike RBD to aspartate increased spike protein's ability to express, then that substitution for that residue would be given a negative score. This positive score for alanine and negative score for aspartate would be two of the 21 spike expression score values for that spike protein residue, so this analysis of the biochemical data takes information about both deleterious and compensatory substitutions. The average score (of all 21 variants) was obtained for each spike RBD residue to determine whether substituting this residue was more or less deleterious or compensatory on average compared to the substitution of other spike RBD residues. Taking position 331 for example, the average score for position 331 was obtained by

summing the values in column 12 of the 21 variants of residue position 331 and dividing that sum by 21.

2. This was repeated for each RBD residue position (positions 331-531) such that each position had an average spike expression score. For position 331, for example, the average spike expression score was -0.73619.
3. The average spike expression score with the highest absolute value - 2.80238) belonged to position V433. To normalize all the average spike expression scores to a value of 1, each average spike expression score was divided by -2.80238. For residue 331, this yields a normalized average expression score of 0.262702.
4. The resulting normalized average spike expression scores (for residue positions 331-531) were used in the analyses in this research.

This makes RBD residues in Figures 3.6-3.9 with positive scores near 1 the most crucial for spike expression or spike:ACE2 binding.

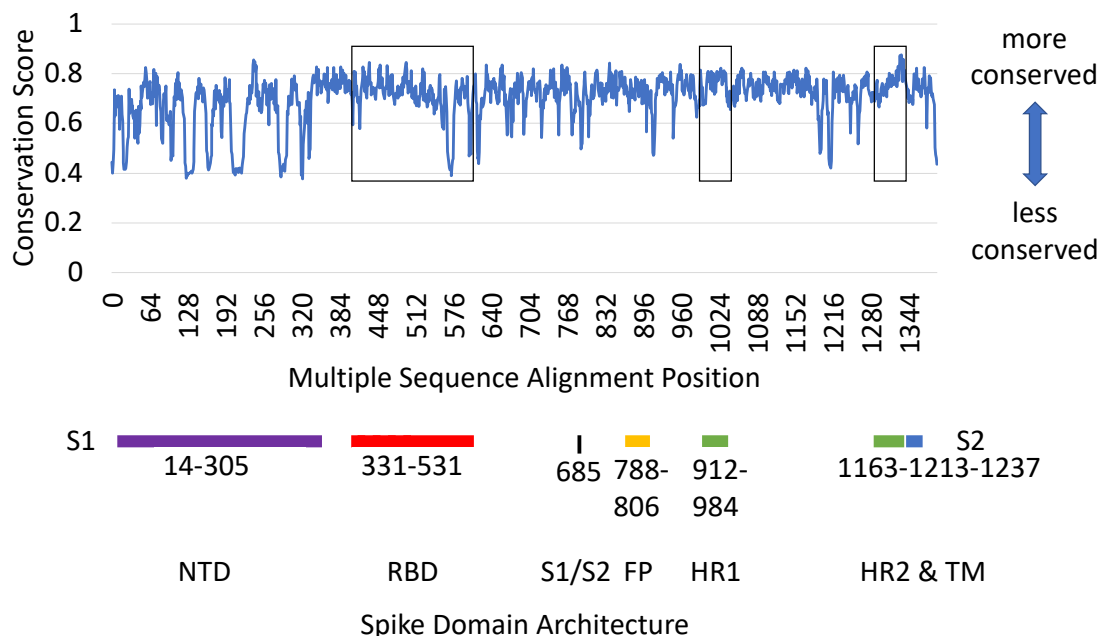
#### 4.5.0 Results

##### 4.5.1 Spike protein regions vary in amino acid conservation across SARS, MERS, and CoV-2 strains

To visualize if there are any regions of spike protein that stand out as highly conserved or variable, a multiple sequence alignment of spike proteins from SARS, MERS, and CoV-2 was made and the amino acid residue positions were scored for conservation using a Jensen–Shannon divergence-based algorithm that quantifies the difference between the amount of amino acid distribution present in each column of the MSA (Capra and Singh, 2007). The Jensen–Shannon divergence-based algorithm uses amino acid identity (not similarity in function as is used by the BLOSUM matrix method) as the criterion for conservation and is as accurate as other more computationally intensive methods, such as those that build phylogenetic trees, at identifying sequence conservation (Capra and Singh, 2007). SARS, MERS, and CoV-2 spike sequences were selected for analysis because they cause severe symptoms in

humans. There are other CoVs such as 229E, OC43, NL63, and HKU1 that cause mild symptoms in humans, and still others that infect zoonotic hosts only. Since the main application of this study is identification of potential antiviral drug binding pockets that are highly conserved, we prioritized viruses that cause severe symptoms in humans to address the most pressing health concern.

The conservation scores for each position in the multiple sequence alignment that contained a CoV-2 spike residue was plotted in Figure 4.4. The MSA contained spike sequences from the 7 CoV-2, 4 SARS, and 9 MERS spike sequences making up the dataset of 20 spike sequences from the severe, human-infecting coronaviruses, which explains why the conservation score tends to be approximately between 0.4 and 0.8 since the sequences from the same CoV species are anywhere from 98.5% to 99.9% similar as shown in Figure 4.3 (leading to the minimum score of ~0.4) and the sequences from different species are variable (77.38% similarity between SARS and CoV-2 and 31.98% similarity between MERS and CoV-2 (Agrawal et al., 2021); leading to a maximum score of ~0.8). Figure 4.4 shows that the N-terminal domain (NTD, CoV-2 spike residue positions 14-305) has multiple regions that are highly variable, the receptor binding domain (RBD, CoV-2 spike residue positions 331-531), fusion peptide (FP, CoV-2 spike residue positions 788-806), and heptapeptide repeat 1 (HP1, CoV-2 spike residue positions 912-984) regions are mostly uniformly conserved, and that the heptapeptide repeat 2 (HP2, CoV-2 spike residue positions 1163-1213) region has some highly conserved residues near the transition to the transmembrane (TM, CoV-2 spike residue positions 1214-1237) domain (Figure 4.2A). The RBD and heptapeptide repeat regions are of interest because their roles in the CoV-2 infection mechanism is well known, and structural studies have confirmed that the RBD is conserved while the NTD is variable for CoV-2 variants (Zhang et al., 2021). Table 4.1 shows the top five most conserved amino acid residues in these regions.



**Figure 4.4. Conservation scores from a Jensen–Shannon divergence-based algorithm used to analyze the multiple sequence alignment of a SARS, MERS, and CoV-2 spike protein dataset collected from NCBI in 2020.** The program that uses this Jensen–Shannon divergence-based algorithm to score the conservation is called TRAnsient Pockets in Proteins with Druggability Calculation (TRAPP) (Capra and Singh, 2007). The position in the multiple sequence alignment (MSA) is shown on the x axis, and below that is shown the domain architecture of CoV-2 spike (protein accession number QHO62112) with its amino acid numbers (Huang et al., 2020b). NTD is N-terminal domain, RBD is receptor binding domain, S1/S2 is the interface between the S1 and S2 regions of spike protein, FP is the fusion peptide, HR1 and HR2 are the heptapeptide repeat domains, and TM is the transmembrane domain. The black boxes show the position of the CoV-2 RBD and heptapeptide repeat domains in the multiple sequence alignment.

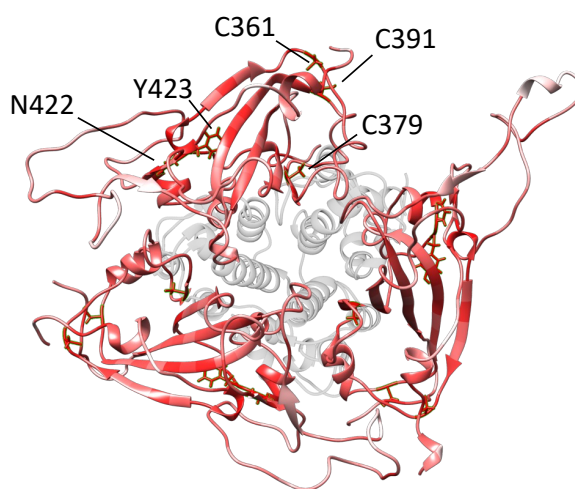
| Receptor Binding Domain |                    | Heptapeptide Repeat Domains |                    |
|-------------------------|--------------------|-----------------------------|--------------------|
| Residue                 | Conservation Score | Residue                     | Conservation Score |



|      |         |       |         |
|------|---------|-------|---------|
| Y423 | 0.84532 | W1212 | 0.87325 |
| C361 | 0.84416 | Y1209 | 0.83010 |
| C379 | 0.83596 | P1213 | 0.82929 |
| N422 | 0.83355 | K1211 | 0.82530 |
| C391 | 0.83311 | N955  | 0.82129 |

**Table 4.1. Top five conserved amino acid residues from receptor binding domain and heptapeptide repeat domains using the conservation analysis of the severe, human-infecting coronavirus (SARS, MERS, and CoV-2) spike proteins dataset from 2020.** Receptor binding domain is residues 331-531 and heptapeptide repeat domains are residues 912–984 and 1163–1213.

The conserved RBD residues were especially of interest because these are the ones that make initial contact with the host cell receptor protein, which is ACE2 for CoV-2. This makes the RBD an appealing drug target since disrupting its function can preclude contact with host cells. Figure 4.5 was made to visualize the locations of the conserved RBD residues on the CoV-2 spike structure. The structural representation shows that four out of the five most conserved residues (N422, Y423, C361, and C391) are located in loop regions and the remaining residue (C379) is located in a beta sheet (Figure 4.5).

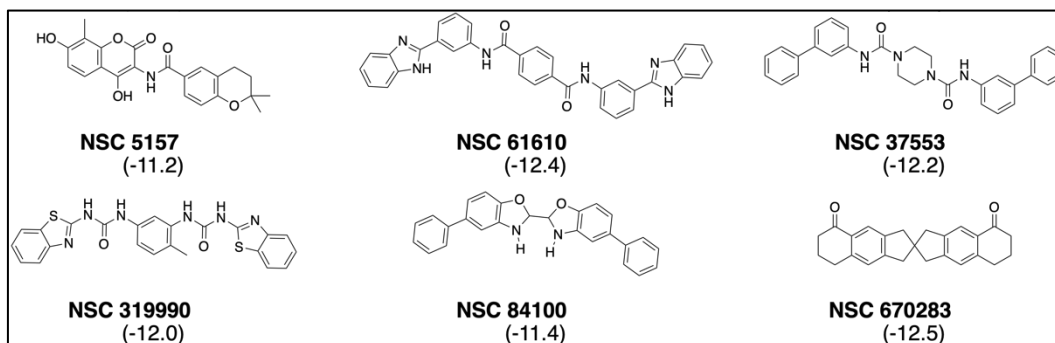


**Figure 4.5. Structural representation of level of conservation of spike receptor binding domain (RBD) residues among SARS, MERS, and CoV-2 spike sequences collected from the National Center for Biotechnology information in 2020.** Red color indicates the residue is highly conserved, white color indicates intermediate conservation, blue color indicates lower conservation, and gray color indicates the residue was not analyzed for conservation. The color attribute was applied using the `defattr` (define attribute) command in Chimera (Pettersen et al., 2004) with the attribute value for each residue assigned as the corresponding conservation score. The top five most conserved RBD residues listed in Table 4.1 (Y423, C361, C379, N422, C391) are labeled in one monomer of the trimer and have side chains shown in all the three monomers of the trimer.

#### 4.5.2 Molecular docking studies reveal small molecules that can bind to different pockets in the S1 domain of spike protein

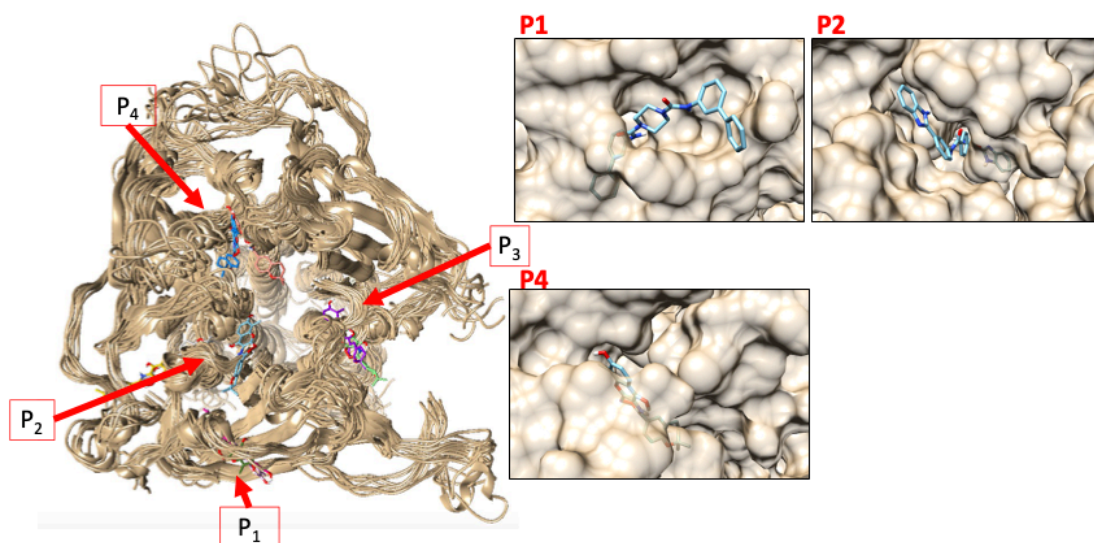
The sequence conservation analysis has identified which residues may be effective antiviral targets that will persist even as new viral strains emerge; however, it remains to be investigated which residues are structurally targetable using small-molecule drugs. To identify compounds and binding pockets of these compounds within the spike protein, molecular docking was performed using a molecular model provided the D.E. Shaw research group ([Creative Commons Attribution 4.0 International Public License](#)). This structure based on PDB IDs 6VXX and 6VW1 contained only the RBD and a short region of S1 fusion region. 11 different snapshots representing different conformations of the protein was used for molecular docking. We screened a small molecule drug library (NCI diversity Set V compound library - 5184 FDA-investigational drugs) against all these 11 structures. Molecular docking was performed using AutoDock Vina (version 1.1.2). We used the following criteria to identify compounds with high binding affinity: (i) a cutoff score of -10.0 kcal/mol was used to select high affinity compounds; (ii) binding site for more than 5 conformations; (iii) rank and select

the highest binding affinity. This gave us top 6 compounds ranked based on highest binding affinity (NSC 5157, NSC 61610, NSC 37553, NSC 319990, NSC 84100, and NSC 670283) (Figure 4.6).



**Figure 4.6. Six compounds with the highest predicted binding affinity to the spike receptor binding.** The compounds are NSC 5157 (binds pocket 4), NSC 61610 (binds pocket 2), NSC 37553 (binds pocket 1), NSC 319990 (binds pocket 2), NSC 84100 (binds pocket 1), and NSC 670283 (binds pockets 1 at -12.4 kcal/mol and 4 at -12.3 kcal/mol). The predicted binding affinity (in units of kcal/mol) is shown in parentheses below each molecule.

Our analysis revealed that there are four binding pockets that are located in the RBD regions of the spike protein timer (Figures 3.3 and 3.4 and Table 4.2). Each pocket can bind to more than one compound indicating the accessibility of the different binding cavities towards diverse chemical compounds (Figure 4.7).



**Figure 4.7. The four pockets identified in the spike RBD that are predicted to bind small molecules.** P stands for pocket. The molecules are shown in the spike protein structure binding in the pockets, and a zoomed in view of the molecules in pockets 1, 2, and 4 is shown on the right.

#### 4.5.3 Mapping of the binding pockets onto the spike protein trimer shows different interfaces of RBD as effective drug binding sites

We analyzed the amino acids that were located within 4 Å of each of the four pockets (Table 4.2.) Pockets 2 and 4 have more interactions with spike protein compared to pockets 1 and 3.

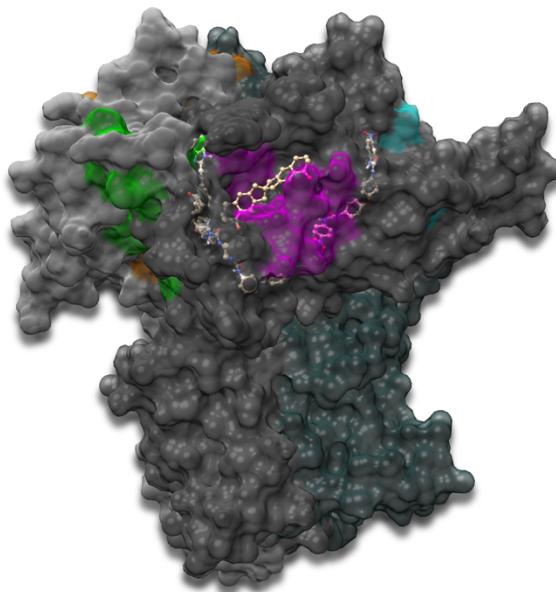
| Pocket 1                                                                                                                   | Pocket 2                                                                                                             | Pocket 3                                                                                                             | Pocket 4                                                                                                             |
|----------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| F338, G339,<br>E340, V341,<br>F342, W353,<br>E354, R355,<br>K356, K357, I358,<br>T385, K386,<br>L387, N388,<br>D389, V395, | D364, Y365,<br>S366, V367,<br>L368, Y369,<br>S375, T376,<br>F377, K378,<br>C379, Y380,<br>A384, T385,<br>K386, L387, | Y369, N370,<br>S371, S383,<br>A384, T385,<br>K386, K403,<br>G404, D405,<br>D406, V407,<br>R408, Q409,<br>Q414, T415, | F342, N343,<br>A344, T345,<br>S371, T372,<br>F373, F374,<br>S375, T376,<br>Y380, R408,<br>Q409, I410,<br>A411, P412, |

|                           |                                                                                                                                                                                 |                                          |                                                                                                                                                                                          |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Y396, A397,<br>D398, S399 | N388, D389,<br>G404, D405,<br>D406, V407,<br>R408, Q409,<br>G416, V417,<br>I418, A419, D420,<br>Y421, Q474,<br>A475, G476,<br>S477, T478,<br>P479, N487,<br>C488, Y489,<br>F490 | G416, V417,<br>K452, Y453,<br>R454, L455 | G413, Q414,<br>T415, V433,<br>L434, A435,<br>W436, N437,<br>T438, R439,<br>N440, I441,<br>D442, Q498,<br>P499, T500,<br>N501, G502,<br>V503, G504,<br>Y505, Q506,<br>P507, Y508,<br>R509 |
|---------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

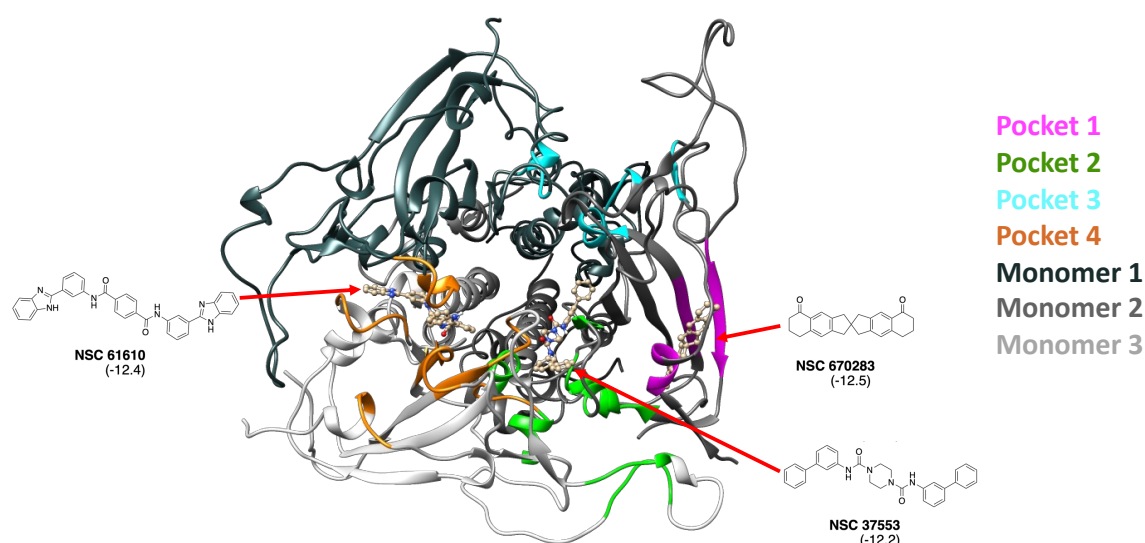
Table 4.2. Index of the residues in the four small-molecule binding pockets in the spike receptor binding domain (RBD).

Mapping these pockets onto the spike protein trimer (Figures 4.7 and 4.8) shows interesting features about the distribution of these pockets. While three of the pockets are in the RBD interfaces of the trimer (pockets 2, 3, and 4), one of them (pocket 1) is covering only one RBD region (Figure 4.9). The location of pocket 2 on the spike structure as shown in Figure 4.9 is especially interesting because the peripheral loop region of each monomer can open (as seen in the monomer where pocket 3 is shown) and close (as seen in the monomer where pocket 2 is shown) (Khare et al., 2021). The open conformation is the one that facilitates recognition for the host receptor protein (Gui et al., 2017), the recognition of which is a crucial step of the CoV-2 viral infection mechanism (Wrapp et al., 2020). Small molecule-binding in pocket 2 has the potential to tether the peripheral loop region to the central helices (Figure 4.9), which would trap the protein in the closed conformation and disrupt the CoV-2 infection mechanism. Furthermore, pocket 2 is located at the interface of two monomers in the trimeric

spike protein (Figure 4.9), so small molecule-binding in this region could disrupt the overall stability of the trimer. As in pocket 2, small molecule binding at the monomer interface in pockets 3 and 4 could disrupt the overall stability of the spike trimer.



**Figure 4.8. Surface representation of the spike protein S1 domain used for molecular dynamics simulation to predict structural conformations that can be used as molecular docking models.** The four molecules predicted to have the strongest binding interactions with spike are also shown. The molecules are NSC 670283, NSC 37553, NSC 80735, and NSC 61610, which are the molecules with the strongest predicted binding interaction in pockets 1-4 respectively. The residues colored magenta, green, teal, and orange are located in pockets 1, 2, 3, and 4 respectively. These pockets were identified because they are regions in the spike protein that are predicted to bind a variety of small molecules. Each of the trimeric subunits of the spike protein are colored in a different shade of gray.



**Figure 4.9. Ribbon representation of the spike protein structure obtained from molecular dynamics simulation.** The molecules with the best binding energy from pockets 1, 2, and 4 (NSC 670283, NSC 37553, and NSC 61610 respectively) are also shown. The units of predicted binding energy are kcal/mol. Each of the trimeric subunits of the spike protein are colored in a different shade of gray.

#### 4.5.4 Conservation of spike protein residues in small-molecule-binding-pockets

One reason why conserved residues are desirable antiviral targets is that they are less likely to change as new variants of the virus emerge, so the antiviral is more likely to be effective against emerging strains. Another reason why conserved residues are promising antiviral targets is that they are often essential for the viral infection mechanism, so disrupting their protein-protein interactions using small molecule drugs has the potential to prevent host infection. Table 4.3 was constructed to identify which pocket is the most advantageous antiviral target based on level of amino acid conservation between the different severe human infecting strains (SARS, CoV-2, and MERS).

| Pocket 1 | Pocket 2 | Pocket 3 | Pocket 4 |
|----------|----------|----------|----------|
|----------|----------|----------|----------|

| Res. | Cons.<br>Score | Res. | Cons.<br>Score | Res. | Cons.<br>Score | Res. | Cons.<br>Score |
|------|----------------|------|----------------|------|----------------|------|----------------|
| W353 | 0.81033        | C379 | 0.83596        | Y453 | 0.80671        | Y380 | 0.78321        |
| F338 | 0.80779        | F377 | 0.81991        | S383 | 0.78390        | P507 | 0.78016        |
| D398 | 0.79528        | Y421 | 0.81052        | G416 | 0.75903        | T376 | 0.76805        |
| R355 | 0.78797        | D420 | 0.78964        | Q409 | 0.75045        | Y508 | 0.76364        |
| I358 | 0.76760        | I418 | 0.78589        | Q414 | 0.74488        | N440 | 0.75130        |
| S399 | 0.76380        | Y380 | 0.78321        | P384 | 0.74315        | S375 | 0.75126        |
| Y396 | 0.75470        | P479 | 0.78314        | R454 | 0.74221        | Q409 | 0.75045        |
| E340 | 0.74978        | T376 | 0.76805        | N370 | 0.73969        | R509 | 0.74758        |
| D389 | 0.74896        | D364 | 0.76386        | R408 | 0.73278        | Y505 | 0.74564        |
| N388 | 0.73762        | G416 | 0.75903        | T415 | 0.72780        | Q414 | 0.74488        |

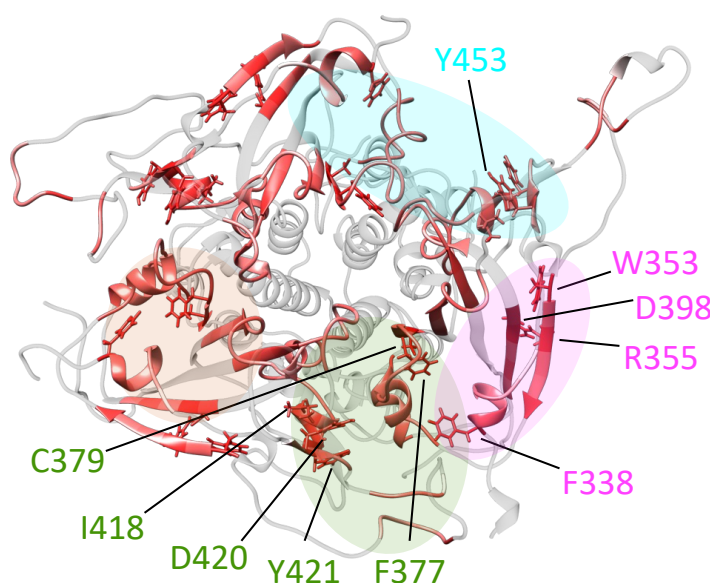
**Table 4.3. Top ten conserved amino acid residues from the four small-molecule binding pockets in the spike receptor binding domain (RBD).**

Conservation analysis was done with the dataset of spike proteins from the severe, human-infecting coronavirus (SARS, MERS, and CoV-2) obtained from the National Center for Biotechnology Information in 2020. All the residues in the pockets are shown in Table 4.2. The bold and colored (magenta for pocket 1, green for pocket 2, and cyan for pocket 3) residues and scores are the top ten most conserved residues in the small-molecule binding pockets. They are labeled and have their sidechains shown in the spike structure in Figure 4.10.

Res. – amino acid residue. Cons. score – conservation score.

The most conserved residues in the small-molecule-binding pockets are found primarily in pockets 2 and 1 (Figure 4.10). Altogether, the molecular docking, structural observation, and amino acid conservation analyses have indicated that pocket 2 is a promising antiviral target based on its predicted susceptibility to small molecule-binding, potential structural effects of small molecule binding on the CoV-2 infection mechanism, and high conservation of amino acid residues.





**Figure 4.10. Structural representation of the level of conservation of the residues in the small molecule binding pockets in the spike protein receptor binding domain.**

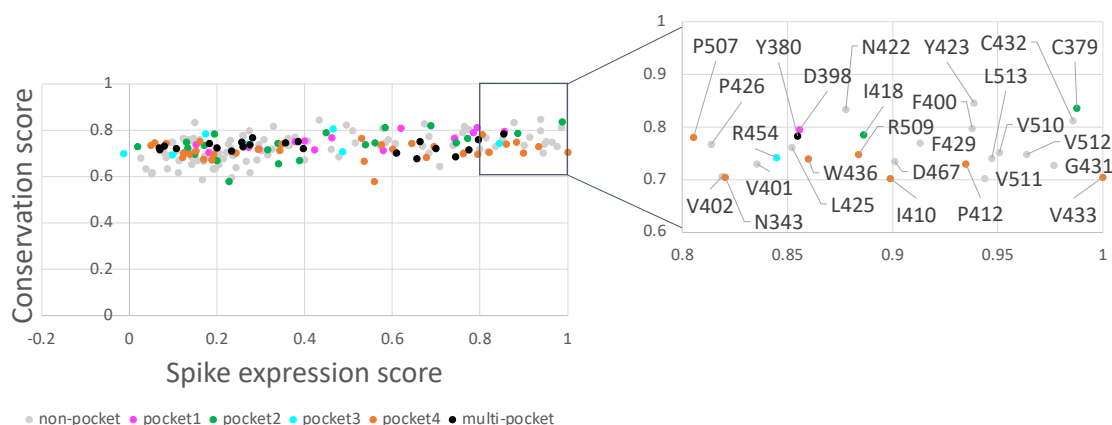
Conservation analysis was done with the dataset of spike proteins from the severe, human-infecting coronavirus (SARS, MERS, and CoV-2) obtained from the National Center for Biotechnology Information in 2020. The top ten most conserved residues in the pockets are labeled in magenta for pocket 1, green for pocket 2, and cyan for pocket 3 (see bolded and colored entries in Table 4.3). The pocket 1 (magenta), pocket 2 (green), pocket 3 (cyan), and pocket 4 (orange) regions are shaded with color.

#### 4.5.5 Conserved spike residues that are important for expression and spike:ACE2-binding are found in small-molecule-binding pockets in the receptor binding domain

Both the conservation analysis and the small-molecule-binding pocket prediction are computational approaches. In order to incorporate experimental data in our analysis, we integrated biochemical data from Starr et al. 2020 which analyzed the role of each amino acid in the CoV-2 spike RBD (residues 331-530) in protein expression, which correlates with the stability and folding of the protein, and in spike:ACE2-binding interactions (Starr et al., 2020).

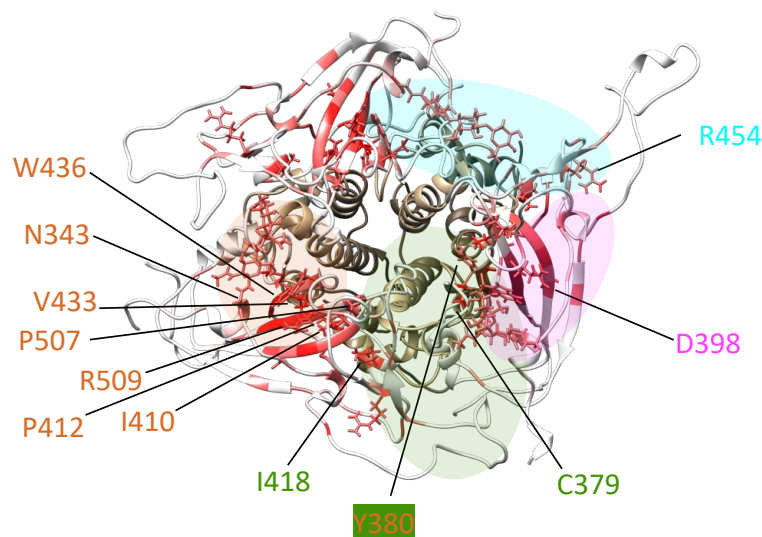
#### 4.5.5.1 Conserved residues in the small-molecule-binding pockets of the receptor binding domain affect spike protein expression

The first set of biochemical data from Starr et al. 2020 to complement the sequence conservation and molecular docking analyses of spike protein is measurements of spike expression after substitution of individual RBD residues (Starr et al., 2020). In this study an *in vivo* approach was used to mutate each individual amino acid in the RBD domain to all other 19 amino acids individually, followed by protein expression analysis (Starr et al., 2020). The spike protein expression score was obtained by taking the average spike expression signal intensity value of all amino acid substitutions from Table S2 in Starr et al. 2020 (Starr et al., 2020), and then normalizing those values by dividing each by the spike expression score value with the highest absolute value. We then plotted a graph with protein expression score on the X-axis and the level of sequence conservation (among spike protein from SARS, MERS, and CoV-2) on the Y-axis. A higher spike protein expression score (X-axis) indicated that spike cannot be properly expressed when that RBD residue is substituted to another amino acid. A higher value on the Y-axis shows that a particular amino acid is highly conserved across all the severe human infecting viruses. As shown previously (Figure 4.4.), RBD residues are highly conserved compared to other regions of the spike protein. Excluding the residues not located in a pocket, the upper right corner of the spike expression analysis plot is dominated by pocket 4 (Figure 4.11). This indicates that targeting pocket 4 with small molecule drugs could disrupt the ability of CoV-2 to express spike protein, thus preventing a successful viral infection.



**Figure 4.11. Scatterplot of spike protein receptor binding domain (RBD) residues showing their degree of influence on spike protein expression and level of conservation.** The color of the datapoint indicates which small-molecule binding pocket contains that residue. The residues in the upper-right corner of the plot (residues C379, P507, R454, N343, D398, Y380, I418, V410, R509, W436, P412, and V433) are both crucial for spike protein expression and highly conserved among spike proteins from SARS, MERS, and CoV-2.

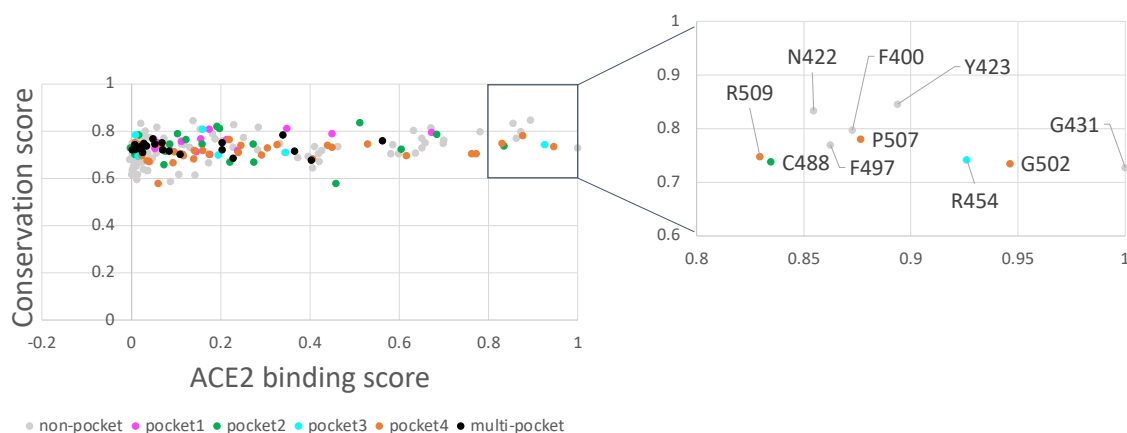
Structural representation of this biochemical data from Starr et al. 2020 shows that residues crucial for spike expression (colored red) tend to appear in the beta sheet regions (Figure 4.12). Figure 4.12 also reinforces the observation that pocket 4 has a high density of conserved residues that are crucial for spike protein expression.



**Figure 4.12. Structural representation of conserved small-molecule-binding pocket residues crucial for spike protein expression.** Red color indicates substituting the residue hinders spike protein expression, white color indicates substituting the residue does not affect spike expression, and tan color indicates the residue was not biochemically tested by Starr et al. in 2020 (Starr et al., 2020). The color intensity in this figure cannot be compared to that of other figures because signal intensity of this assay cannot be normalized to that of the spike ACE2-binding assay (Figure 4.14) or sequence conservation analyses (Figures 3.2, 3.5, and 3.11). The residues that are labeled and have sidechains shown are found in the zoom-in box in Figure 4.11 (residue positions 379, 507, 454, 343, 398, 380, 418, 410, 509, 436, 412, and 433), and the label colors and shaded regions correspond to the small-molecule-binding pockets: pocket 1 (magenta), pocket 2 (green), pocket 3 (cyan), pocket 4 (orange). Y380 is found in both pockets 3 and 4, so it is colored both cyan and orange. The PDB ID is 6VXX, and the color attribute was applied using the `defattr` (define attribute) command in Chimera (Pettersen et al., 2004) with the attribute value for each residue assigned as the corresponding average spike expression signal intensity value of all amino acid substitution from Table S2 in Starr et al. 2020 (Starr et al., 2020).

#### 4.5.5.2 Conserved residues in the small-molecule-binding pockets of the receptor binding domain affect spike:ACE2-interactions

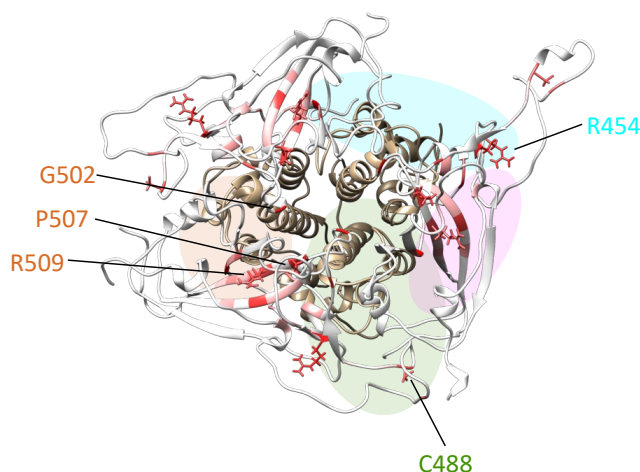
The second set of biochemical data from Starr et al. 2020 to complement the sequence conservation and molecular docking analyses of spike protein is measurements of spike:ACE2-binding interactions after substitution of individual RBD residues (Starr et al., 2020). The spike:ACE2-binding score was obtained by taking the average spike:ACE2-binding signal intensity value of all amino acid substitution from Table S2 in Starr et al. 2020 (Starr et al., 2020), and then normalizing those values by dividing each by the ACE2 binding score with the highest absolute value. RBD residues are plotted in Figure 4.13 with spike:ACE2-binding score on the X-axis and level of conservation (among spike protein from SARS, MERS, and CoV-2) on the Y-axis. High ACE2 binding score indicated that spike cannot properly bind ACE2 when that RBD residue is substituted to another amino acid. As with the spike expression analysis plot in Figure 4.11, pocket 4 residues dominate the upper right corner of the amino acid conservation spike:ACE2-binding analysis plot (Figure 4.13), which indicates that targeting pocket 4 with small molecule drugs could disrupt the ability of CoV-2 to bind ACE2 for infection.



**Figure 4.13. Scatterplot of spike protein receptor binding domain (RBD) residues showing their degree of influence on spike:ACE2-binding and level of sequence conservation. The color of the datapoint indicates which**

small-molecule binding pocket contains that residue. The residues in the upper-right corner of the plot (residues G502, P507, R509, C488, and R454) are both crucial for ACE2-binding and highly conserved among spike proteins from SARS, MERS, and CoV-2.

Structural representation of this biochemical data from Starr et al. 2020 shows that residues crucial for spike:ACE2-binding (colored red) tend to appear in the beta sheet regions (Figure 4.14), which is similar to the spike expression results in Figure 4.12; however, there is a noticeable increase in importance of the peripheral loop regions for spike:ACE2 binding relative to spike expression (Figures 3.7 and 3.9).

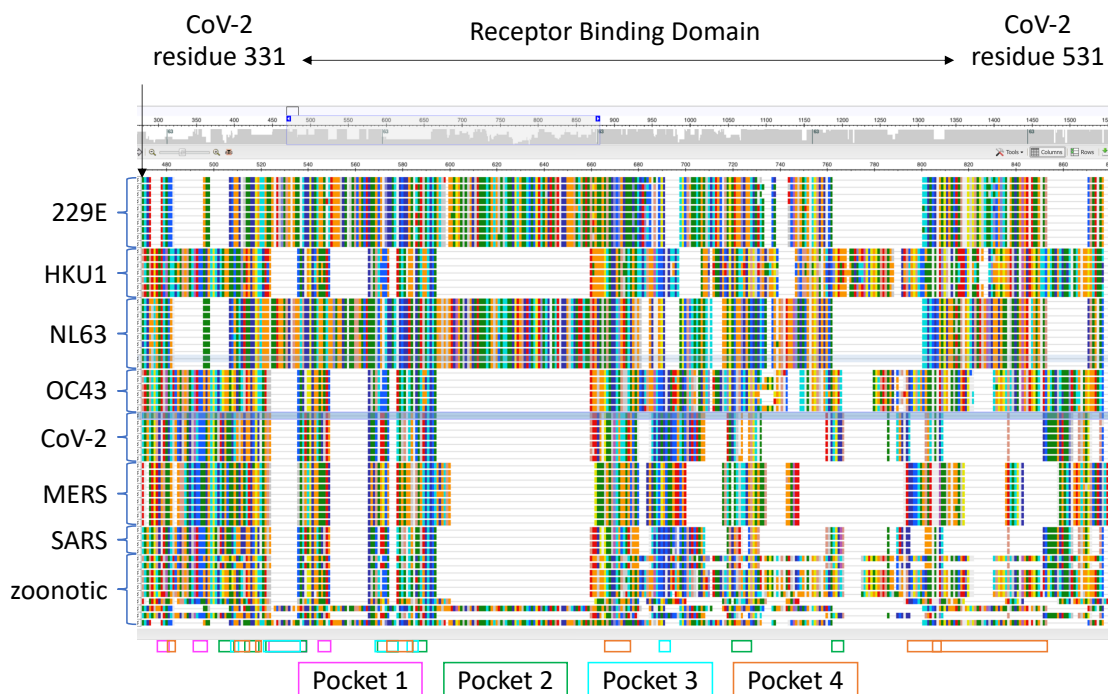


**Figure 4.14. Structural representation of conserved small-molecule-binding pocket residues crucial for spike:ACE2-binding.** Red color indicates substituting the residue for other amino acids hinders ACE2 binding, white color indicates substituting the residue does not affect ACE2 binding, and tan color indicates the residue was not biochemically tested by Starr et al. in 2020 (Starr et al., 2020). The color intensity in this figure cannot be compared to that of other

figures because signal intensity of this assay cannot be normalized to that of the spike expression assay (Figure 4.12) or sequence conservation analyses (Figures 3.2, 3.5, and 3.11). The residues that are labeled and have sidechains shown are found in the zoom-in box in Figure 4.13 (residue positions 502, 507, 509, 488, and 454), and the label colors and shaded regions correspond to the small-molecule-binding pockets: pocket 1 (magenta), pocket 2 (green), pocket 3 (cyan), pocket 4 (orange). The PDB ID is 6VXX and the color attribute was applied using the `defattr` (define attribute) command in Chimera (Pettersen et al., 2004) with the attribute value for each residue assigned as the corresponding average ACE2 binding signal intensity value of all amino acid substitution from Table S2 in Starr et al. 2020 (Starr et al., 2020).

#### 4.5.6 Conservation of some spike residues changes when zoonotic-isolate spike sequences are added to a dataset of human-infecting CoV spikes

Since CoV-2 probably originated in a zoonotic source and then adapted to infecting humans (Qiu et al., 2020), it is of interest what changes may have occurred between spike proteins from coronaviruses isolated from humans and closely related coronaviruses isolated from zoonotic sources (Gussow et al., 2020). To investigate this, a dataset of spike proteins from human-isolated CoVs, including both severe and non-severe disease-causing strains, (229E, OC43, NL63, HKU1, SARS, MERS, CoV-2) was collected, and then an expanded dataset was made by adding 10 spike proteins from zoonotic-isolate CoVs to the 53 from the human-infecting strains. Conservation was scored for both datasets, and the top five conserved residues from each dataset are shown in Table 4.4.



**Figure 4.15. Multiple sequence alignment (MSA) of the dataset containing human and zoonotic spike proteins.** The MSA is zoomed in on CoV-2 sequence QHO62112 residues 331-531, which encompass the receptor binding domain. The colors in the MSA represent different amino acids according to the RasMol amino color scheme. The magenta, green, cyan, and orange boxes at the bottom of the MSA mark which MSA positions are in which pockets according to the QHO62112 amino acid residue positions. CoV-2 sequence QHO62112 is highlighted in blue. Figure made using the [NCBI MSA viewer](#).

Examination of the multiple sequence alignment of the spike protein dataset with both human and zoonotic strains reveals that eight of the ten spike sequences from zoonotic-infecting coronaviruses (with exception to 229E-related QHA24665.1 and camel alphacoronavirus YP\_009194639.1) resemble the beta coronavirus spikes (from SARS, MERS, CoV-2, OC43 and HKU1) which constitute the majority of the dataset of 53 spikes from human-infecting coronaviruses. These 8 zoonotic sequences are different from the alpha coronavirus spikes (from 229E and NL63) which constitute the minority of the



dataset of 53 spikes from human-infecting coronaviruses (Figure 4.15), which may explain why adding the 10 zoonotic-isolate sequences increased the overall conservation score for RBD residues relative to the dataset with just the 53 sequences from human-infecting coronaviruses only (Table 4.4 and the datapoints above the green line in Figure 4.5).

Table 4.4 shows that pocket 4 tends to be the most conserved in the human-isolates while pocket 2 is more conserved in human+ zoonotic-isolates. This indicates that pocket 4 of spikes from human-isolates has a different amino acid composition than that of zoonotic-isolates; however, pocket 2 has similar amino acid composition in spikes from human- and zoonotic-isolates. This means that pocket 4 may be a region where changes occur that help CoVs adapt to new hosts, while pocket 2 is a region that stays constant among the viruses infecting different host strains. Furthermore, pocket 4 may be an effective antiviral target in human-infecting CoVs, but not in animal-infecting CoVs, while pocket 2 may be a broad antiviral effective against both human- and animal-infecting CoVs.

Notably, C488 in pocket 2 is conserved in the human-isolate only dataset but not in the zoonotic-isolate added dataset (Table 4.4). This is of interest because C488 is disulfide bonded to C480 which is located in a region that is reported to appear as an insertion in the human-infecting CoV spikes whenever they are compared to their closest zoonotic relative (Gussow et al., 2020). The results from the conservation analysis in this study suggest that the regions related to C480 tend to be found in human-isolates and not in zoonotic-isolates, which supports the findings in Gussow et al.

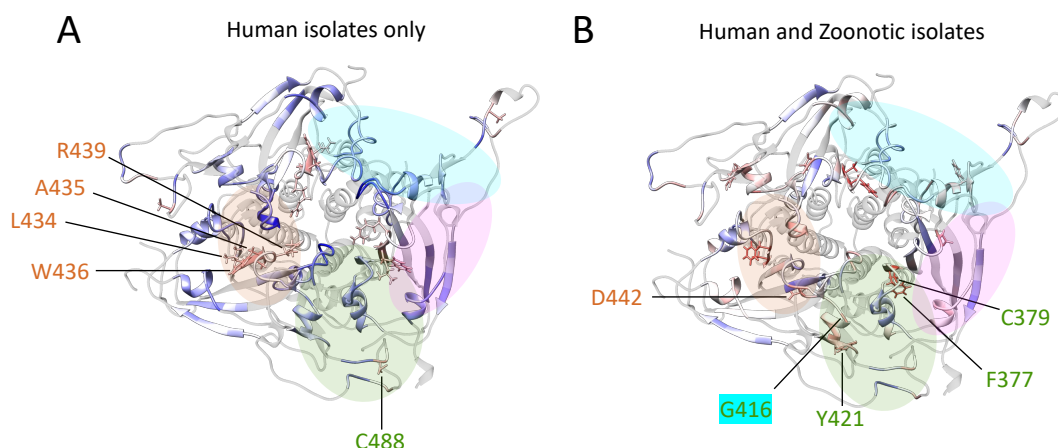
| Human-isolate only |                    | Human-isolates + zoonotic-isolates |                    |
|--------------------|--------------------|------------------------------------|--------------------|
| Residue            | Conservation Score | Residue                            | Conservation Score |
| W436               | 0.65154            | C379                               | 0.76180            |
| A435               | 0.64082            | F377                               | 0.73763            |

|      |         |      |         |
|------|---------|------|---------|
| C488 | 0.63783 | G416 | 0.68213 |
| L434 | 0.61512 | D442 | 0.66200 |
| R439 | 0.61041 | Y421 | 0.64884 |

**Table 4.4. Top five conserved residues in the small-molecule-binding pockets in the spike receptor binding domain (RBD) for datasets obtained from the National Center for Biotechnology Information in 2020 containing spike proteins from all seven human-infecting coronavirus species (229E, OC43, NL63, HKU1, SARS, MERS, CoV-2) and from coronaviruses isolated from zoonotic hosts.** Residue font color correspond to the small-molecule-binding pockets: pocket 1 (magenta), pocket 2 (green), pocket 3 (cyan), pocket 4 (orange). G416 is in both pocket 2 and 3, so it is colored both green and cyan. A structural representation of these residues and corresponding pockets is shown in Figure 4.16.

Structural representation of the top five conserved residues from the human-isolates (seven species from alpha and beta viruses: 229E, OC43, NL63, HKU1, SARS, MERS, CoV-2) and all human + zoonotic-isolates show that C488 is in the peripheral loop region of pocket 2 and that its disulfide partner C480 is close-by and outside of the pocket 2 (Figure 4.16 A). The white shade of C488 in Figure 4.16 B relative to its red shade in Figure 4.16 A shows that adding the spike protein sequences from the zoonotic-isolates significantly decreased its conservation. Similar to the spike protein expression (Figure 4.12) and ACE2-binding (Figure 4.14) representations analyzing only severe disease causing strains, the conservation analysis with all seven human-infecting species (Figure 4.16 A) has a high concentration of residues of interest in pocket 4. Similar to the severe, human-infecting coronavirus conservation analysis (SARS, MERS, and CoV-2, Figure 4.10), the conservation analysis with integrated human and zoonotic-isolates dataset (229E, OC43, NL63, HKU1, SARS, MERS, CoV-2 plus zoonotic-isolates, Figure 4.16 A) have many conserved residues in pocket 2, which indicates that pocket 2 of the spike proteins from the zoonotic-isolates may be more similar to that of severe human-infecting CoVs (SARS, MERS, and CoV-

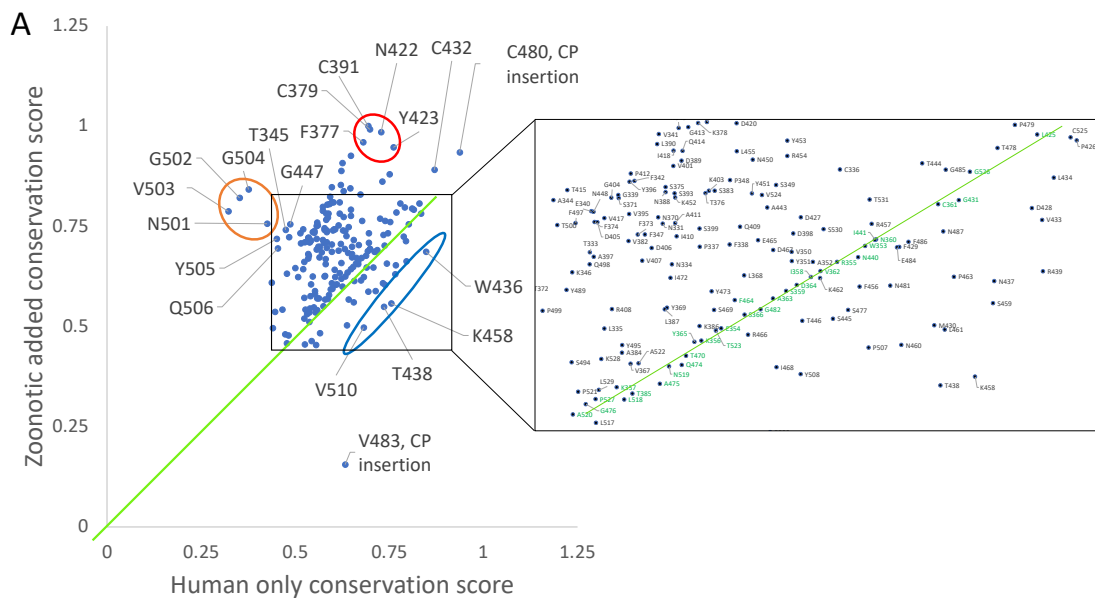
2) than to that of the CoVs that cause common cold-like symptoms in humans (229E, OC43, NL63, HKU1).



**Figure 4.16. Structural representation of the level of conservation of residues in the small molecule binding pockets in the spike protein receptor binding domain.** The datasets collected from the National Center for Biotechnology Information in 2020 are composed of spike proteins from coronaviruses that are **A)** human-isolate only (229E, OC43, NL63, HKU1, SARS, MERS, CoV-2) and **B)** (A) + zoonotic-isolates. The top five most conserved residues (Table 4.4) in the pockets have their atoms shown and are labeled in green for pocket 2, cyan for pocket 3, and orange for pocket 4. G416 is in pockets 2 and 3, so its label is colored both green and cyan. The pocket 1 (magenta), pocket 2 (green), pocket 3 (cyan), and pocket 4 (orange) regions are shaded with color.

To better visualize which spike RBD residues have major shifts in conservation when the zoonotic-isolate sequences are added to all human-infecting isolates (7 species), the RBD residues were plotted with the conservation scores from the human-isolate (7 species) only analysis on the x-axis and the conservation scores from all human+zoonotic-isolates analysis on the y-axis (Figure 4.17). The result of such a plot is that residues with little change in conservation before and

after adding in the spike sequences from zoonotic-isolates will form a line with a slope of 1 while those residues whose conservation was greatly affected by the addition of spike protein sequences from the zoonotic-isolate will deviate from that line. Residues that deviate above the line are those that spike proteins from zoonotic- and human-isolates have in common while those that deviate below the line are those that are dissimilar between spike proteins from zoonotic- and human-isolates. Groups of deviant residues are circled in orange (because all those residues are in pocket 4, Table 4.2), red, and blue (red and blue do not show a strong trend of belonging to a particular small-molecule binding pocket; Figure 4.17 A). Residues along the line with a slope of 1 are highlighted in green (Figure 4.17 A). The colors are then used to structurally show where these groups of residues are located in the spike RBD (Figure 4.17 B). The residues whose conservation scores were unaffected by the addition of zoonotic-isolate spike sequences (green) are found on the periphery of the spike RBD, the pocket 4 residues (orange) are mainly found in central loop and alpha-helical regions, the residues in red are mainly in loop regions, and the residues in blue are mainly in beta-sheets (Figure 4.17).



B



**Figure 4.17. Visualization of how adding spike protein sequences from zoonotic-isolates to those from all human-infecting species changes conservation of receptor binding domain (RBD) residues.** A) Plot of conservation scores of the spike RBD residues (including those that are not in small-molecule binding pockets). The x-axis scores are from a dataset containing only human-infecting coronavirus spikes (7 coronavirus species: 229E, OC43, NL63, HKU1, SARS, MERS, CoV-2), and the y-axis scores are from a dataset containing additional spike protein sequences from zoonotic-isolates. The green line is drawn with a slope of one through the origin such that it will pass through datapoints that have nearly the same conservation score before and after adding the zoonotic isolate spike sequences. The residues that deviate from the green line had a change in conservation score once spike protein sequences from the zoonotic-isolates were added. Groups of residues that deviate from the green line are circled in orange (residue numbers 503, 502, 504, and 501), red (residue numbers 377, 379, 391, 422, and 423), and blue (residue numbers 510, 458, 438, and 436). Orange was chosen for residues N501, G502, V503, and G504 because they are all in pocket 4 (Table 4.2). The note for “CP insertion” on V483 and C480 indicates that these residues are involved in a cysteine-proline (CP) insertion event associated with the adaptation of CoV-2 to infecting humans

(Gussow et al., 2020). The zoomed-in panel shows the labels for the residues that did not change conservation score after adding the zoonotic-isolate spike sequences (residue numbers 520, 476, 518, 527, 385, 475, 357, 519, 470, 474, 365, 356, 523, 354, 366, 482, 363, 359, 364, 358, 362, 355, 440, 353, 360, 431, 441, 361, 526, and 425 are highlighted in green because they are the closest to the green line). **B)** Residues from panel A visualized on the spike protein structure. The residues in green are the green highlighted residues in panel A that did not deviate from the green line. Those in orange, red, and blue are those circled with the corresponding color in panel A.

#### 4.5.7 Conserved residues have been shown to be important for biological function, viral evolution, and antibody binding

A literature search analyzing the residues of interest from our conservation analysis in each of the categories of deviant residues (orange, red, and blue) showed them to be critical for biological functions and antibody interactions (Table 4.5). Some residues (indicated with an asterisk) are discussed in more detail in paragraph form below Table 4.5.

| <b>Color-coded location in Figure 4.17</b> | <b>Residue</b>             | <b>Function</b>                                                                                                                                                                         | <b>References</b>   |
|--------------------------------------------|----------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| Orange                                     | Y495 to G504               | affect RBD–ACE2 binding affinity                                                                                                                                                        | (Teng et al., 2021) |
| Orange                                     | G502, Y505, Q498, and N501 | CoV-2 spike residues that interact with human ACE2 residue K353 which is conserved in multiple zoonotic ACE2 proteins that are receptors for CoV-2 spike, but different (H353) in mouse | (Zhao et al., 2020) |

|                                |                                 |                                                                                                                                                                                                     |                                           |
|--------------------------------|---------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------|
|                                |                                 | ACE2 that is not a receptor for CoV-2 spike                                                                                                                                                         |                                           |
| Orange                         | G502                            | interacts with antibody B38; G502 point mutants cause severe destabilizing effect on the RBD:ACE2 complex, with G502P causing the most destabilizing effect; in pocket 4                            | (Lim et al., 2020)<br>(Teng et al., 2021) |
|                                | G502 (along with N501 and Q506) | main protein:protein interaction hotspot between spike and ACE2 and specifically interacts with ACE2 residues G354 through a hydrogen bond and D355 through electrostatic interactions; in pocket 4 | (Lim et al., 2020)                        |
| Orange                         | N501*                           | evolves in highly transmissible strains; in pocket 4                                                                                                                                                | (Aleem et al., 2022)                      |
| Orange                         | N501-Y505                       | targets of neutralizing antibodies; all in pocket 4                                                                                                                                                 | (Huang et al., 2020a; Ku et al., 2021)    |
| Just outside the orange circle | Y505*                           | role in host specificity through interactions with ACE2; one of the top 10 conserved residues in pocket 4 (Table 4.3)                                                                               | (Dimonte et al., 2021)                    |
| Just outside the orange circle | Q506*                           | in pocket 4 (Table 4.2)                                                                                                                                                                             |                                           |

|                                |                                        |                                                                                                                                                                                                                                                                                                    |                                                                         |
|--------------------------------|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------|
|                                | Q506 (along with G502 and N501)        | in the main ACE2 protein-protein interaction hotspot                                                                                                                                                                                                                                               | (Chowdhury et al., 2020; Lim et al., 2020)                              |
| Just outside the orange circle | G447                                   | B-cell epitope; binds negatively charged linear polysaccharide heparan sulfate proteoglycans used by viruses for cell surface attachment                                                                                                                                                           | (Verma and Subbarao, 2021; Wintjens et al., 2020) (Seyran et al., 2021) |
| Just outside the orange circle | T345                                   | antibody epitope                                                                                                                                                                                                                                                                                   | (Huang et al., 2020a; Ku et al., 2021)                                  |
|                                | T345 (along with Y505, N501, and G502) | binds heparin as do CoV-2 spike residues Y505, N501, and G502, and this heparin binding capability affects the transition from closed to open conformation that allows the spike RBD to bind the host receptor; in pocket 4 (Table 4.2)                                                            | (Mycroft-West et al., 2020; Seyran et al., 2021)                        |
| Red                            | F377                                   | in pocket 2 (Tables 3.2, 3.3, and 3.4); contacts linoleic acid that stabilizes the monomers of the spike trimer such that the closed conformation of the spike RBD is favored due to changes in flexibility, solvent accessibility, and pH; located in a hydrophobic pocket that can bind steroids | (Bangaru et al., 2020) (Warwicker, 2021) (Carino et al., 2020)          |



|     |       |                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                      |
|-----|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Red | C379* | in pocket 2 and highly conserved among the severe, human-infecting coronaviruses (Tables 3.1, 3.2, 3.3, and 3.4 and Figure 4.5); crucial for CoV-2 spike expression (Figures 3.6 and 3.7); potential target of antibody neutralizing therapies; interacts with antiviral drugs; disulfide bonds with C432 for structural stabilization; located in a steroid-binding hydrophobic pocket | (Starr et al., 2020) (Gadanec et al., 2021) (Toor et al., 2021) (Hati and Bhattacharyya, 2020) (Carino et al., 2020) |
| Red | C391  | highly conserved among the severe, human-infecting coronaviruses (Table 4.1 and Figure 4.5); binds GRP78 which is an alternate receptor other than ACE2; disulfide bonds to C525 to enhance the stability of RBD beta sheets                                                                                                                                                            | (Allam et al., 2020; Gadanec et al., 2021) (Wang et al., 2020a)                                                      |
| Red | N422  | highly conserved among the severe, human-infecting coronaviruses (Table 4.1 and Figure 4.5); identified as a key residue by the stress metric which gives insight into the mechanisms of protein                                                                                                                                                                                        | (Rezaei et al., 2022)                                                                                                |

|      |      |                                                                                                                                                                                                                                                                                                         |                                                                          |
|------|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------|
|      |      | folding, conformational alterations, and the effects of amino acid mutations on the protein structure and function                                                                                                                                                                                      |                                                                          |
| Red  | Y423 | highly conserved among the severe, human-infecting coronaviruses (Table 4.1 and Figure 4.5); key for spike protein expression and spike:ACE2-binding; works with D398 to balance stability and infectivity of CoV-2 <i>via</i> pH-dependent transitions between the open and closed spike conformations | (Starr et al., 2020)<br>(Warwicker, 2021)                                |
| Blue | V510 | located in a 93-amino acid long sequence (N318-V510) in the spike RBD that corresponds to SARS-CoV residues that are key targets for neutralizing antibodies                                                                                                                                            | (Jiang et al., 2021)                                                     |
| Blue | K458 | infection mechanism-relevant electrostatic interaction with E471; K458 (along with K417, L455, N487, S477, and G476) is important for RBD–antibody interactions; heparin binding                                                                                                                        | (Arbeitman et al., 2020) (Luan and Huynh, 2020)<br>(Seyran et al., 2021) |
| Blue | W436 | located in pocket 4 (Table 4.2), highly conserved among the human-isolate coronavirus spikes (Table 4.4), and crucial                                                                                                                                                                                   | (Rezaei et al., 2022) (Barnes et al., 2020)                              |

|  |  |                                                                                                                                                                                                                                                       |  |
|--|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|
|  |  | for spike protein expression (Figures 3.6 and 3.7); predicted by the betweenness metric (Hu et al., 2014) to be key for stability of the spike protein and its affinity for the ACE2 receptor; located in the neutralizing antibody C144 binding site |  |
|--|--|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|

**Table 4.5 Summary of literature findings for each residue of interest shown in Figure 4.17.** Some residues (indicated by an asterisk) require more in-depth discussion, so paragraphs about some of the residues can be found below.

Orange – Since the Spike RBD residues circled in orange are above the green line and shifted to the left of the graph, they were not very conserved in the human-isolate only analysis, but became more conserved after the zoonotic-isolate sequences were added. Interestingly, they all happen to be found in pocket 4.

Spike residue N501 is mutated to tyrosine in the following highly transmissible variants of CoV-2: Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), and Omicron (B.1.1.529). The first three of these variants were known as the UK, South African, and Brazil strains respectively and were first detected in late 2020 and early 2021 (Aleem et al., 2022). The N501Y mutation is one of several spike protein mutations that emerged in an immunocompromised individual who had prolonged viral replication for over 20 weeks (Choi et al., 2020; Tegally et al., 2021). The CoV-2 spike residue N501 corresponds to the SARS spike residue T487, which is one of the six amino acids in the RBD crucial for ACE2-binding (Andersen et al., 2020). Computational analyses predict that CoV-2 spike residue

N501 is important for RBD:ACE2 hydrogen-bonding network and hydrophobic interactions (Wang et al., 2020b), and Teng et al. found that the N501E substitution in spike causes the largest negative (stabilizing) change in ACE2-binding free energy ( $-2.490$  kcal/mol) (Teng et al., 2021). Electrostatic interactions between ACE2 residue G354 and spike residue N501 occur in the main ACE2 hotspot, which also contains spike residues G502 and Q506 (Lim et al., 2020; Liu et al., 2021).

The CoV-2 spike residue Y505 is substituted to phenylalanine in most animal coronaviruses, so amino acid position 505 might have a role in host specificity (Dimonte et al., 2021). CoV-2 spike residue Y505 corresponds to SARS spike residue Y4911 which is essential for ACE2 binding (Andersen et al., 2020). As expected, CoV-2 spike residue Y505 does contact ACE2, and the specific ACE2 residues it interacts with are E37, K353, G354 and R393 (Durmaz et al., 2020). The ACE2 binding event disrupts the spike H-bond network comprising multiple residues near the orange circled region in Figure 4.17 (V504, Y505, and Q506) (Meirson et al., 2021). Although these observations indicate Y505 is directly involved in binding ACE2, there is controversy whether amino acid position 501 or 505 is key for ACE2 binding (Xu et al.).

CoV-2 spike residue Q506 is in the main ACE2 PPI hotspot along with G502 and N501, and it specifically interacts with ACE2 residue Q325 *via* hydrogen bonding (Chowdhury et al., 2020; Lim et al., 2020). ACE2-binding induces destabilization of CoV-2 spike hydrogen bond networks comprising CoV-2 spike residues V504 and Y50 in addition to Q506 (Meirson et al., 2021). Along with CoV-2 residue Y505, Q506 is frequently substituted for other residues in animal-infecting coronavirus spikes (Dimonte et al., 2021). There is also a predicted Q506 (CoV-2 spike):E329 (human-AE2) hydrogen bond that appears to be disrupted by bat-ACE2 substitution E329N, which may explain why CoV-2 does not infect bats (Zhao et al., 2020).

Red – Since the Spike RBD residues circled in red are above the green line and shifted to the right of the graph, they are Spike RBD residues that were already highly conserved in the human-isolates only analysis, and became even more conserved after the zoonotic-isolate sequences were added.

CoV-2 spike residue C379 is located in pocket 2 and highly conserved among the severe, human-infecting coronaviruses (Tables 3.1, 3.2, 3.3, and 3.4 and Figure 4.5). It is also crucial for CoV-2 spike expression (Figures 3.6 and 3.7) (Starr et al., 2020). The residues in SARS and MERS spike that correspond to CoV-2 spike residue C379 are targets of antibody neutralizing therapies against SARS-CoV and MERS (Gadanec et al., 2021), and CoV-2 residue C379 itself interacts with antiviral drugs (Toor et al., 2021). C379 disulfide bonds with C432 for structural stabilization (Hati and Bhattacharyya, 2020), and the location of C379 is a steroid-binding hydrophobic pocket (Carino et al., 2020).

Blue – Since the Spike RBD residues circled in blue are below the green line and shifted to the right of the graph, they are Spike RBD residues that were highly conserved in the human-isolates only analysis, but became even less conserved after the zoonotic-isolate sequences were added.

In general, the residues in the orange circled region of Figure 4.17 are located in pocket 4 of the spike RBD and are involved in ACE2 interactions, those in the red circled region are located in pocket 2 and are involved in trimer stability and spike conformation, and those in the blue circled region are targets for antibody binding. It is interesting that analyzing differences in conservation of these RBD residues using human-isolate only and an expanded dataset of human- and zoonotic-isolates can reveal groups of residues that each share a functional theme. This may indicate that the identity of the residues in the orange, red, and blue circled regions in Figure 4.17 are key determinants of spike binding of the host receptor, conformation/stability/flexibility, and host immune-system

interactions that differ among various coronavirus species and their corresponding hosts.

#### 4.6.0 Discussion and future directions

##### 4.6.1 Reasoning behind targeting highly conserved *versus* variable proteins or amino acid residues

The rationale behind investigating spike protein as an antiviral target is the idea that conserved residues in a variable protein are potentially effective antiviral targets. The reasoning behind targeting conserved residues is that they are likely to be essential for protein structure and function, so drugs binding in these regions are likely to effectively disrupt the viral infection mechanism.

Furthermore, conserved residues are more stable drug targets, so the drug is likely to be effective against various viral strains in the future. Residues considered effective, stable drug targets in spike due to their high conservation and necessity for spike protein expression or spike:ACE-2 binding are C379, P507, R454, N343, D398, Y380, I418, V410, R509, W436, P412, V433, G502, P507, R509, C488, and R454, most of which are in spike RBD small-molecule-binding pocket 4 or 2 (Figures 3.6-3.9).

It is, however, interesting to consider the advantages of targeting variable residues since these are the most likely to have recently changed for the adaptation of CoV-2 to human hosts. Analyses were done with animal-infecting coronaviral spike protein sequences to investigate such potential adaptations. Residues that vary with host were found to be CoV-2 spike residue position numbers 503, 502, 504, and 501, 377, 379, 391, 422, and 423, 510, 458, 438, and 436 (Figure 4.17). These residues fell into three groups with similar conservation profiles and mechanistic roles, which indicates that they may be related to key adaptations in spike protein that contribute to spillover of coronaviruses from zoonotic sources into the human population.

The reasoning behind targeting a variable protein like spike is that it is likely be involved in the adaptation process that facilitated the entry and spread of CoV-2 in the human population. A conserved protein such as 3-chymotrypsin like protease (Roe et al., 2021), however, is also a promising drug target since it is likely needed for the viral infection mechanism in a wide variety of coronavirus species.

#### 4.6.2 Future Prospects

Since there is now a nearly three years' worth of CoV-2 spike protein sequence data, sequence conservation analysis with respect to time (especially correlating sequence changes with known surges in CoV-2 cases) is a further investigation that could be done to identify the CoV-2 spike residues that are currently most active in helping CoV-2 adapt to human hosts. To introduce biophysical studies, isothermal titration calorimetry could be used to measure which drugs best disrupt spike:ACE2 binding. Alternatively, pseudoviral entry assays could be used to study which molecules hinder CoV-2 entry into the host cell. Pocket 4 was identified as containing highly conserved residues crucial for Spike:ACE2 binding, and the molecule with the best predicted binding score is NSC 670283, so that may be one of the first molecules to test. Long-term future work would be to modify small molecule drugs to improve their ability to disrupt the spike:ACE2 binding. Another future work possibility is to do conservation analysis on 3-chymotrypsin like protease protein to identify which of its residues are conserved potential drug targets.

## 5 Chapter 5: Outlook

### 5.1.0 Summary

#### 5.1.1 CRISPR-Cas-related projects

Shortly after the characterization of SpyCas9 as a programmable endonuclease, it began to be used in cell-based assays as a gene editing tool. Various unintended effects, such as chromosome truncations and cytotoxicity, were observed in mammalian cells treated with SpyCas9. These unintended effects could be attributed to off-target DNA cleavage in instances where the gRNA partially matches the genome at sites other than the target site. Another potential culprit is gRNA-free DNA cleavage from Cas9 proteins that are not bound to a gRNA, but are free in the mammalian cell nucleus damaging DNA that is transiently single stranded during processes such as DNA replication. To mitigate the possibility of non-specific DNA cleavage by SpyCas9, variants that are devoid of gRNA-free ssDNA degradation activity (SpyCas9a<sup>H982A</sup>) and more selective for on-target DNA linearization (SpyCas9<sup>2Pro</sup>) were developed.

##### 5.1.1.1 Engineering SpyCas9 without guide RNA-free DNA cleavage (SpyCas9<sup>H982A</sup>)

In this work, SpyCas9<sup>WT</sup> and SpyCas9<sup>H982A</sup> was assayed for metal dependence and kinetics of gRNA-free ssDNA degradation. The H982A substitution caused ~167 and ~17-fold reductions in the Mn<sup>2+</sup>- and Co<sup>2+</sup>-mediated gRNA-free ssDNA cleavage rate constants respectively. Since the SpyCas9<sup>H982A</sup> variant is intended for use in gene therapy applications, another important activity to characterize is its ability to do RNA-guided DNA cleavage. The H982A substitution causes a ~1.2-fold reduction in the RNA-guided DNA cleavage fast rate constant, which is insignificant compared to the ~167- and ~17-fold reductions in the Mn<sup>2+</sup>-mediated gRNA-free ssDNA cleavage rate constant. In all, the SpyCas9<sup>H982A</sup> variant is a safer and comparably efficient for gene therapy applications relative to SpyCas9<sup>WT</sup>.



#### 5.1.1.2 Elucidating the Mechanism of SpyCas9<sup>2Pro</sup>'s reduced off-target DNA cleavage

The SpyCas9<sup>2Pro</sup> variant has already been established as being more selective for on-target RNA-guided dsDNA linearization relative to the WT when there is a PAM-proximal mismatch between the gRNA and the target DNA. Kinetic analyses also have established that the increase in on-target selectivity is due to SpyCas9<sup>2Pro</sup> being forced to use an alternative, and slower, linearization pathway than is used by SpyCas9<sup>WT</sup> to linearize dsDNA with a mismatch in a PAM-proximal position. The reasons behind why SpyCas9<sup>2Pro</sup> has to use this slower pathway to cleave dsDNA with PAM-proximal mismatches is investigated in this work, and it was hypothesized that the 2Pro substitution impairs a loop-to-helix transition that is necessary to allow the PAM-proximal mismatched dsDNA cleavage reaction to proceed down the faster pathway. An alternative hypothesis is that the loss of intra-protein interactions from the K65 residue that is substituted for proline in SpyCas9<sup>2Pro</sup> is what causes the change in on-target DNA cleavage selectivity.

The experiments necessary to test these hypotheses are 1) *in vitro* matched and mismatched DNA cleavage assays with SpyCas9<sup>WT</sup>, SpyCas9<sup>2Pro</sup> (a variant with impaired loop-to-helix transition in the BH upon gRNA-binding and no intra-protein electrostatic interactions in the K65P substituted site), and SpyCas9<sup>2Ala</sup> (a variant with a robust loop-to-helix transition in the BH upon gRNA-binding, but still no intra-protein electrostatic interactions in the K65A substituted site); and 2) EPR studies to measure the loop-to-helix transition upon gRNA-binding by detecting the change in distance between two labels on either side of the loop region expected to transition to a helix. This experiment with the EPR measurements is necessary to confirm that the 2Pro and 2Ala substitutions have a measurable effect on the BH structure as expected. These studies may also give insight into how the 2Pro and 2Ala substitutions affect the Cas9 population distribution of various structural states. For example, the 2Pro substitution may broaden the distribution of the population of Cas9s in either the loop or the helix

form because the protein is slipping in and out of the helix form due to the proline disruption.

Towards the end of making these distance measurements using EPR, eight SpyCas9 variants amenable to nitroxide labeling at the specific residues (E60C and C80) on either side of the loop region that transitions to a helix were constructed, purified, and assayed for RNA-guided DNA cleavage activity. Confirming that the SpyCas9 variants are active for RNA-guided DNA cleavage ensured that the substitutions that prepare the protein for labeling do not affect the structure of the active site, so distance measurements obtained from these mutants will accurately represent the effects of the 2Pro and 2Ala substitutions and not the substitutions that facilitate nitroxide labeling.

The result of the *in vitro* assays is that, regarding RNA-guided DNA cleavage kinetics, an unexpected trend is observed where the selectivity of SpyCas9<sup>2Ala</sup> for on-target DNA cleavage over substrates with a PAM-proximal mismatch is greater than that of both SpyCas9<sup>WT</sup> and SpyCas9<sup>2Pro</sup>. Furthermore, the RNA-guided DNA cleavage rate constants themselves showed that SpyCas9<sup>2Ala</sup> cleaves all types of DNA substrates faster than does SpyCas9<sup>WT</sup>. A different trend is observed regarding the RNA-guided DNA cleavage linearization efficiency, which indicates which substrates the SpyCas9 variants tend to fully convert to product (as opposed to the rate constants which indicate how quickly product was produced). The observed RNA-guided DNA cleavage linearization efficiency selectivity trend matched the expected inverse relationship with the BH helicity trend, which is that SpyCas9<sup>2Pro</sup> < SpyCas9<sup>WT</sup> < SpyCas9<sup>2Ala</sup> in term of helical tendency and SpyCas9<sup>2Ala</sup> < SpyCas9<sup>WT</sup> < SpyCas9<sup>2Pro</sup> in terms of selectivity for matched over mismatched substrate in terms of DNA cleavage linearization efficiency.

For biotechnology application purposes, the DNA cleavage linearization efficiency is more relevant than the cleavage kinetics because the main concern

is the amount of off-target DNA cleavage that occurs in the mammalian cell genome. Since linearization efficiency will be used as the main indicator for on-target RNA-guided DNA cleavage selectivity, it appears that integrity of the BH (not lost intra-protein interactions from K65 substitution) is the determinant of SpyCas9<sup>2Pro</sup>'s selectivity. Further studies of different BH substitution variants (besides 2Pro and 2Ala), such as substituting a negatively charged residue in the place of positively charged K65, can be done to develop even safer SpyCas9-based gene therapy options. This negatively charged residue-based future research is especially interesting because it could exaggerate the effect that 2Ala had on the RNA-guided DNA cleavage rate constants. The 2Ala substitution changed K65 from positive to neutral, which caused an increase in on-target DNA cleavage selectivity in terms of RNA-guided DNA cleavage rate constants. If K65 is substituted for a negatively charged residue as opposed to a neutral one, the effect may be even more extreme. This is relevant for biotechnology applications because one of the effects of the 2Ala substitution was a drastic increase in overall RNA-guided DNA cleavage rate constant accompanied by some increase in on-target selectivity, which could be useful in some applications.

#### 5.1.2 Coronavirus spike protein project: Prioritizing antiviral targets in coronavirus-2 (CoV-2) spike using sequence conservation analysis

Four approaches were used to identify antiviral target sites in coronavirus spike protein. Firstly, spike sequence conservation analysis was used to identify residues that are predicted to be important to the coronavirus infection mechanism or specifically to the adaptation of CoV-2 to infecting human as opposed to zoonotic hosts. Secondly, molecular docking analysis was used to identify residues that could be targeted by small molecules. Thirdly, biochemical data on how crucial RBD residues are for spike-ACE2 binding was used to identify which residues mediate the initial stage of infection with CoV-2. In all, it was found that residues such as G502 are highly conserved, essential for ACE2-

binding, and located in a pocket in spike that is targetable using small-molecule drugs.

## 5.2.0 Future directions

### 5.2.1 High fidelity Cas9 variants for safer gene therapy tools

#### 5.2.1.1 SpyCas9 variants with reduced gRNA-free DNA cleavage activity

##### 5.2.1.1.1 Visualizing the effect of SpyCas9<sup>H982A</sup> in mammalian cell nuclei using fluorescence microscopy

A fluorescence microscopy experiment was done by Saha *et al.* with CjeCas9 and SpyCas9 to observe whether guide-free Cas9 damages genomic DNA in mammalian cell nuclei (Saha et al., 2020c). Both proteins damage the genomic DNA in the mammalian cell nuclei, so to test if the H982A substitution in SpyCas9 has significant impact on the safety of SpyCas9 for gene therapy applications, a next step in the Rajan lab is to do this fluorescence microscopy assay with SpyCas9<sup>H982A</sup>. Based on the reduction in gRNA-free DNA cleavage activity of SpyCas9<sup>H982A</sup> *in vitro*, the expected result is that the H982A variant will cause less damage to the mammalian cell genome relative to that caused by the wild-type SpyCas9.

##### 5.2.1.1.2 Different substitution variants to further reduce gRNA-free DNA cleavage activity (especially in the presence of Co<sup>2+</sup>)

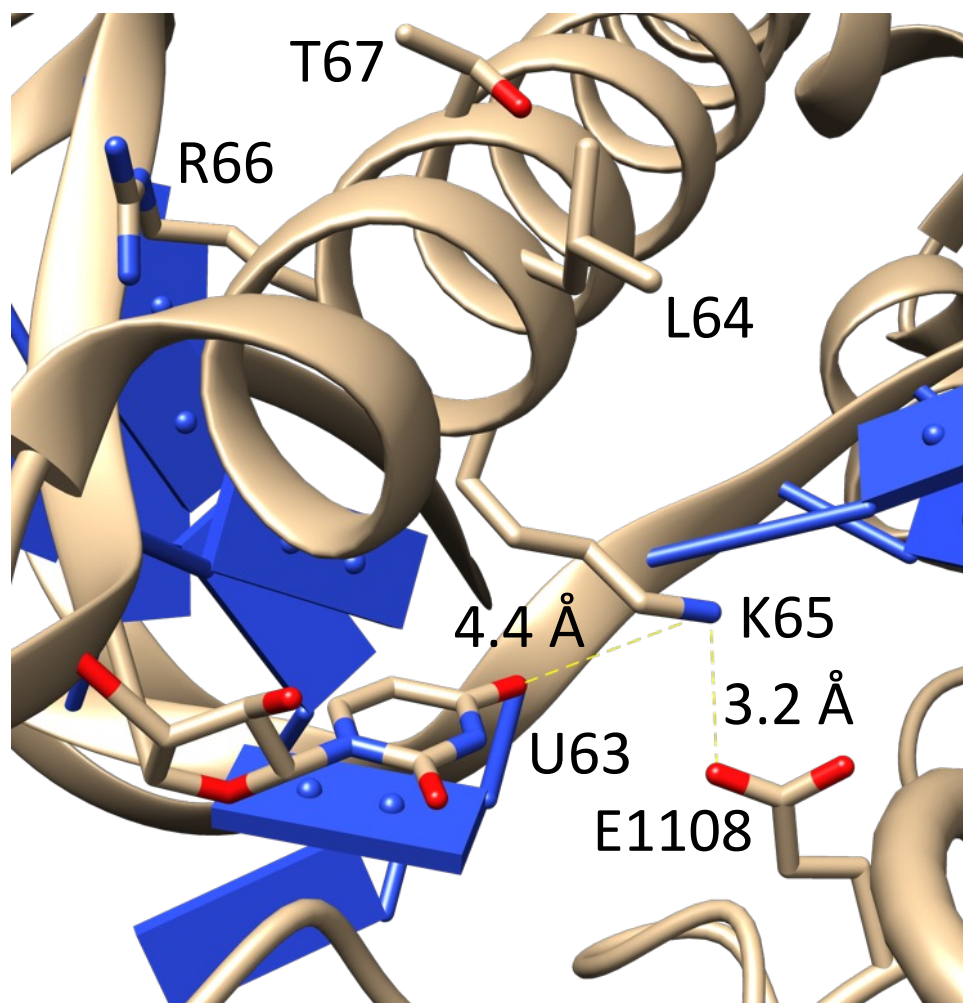
Since the H982A substitution reduced the Mn<sup>2+</sup>-mediated RNA-guided DNA cleavage rate constant ~167-fold, and the Co<sup>2+</sup>-mediated RNA-guided DNA cleavage rate constant ~17-fold, it is of interest to identify a variant that can further reduce the Co<sup>2+</sup>-mediated activity. This could possibly be achieved using different substitutions (besides that of alanine) in the H982 position. Another approach is to use molecular dynamics simulations to identify different residue positions that may be helping structure the RuvC active site for gRNA-free DNA cleavage. Using this method, S15A and L52A have been identified as substitutions with potential to remove the SpyCas9 gRNA-free DNA cleavage activity. A future direction in the Rajan lab is to purify these

variants and assay them for gRNA-free and RNA-guided DNA cleavage activities.

#### 5.2.1.2 SpyCas9 variants with reduced off-target DNA cleavage activity

##### 5.2.1.2.1 Negatively charged substitution variants of K65 in the SpyCas9 BH

To plan what type of substitution may elicit this even more drastic effect on RNA-guided DNA cleavage rate constant and on-target selectivity, the intra-protein and protein-nucleic acid interactions of K65 were structurally observed in SpyCas9 bound to gRNA (PDB ID 4ZT0; note that the same results were obtained with PDB ID: 5F9R (Jiang et al., 2016), which is SpyCas9 bound to gRNA and substrate DNA) (Figure 5.1).



**Figure 5.1. Structural observation of the K65 intra-protein and protein-nucleic acid interactions.** This structure is PDB ID 4ZT0 (Jiang et al., 2015),

which is SpyCas9 bound to gRNA. The residues within a 5 Å zone of the K65 sidechain nitrogen are Uracil 63 (U63) and glutamate 1108 (E1108), which are 4.4 Å and 3.2 Å away respectively. E1108 is in the phosphate lock loop region of SpyCas9 and U63 is in the stem loop region of the gRNA.

The two interactions of K65 are in the SpyCas9 phosphate lock loop, which facilitates substrate DNA-gRNA hybridization, and the gRNA stem loop 1 which is from the tracrRNA and is alternatively named the nexus (Jiang et al., 2015) (Briner et al., 2014). The nexus (stem loop 1) extensively interacts with SpyCas9 in the pre- and post-DNA bound structure and is essential for formation of the SpyCas9-gRNA complex (Jiang et al., 2015; Nishimasu et al., 2014). Some substitutions in the nexus reduce SpyCas9 RNA-guided DNA cleavage activity but substituting nearby nucleic acid residue cytosine 61 is tolerated by SpyCas9 (Briner et al., 2014). This information about phosphate lock loop and the nexus indicates that substituting K65 for a negatively charged residue has potential to modulate formation of a functional Cas9-gRNA complex that can hybridize with the substrate DNA at the 20 nt guide region of the gRNA, so this is a future direction for the Rajan lab to further develop a safer SpyCas9 variant for gene therapy applications.

- 5.2.2 Small molecule drugs to prevent infection from present and future CoV-2 strains  
To identify if targeting the pocket where G502 is located can impair the ability of CoV-2 to infect humans, a future direction in the Rajan lab is to do pseudoviral entry assays with the small molecule that binds in that pocket (NSC 61610 in pocket 4).

An additional future direction for this spike project is to take advantage of the spike sequence and CoV-2 infection case number data that has been collected over time since 2020 to correlate mutations in spike protein with the emergence of new CoV-2 variants that caused surges in the number of infection cases. MSAs will be constructed with CoV-2 spike sequences collected in a time range

immediately following a case surge (which is expected to correlate with the emergence of a new CoV-2 variant). The changes in conservation between the various MSAs will be used to identify which sites are more stable drug targets and which sites are more variable. These results will be cross checked with the sequence conservation findings from this study of CoV-2 sequences from the year 2020, which are the CoV-2 sequences that were compared to SARS and MERS sequences to identify conserved and stable sites. If the sites identified as stable in this study of CoV-2 spike sequences from 2020 are confirmed to remain stable even in new CoV-2 variants, then we will have established sequence conservation analysis with related viral species as a useful method for identifying stable drug targets in novel viruses.

## References

- Abudayyeh, O.O., Gootenberg, J.S., Essletzbichler, P., Han, S., Joung, J., Belanto, J.J., Verdine, V., Cox, D.B.T., Kellner, M.J., Regev, A., *et al.* (2017). RNA targeting with CRISPR–Cas13. *Nature* 550, 280-284.
- Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., *et al.* (2016a). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, aaf5573.
- Abudayyeh, O.O., Gootenberg, J.S., Konermann, S., Joung, J., Slaymaker, I.M., Cox, D.B.T., Shmakov, S., Makarova, K.S., Semenova, E., Minakhin, L., *et al.* (2016b). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* 353, aaf5573.
- Acharya, S., Mishra, A., Paul, D., Ansari, A.H., Azhar, M., Kumar, M., Rauthan, R., Sharma, N., Aich, M., Sinha, D., *et al.* (2019). Francisella novicida Cas9 interrogates genomic DNA with very high specificity and can be used for mammalian genome editing. *Proc Natl Acad Sci U S A* 116, 20959-20968.
- Agrawal, A., Varshney, R., Pathak, M., Patel, S.K., Rai, V., Sulabh, S., Gupta, R., Solanki, K.S., Varshney, R., and Nimmanapalli, R. (2021). Exploration of antigenic determinants in spike glycoprotein of SARS-CoV2 and identification of five salient potential epitopes. *VirusDisease* 32, 774-783.
- Aklujkar, M., and Lovley, D.R. (2010). Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of *Pelobacter carbinolicus*. *BMC Evolutionary Biology* 10, 230.
- Aleem, A., Akbar Samad, A.B., and Slenker, A.K. (2022). Emerging Variants of SARS-CoV-2 And Novel Therapeutics Against Coronavirus (COVID-19). In StatPearls (Treasure Island (FL): StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.).
- Ali, Z., Abulfaraj, A., Idris, A., Ali, S., Tashkandi, M., and Mahfouz, M.M. (2015). CRISPR/Cas9-mediated viral interference in plants. *Genome Biology* 16, 238.
- Alkhnbashi, O.S., Saunders, S.J., Costa, F., Shah, S.A., Garrett, R.A., and Backofen, R. (2016). Characterizing leader sequences of CRISPR loci. *Bioinformatics* 32, i576-i585.
- Allam, L., Ghrifi, F., Mohammed, H., El Hafidi, N., El Jaoudi, R., El Harti, J., Lmimouni, B., Belyamani, L., and Ibrahimi, A. (2020). Targeting the GRP78-Dependant SARS-CoV-2 Cell Entry by Peptides and Small Molecules. *Bioinformatics and Biology Insights* 14, 1177932220965505.
- Allen, D., Rosenberg, M., and Hendel, A. (2020). Using Synthetically Engineered Guide RNAs to Enhance CRISPR Genome Editing Systems in Mammalian Cells. *Front Genome Ed* 2, 617910.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403-410.
- Aman, R., Ali, Z., Butt, H., Mahas, A., Aljedaani, F., Khan, M.Z., Ding, S., and Mahfouz, M. (2018). RNA virus interference via CRISPR/Cas13a system in plants. *Genome Biology* 19, 1.



Amitai, G., and Sorek, R. (2016). CRISPR-Cas adaptation: insights into the mechanism of action. *Nat Rev Microbiol* 14, 67-76.

Anders, C., Niewoehner, O., Duerst, A., and Jinek, M. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 513, 569-573.

Andersen, K.G., Rambaut, A., Lipkin, W.I., Holmes, E.C., and Garry, R.F. (2020). The proximal origin of SARS-CoV-2. *Nature Medicine* 26, 450-452.

Arbeitman, C.R., Auge, G., Blaustein, M., Bredeston, L., Corapi, E.S., Craig, P.O., Cossio, L.A., Dain, L., D'Alessio, C., Elias, F., *et al.* (2020). Structural and functional comparison of SARS-CoV-2-spike receptor binding domain produced in *Pichia pastoris* and mammalian cells. *Sci Rep-Uk* 10, 21779.

Ashour, H.M., Elkhatab, W.F., Rahman, M.M., and Elshabrawy, H.A. (2020). Insights into the Recent 2019 Novel Coronavirus (SARS-CoV-2) in Light of Past Human Coronavirus Outbreaks. In *Pathogens*.

Babu, K., Amrani, N., Jiang, W., Yogesha, S.D., Nguyen, R., Qin, P.Z., and Rajan, R. (2019). Bridge Helix of Cas9 Modulates Target DNA Cleavage and Mismatch Tolerance. *Biochemistry* 58, 1905-1917.

Babu, K., Kathiresan, V., Kumari, P., Newsom, S., Parameshwaran, H.P., Chen, X., Liu, J., Qin, P.Z., and Rajan, R. (2021). Coordinated Actions of Cas9 HNH and RuvC Nuclease Domains Are Regulated by the Bridge Helix and the Target DNA Sequence. *Biochemistry* 60, 3783-3800.

Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., *et al.* (2011). A dual function of the CRISPR-Cas system in bacterial antiviral immunity and DNA repair. *Mol Microbiol* 79, 484-502.

Bangaru, S., Ozorowski, G., Turner, H.L., Antanasijevic, A., Huang, D., Wang, X., Torres, J.L., Diedrich, J.K., Tian, J.-H., Portnoff, A.D., *et al.* (2020). Structural analysis of full-length SARS-CoV-2 spike protein from an advanced vaccine candidate. *Science* 370, 1089-1094.

Barnes, C.O., Jette, C.A., Abernathy, M.E., Dam, K.-M.A., Esswein, S.R., Gristick, H.B., Malyutin, A.G., Sharaf, N.G., Huey-Tubman, K.E., Lee, Y.E., *et al.* (2020). SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* 588, 682-687.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science* 315, 1709-1712.

Batra, R., Nelles, D.A., Pirie, E., Blue, S.M., Marina, R.J., Wang, H., Chaim, I.A., Thomas, J.D., Zhang, N., Nguyen, V., *et al.* (2017). Elimination of Toxic Microsatellite Repeat Expansion RNA by RNA-Targeting Cas9. *Cell* 170, 899-912.e810.

Behler, J., and Hess, W.R. (2020). Approaches to study CRISPR RNA biogenesis and the key players involved. *Methods* 172, 12-26.

Beloglazova N, B.G., Zimmerman MD, Proudfoot M, Makarova KS, Kudritska M, Kochinyan S, Wang S, Chruszcz M, Minor W, Koonin EV, Edwards AM, Savchenko A, Yakunin AF. (2008). A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. *J Biol Chem* 283, 20361-20371.

Bernheim, A., Calvo-Villamañán, A., Basier, C., Cui, L., Rocha, E.P.C., Touchon, M., and Bikard, D. (2017). Inhibition of NHEJ repair by type II-A CRISPR-Cas systems in bacteria. *Nat Commun* 8, 2094.

Blaber, M., Zhang, X.-j., and Matthews, B.W. (1993). Structural Basis of Amino Acid  $\alpha$  Helix Propensity. *Science* 260, 1637-1640.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005a). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151, 2551-2561.

Bolotin, A., Quinquis, B., Sorokin, A., and Ehrlich, S.D. (2005b). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology* 151, 2551-2561.

Brandsma, I., and van Gent, D.C. (2012). Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integrity* 3, 9.

Breinig, M., Schweitzer, A.Y., Herianto, A.M., Revia, S., Schaefer, L., Wendler, L., Cobos Galvez, A., and Tschaharganeh, D.F. (2019). Multiplexed orthogonal genome editing and transcriptional activation by Cas12a. *Nature Methods* 16, 51-54.

Briner, Alexandra E., Donohoue, Paul D., Goma, Ahmed A., Selle, K., Slorach, Euan M., Nye, Christopher H., Haurwitz, Rachel E., Beisel, Chase L., May, Andrew P., and Barrangou, R. (2014). Guide RNA Functional Modules Direct Cas9 Activity and Orthogonality. *Molecular Cell* 56, 333-339.

Brouns, S.J., Jore, M.M., Lundgren, M., Westra, E.R., Slijkhuis, R.J., Snijders, A.P., Dickman, M.J., Makarova, K.S., Koonin, E.V., and van der Oost, J. (2008). Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science* 321, 960-964.

Burstein, D., Harrington, L.B., Strutt, S.C., Probst, A.J., Anantharaman, K., Thomas, B.C., Doudna, J.A., and Banfield, J.F. (2017). New CRISPR-Cas systems from uncultivated microbes. *Nature* 542, 237-241.

Cameron, P., Coons, M.M., Klompe, S.E., Lied, A.M., Smith, S.C., Vidal, B., Donohoue, P.D., Rotstein, T., Kohrs, B.W., Nyer, D.B., *et al.* (2019). Harnessing type I CRISPR–Cas systems for genome engineering in human cells. *Nature Biotechnology* 37, 1471-1477.

Capra, J., and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics* 23, 1875-1882.

Carino, A., Moraca, F., Fiorillo, B., Marchianò, S., Sepe, V., Biagioli, M., Finamore, C., Bozza, S., Francisci, D., Distrutti, E., *et al.* (2020). Hijacking SARS-CoV-2/ACE2 Receptor Interaction by Natural and Semi-synthetic Steroidal Agents Acting on Functional Pockets on the Receptor Binding Domain. *Frontiers in Chemistry* 8.

Casalino, L., Nierzwicki, L., Jinek, M., and Palermo, G. (2020). Catalytic Mechanism of Non-Target DNA Cleavage in CRISPR-Cas9 Revealed by *Ab Initio* Molecular Dynamics. *Acs Catal* 10, 13596-13605.

Case, D.A.B., R. M.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Giese, T.J.; Gohlke, H.; Goetz, A. W.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Luo, R.; Madej, B.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Nguyen, H. T.; Omelyan, I.; Onufriev, A.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R.C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; Kollman, P. A. (2016). AMBER 2016.

Chakraborty, C., Sharma, A.R., Bhattacharya, M., Agoramoorthy, G., and Lee, S.-S. (2021). The Drug Repurposing for COVID-19 Clinical Trials Provide Very Effective Therapeutic Combinations: Lessons Learned From Major Clinical Studies. *Frontiers in Pharmacology* 12.

Charpentier, E., Richter, H., van der Oost, J., and White, M.F. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. *FEMS Microbiology Reviews* 39, 428-441.

Chatterjee, P., Jakimo, N., and Jacobson, J.M. Minimal PAM specificity of a highly similar SpCas9 ortholog. *Science Advances* 4, eaau0766.

Chen, B., Gilbert, Luke A., Cimini, Beth A., Schnitzbauer, J., Zhang, W., Li, G.-W., Park, J., Blackburn, Elizabeth H., Weissman, Jonathan S., Qi, Lei S., *et al.* (2013). Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* 155, 1479-1491.

Chen, J.S., Dagdas, Y.S., Kleinstiver, B.P., Welch, M.M., Sousa, A.A., Harrington, L.B., Sternberg, S.H., Joung, J.K., Yildiz, A., and Doudna, J.A. (2017). Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature* 550, 407-410.

Chen, J.S., Ma, E., Harrington, L.B., Da Costa, M., Tian, X., Palefsky, J.M., and Doudna, J.A. (2018). CRISPR-Cas12a target binding unleashes indiscriminate single-stranded DNase activity. *Science* 360, 436-439.

Cho, S.W., Kim, S., Kim, J.M., and Kim, J.-S. (2013). Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. *Nature Biotechnology* 31, 230-232.

Choi, B., Choudhary, M.C., Regan, J., Sparks, J.A., Padera, R.F., Qiu, X., Solomon, I.H., Kuo, H.-H., Boucau, J., Bowman, K., *et al.* (2020). Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host. *N Engl J Med* 383, 2291-2293.

Chowdhury, R., Boorla, V.S., and Maranas, C.D. (2020). Computational biophysical characterization of the SARS-CoV-2 spike protein binding with the ACE2 receptor and implications for infectivity. *Computational and Structural Biotechnology Journal* 18, 2573-2582.

Chu, V.T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., and Kühn, R. (2015). Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature Biotechnology* 33, 543-548.

Citorik, R.J., Mimee, M., and Lu, T.K. (2014). Sequence-specific antimicrobials using efficiently delivered RNA-guided nucleases. *Nature Biotechnology* 32, 1141-1145.

Cocozaki, A.I., Ramia, N.F., Shao, Y., Hale, C.R., Terns, R.M., Terns, M.P., and Li, H. (2012). Structure of the Cmr2 subunit of the CRISPR-Cas RNA silencing complex. *Structure* 20, 545-553.

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., and Zhang, F. (2017). RNA editing with CRISPR-Cas13. *Science (New York, NY)* 358, 1019-1027.

Crowley, V.M., Catching, A., Taylor, H.N., Borges, A.L., Metcalf, J., Bondy-Denomy, J., and Jackson, R.N. (2019). A Type IV-A CRISPR-Cas System in *Pseudomonas aeruginosa* Mediates RNA-Guided Plasmid Interference In Vivo. *The CRISPR Journal* 2, 434-440.

Cui, L., Wang, X., Huang, D., Zhao, Y., Feng, J., Lu, Q., Pu, Q., Wang, Y., Cheng, G., Wu, M., *et al.* (2020). CRISPR-cas3 of *Salmonella* Upregulates Bacterial Biofilm

Formation and Virulence to Host Cells by Targeting Quorum-Sensing Systems. *Pathogens* 9, 53.

Cullot, G., Boutin, J., Toutain, J., Prat, F., Pennamen, P., Rooryck, C., Teichmann, M., Rousseau, E., Lamrissi-Garcia, I., Guyonnet-Duperat, V., *et al.* (2019). CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat Commun* 10, 1136.

Dagdas, Y.S., Chen, J.S., Sternberg, S.H., Doudna, J.A., and Yildiz, A. (2017). A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Science Advances* 3.

Davis, K.M., Pattanayak, V., Thompson, D.B., Zuris, J.A., and Liu, D.R. (2015). Small molecule-triggered Cas9 protein with improved genome-editing specificity. *Nature Chemical Biology* 11, 316-318.

Deltcheva, E., Chylinski, K., Sharma, C.M., Gonzales, K., Chao, Y., Pirzada, Z.A., Eckert, M.R., Vogel, J., and Charpentier, E. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. *Nature* 471, 602-607.

Dimonte, S., Babakir-Mina, M., Hama-Soor, T., and Ali, S. (2021). Genetic Variation and Evolution of the 2019 Novel Coronavirus. *Public Health Genomics* 24, 54-66.

Ding, Y., Kathiresan, V., Zhang, X., Haworth, I.S., and Qin, P.Z. (2019). Experimental Validation of the ALLNOX Program for Studying Protein-Nucleic Acid Complexes. *J Phys Chem A* 123, 3592-3598.

Dolan, A.E., Hou, Z., Xiao, Y., Gramelspacher, M.J., Heo, J., Howden, S.E., Freddolino, P.L., Ke, A., and Zhang, Y. (2019). Introducing a Spectrum of Long-Range Genomic Deletions in Human Embryonic Stem Cells Using Type I CRISPR-Cas. *Molecular cell* 74, 936-950.e935.

Dow, L.E., Fisher, J., O'Rourke, K.P., Muley, A., Kastenhuber, E.R., Livshits, G., Tschaharganeh, D.F., Socci, N.D., and Lowe, S.W. (2015). Inducible in vivo genome editing with CRISPR-Cas9. *Nature Biotechnology* 33, 390-394.

Durmaz, B., Abdulmajed, O., and Durmaz, R. (2020). Mutations Observed in the SARS-CoV-2 Spike Glycoprotein and Their Effects in the Interaction of Virus with ACE-2 Receptor. *Medeni Med J* 35, 253-260.

Dürr SL, B.O., Suardiaz R, Jambrina PG, Peter C, Shao Y, et al. The Dual Role of Histidine as General Base and Recruiter of a Third Metal Ion in HIV-1 RNase H. *ChemRxiv*. Cambridge: Cambridge Open Engage; 2019; This content is a preprint and has not been peer-reviewed. (2019). The dual role of histidine as general base and recruiter of a third metal ion in HIV-1 RNase H. *ChemRxiv* Cambridge: Cambridge Open Engage; 2019.

East, K.W., Newton, J.C., Morzan, U.N., Narkhede, Y.B., Acharya, A., Skeens, E., Jogl, G., Batista, V.S., Palermo, G., and Lisi, G.P. (2020). Allosteric Motions of the CRISPR-Cas9 HNH Nuclease Probed by NMR and Molecular Dynamics. *J Am Chem Soc* 142, 1348-1358.

East-Seletsky, A., O'Connell, M.R., Knight, S.C., Burstein, D., Cate, J.H.D., Tjian, R., and Doudna, J.A. (2016). Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature* 538, 270-273.

Eberhardt, J., Santos-Martins, D., Tillack, A.F., and Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling* 61, 3891-3898.

Edraki, A., Mir, A., Ibraheim, R., Gainetdinov, I., Yoon, Y., Song, C.-Q., Cao, Y., Gallant, J., Xue, W., Rivera-Pérez, J.A., *et al.* (2019). A Compact, High-Accuracy Cas9 with a Dinucleotide PAM for In Vivo Genome Editing. *Molecular Cell* 73, 714-726.e714.

Engstrom, J.U., and Kmiec, E.B. (2008). DNA Replication, cell cycle progression and the targeted gene repair reaction. *Cell Cycle* 7, 1402-1414.

Faure, G., Makarova, K.S., and Koonin, E.V. (2019). CRISPR–Cas: Complex Functional Networks and Multiple Roles beyond Adaptive Immunity. *Journal of Molecular Biology* 431, 3-20.

Fedorova, I., Vasileva, A., Selkova, P., Abramova, M., Arseniev, A., Pobegalov, G., Kazalov, M., Musharova, O., Goryanin, I., Artamonova, D., *et al.* (2020). PpCas9 from *Pasteurella pneumotropica* — a compact Type II-C Cas9 ortholog active in human cells. *Nucleic Acids Research* 48, 12297-12309.

Feenstra, K.A.H., B.; Berendsen, H. J. C. (1999). Improving Efficiency of Large Time-Scale Molecular Dynamics Simulations of Hydrogen-Rich Systems. *J Comput Chem* 20, 786-798.

Ferenczi, A., Pyott, D.E., Xipnitou, A., and Molnar, A. (2017). Efficient targeted DNA editing and replacement in *Chlamydomonas reinhardtii* using Cpf1 ribonucleoproteins and single-stranded DNA. *Proceedings of the National Academy of Sciences* 114, 13567.

Ferry, Q.R.V., Lyutova, R., and Fulga, T.A. (2017). Rational design of inducible CRISPR guide RNAs for de novo assembly of transcriptional programs. *Nat Commun* 8, 14633.

Flusche, T., and Rajan, R. Molecular Details of DNA Integration by CRISPR-Associated Proteins During Adaptation in Bacteria and Archaea. In (Cham: Springer International Publishing), pp. 1-17.

Fonfara, I., Richter, H., Bratovic, M., Le Rhun, A., and Charpentier, E. (2016). The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* 532, 517-521.

Forcelloni, S., Benedetti, A., Dilucca, M., and Giansanti, A. (2021). Identification of Conserved Epitopes in SARS-CoV-2 Spike and Nucleocapsid Protein. *Current Genomics* 22, 541-549.

Fu, Y., Foden, J.A., Khayter, C., Maeder, M.L., Reyon, D., Joung, J.K., and Sander, J.D. (2013). High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol* 31, 822-826.

Fujita, T., and Fujii, H. (2015). Isolation of Specific Genomic Regions and Identification of Associated Molecules by Engineered DNA-Binding Molecule-Mediated Chromatin Immunoprecipitation (enChIP) Using CRISPR. In *Chromatin Protocols*, S.P. Chellappan, ed. (New York, NY: Springer New York), pp. 43-52.

Furuhata, Y., and Kato, Y. (2021). Asymmetric Roles of Two Histidine Residues in *Streptococcus pyogenes* Cas9 Catalytic Domains upon Chemical Rescue. *Biochemistry* 60, 194-200.

Gadanec, L.K., McSweeney, K.R., Qaradakh, T., Ali, B., Zulli, A., and Apostolopoulos, V. (2021). Can SARS-CoV-2 Virus Use Multiple Receptors to Enter Host Cells? In *International Journal of Molecular Sciences*.

Gaj, T., Gersbach, C.A., and Barbas, C.F., 3rd (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. *Trends Biotechnol* 31, 397-405.

Galindo-Murillo, R., Robertson, J.C., Zgarbova, M., Sponer, J., Otyepka, M., Jurecka, P., and Cheatham, T.E., 3rd (2016). Assessing the Current State of Amber Force Field Modifications for DNA. *J Chem Theory Comput* 12, 4114-4127.

Gao, N., Zhang, C., Hu, Z., Li, M., Wei, J., Wang, Y., and Liu, H. (2020). Characterization of *Brevibacillus laterosporus* Cas9 (BlatCas9) for mammalian genome editing. *Frontiers in cell and developmental biology* 8, 583164.

Gao, N.J., Al-Bassam, M.M., Poudel, S., Wozniak, J.M., Gonzalez, D.J., Olson, J., Zengler, K., Nizet, V., and Valderrama, J.A. (2019). Functional and Proteomic Analysis of *Streptococcus pyogenes* Virulence Upon Loss of Its Native Cas9 Nuclease. *Frontiers in Microbiology* 10, 1967.

Gasiunas, G., Barrangou, R., Horvath, P., and Siksnys, V. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 109, E2579-E2586.

Gaudelli, N.M., Komor, A.C., Rees, H.A., Packer, M.S., Badran, A.H., Bryson, D.I., and Liu, D.R. (2017). Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* 551, 464-471.

Gleditsch, D., Pausch, P., Müller-Esparza, H., Özcan, A., Guo, X., Bange, G., and Randau, L. (2019). PAM identification by CRISPR-Cas effector complexes: diversified mechanisms and structures. *RNA biology* 16, 504-517.

Goldberg, G.W., Jiang, W., Bikard, D., and Marraffini, L.A. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. *Nature* 514, 633-637.

Gomaa, A.A., Klumpe, H.E., Luo, M.L., Selle, K., Barrangou, R., and Beisel, C.L. (2014). Programmable removal of bacterial strains by use of genome-targeting CRISPR-Cas systems. *mBio* 5, e00928.

Gong, S.Z., Yu, H.H., Johnson, K.A., and Taylor, D.W. (2018). DNA Unwinding Is the Primary Determinant of CRISPR-Cas9 Activity. *Cell Reports* 22, 359-371.

Gootenberg, J.S., Abudayyeh, O.O., Kellner, M.J., Joung, J., Collins, J.J., and Zhang, F. (2018). Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6. *Science* 360, 439.

Gui, M., Song, W., Zhou, H., Xu, J., Chen, S., Xiang, Y., and Wang, X. (2017). Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Research* 27, 119-129.

Gunderson, F.F., and Cianciotto, N.P. (2013). The CRISPR-Associated Gene *cas2* of *Legionella pneumophila* Is Required for Intracellular Infection of Amoeboe. *mBio* 4, e00074-00013.

Gussow, A.B., Auslander, N., Faure, G., Wolf, Y.I., Zhang, F., and Koonin, E.V. (2020). Genomic determinants of pathogenicity in SARS-CoV-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences* 117, 15193.

Haapaniemi, E., Botla, S., Persson, J., Schmierer, B., and Taipale, J. (2018). CRISPR–Cas9 genome editing induces a p53-mediated DNA damage response. *Nature Medicine* 24, 927-930.

Harrington, L.B., Paez-Espino, D., Staahl, B.T., Chen, J.S., Ma, E., Kyripides, N.C., and Doudna, J.A. (2017). A thermostable Cas9 with increased lifetime in human plasma. *Nat Commun* 8, 1424.

Hati, S., and Bhattacharyya, S. (2020). Impact of Thiol–Disulfide Balance on the Binding of Covid-19 Spike Protein with Angiotensin-Converting Enzyme 2 Receptor. *ACS Omega* 5, 16292-16298.

Heidrich, N., Hagmann, A., Bauriedl, S., Vogel, J., and Schoen, C. (2019). The CRISPR/Cas system in *Neisseria meningitidis* affects bacterial adhesion to human nasopharyngeal epithelial cells. *RNA biology* 16, 390-396.

Heler, R., Marraffini, L.A., and Bikard, D. (2014). Adapting to new threats: the generation of memory by CRISPR-Cas immune systems. *Mol Microbiol* 93, 1-9.

Heussler, G.E., Cady, K.C., Koeppen, K., Bhujji, S., Stanton, B.A., and O'Toole, G.A. (2015). Clustered Regularly Interspaced Short Palindromic Repeat-Dependent, Biofilm-Specific Death of *Pseudomonas aeruginosa* Mediated by Increased Expression of Phage-Related Genes. *mBio* 6, e00129-00115.

Hopkins, C.W., Le Grand, S., Walker, R.C., and Roitberg, A.E. (2015). Long-Time-Step Molecular Dynamics through Hydrogen Mass Repartitioning. *J Chem Theory Comput* 11, 1864-1874.

Horvath, P., Romero, D.A., Coute-Monvoisin, A.C., Richards, M., Deveau, H., Moineau, S., Boyaval, P., Fremaux, C., and Barrangou, R. (2008). Diversity, activity, and evolution of CRISPR loci in *Streptococcus thermophilus*. *J Bacteriol* 190, 1401-1412.

Hou, X., Zaks, T., Langer, R., and Dong, Y. (2021). Lipid nanoparticles for mRNA delivery. *Nature Reviews Materials* 6, 1078-1094.

Hou, Z., Zhang, Y., Propson, N.E., Howden, S.E., Chu, L.F., Sontheimer, E.J., and Thomson, J.A. (2013). Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proc Natl Acad Sci U S A*.

Hu, G., Yan, W., Zhou, J., and Shen, B. (2014). Residue interaction network analysis of Dronpa and a DNA clamp. *Journal of Theoretical Biology* 348, 55-64.

Hu, Z., Wang, S., Zhang, C., Gao, N., Li, M., Wang, D., Wang, D., Liu, H., Ong, S.-G., *et al.* (2020). A compact Cas9 ortholog from *Staphylococcus Auricularis* (SauriCas9) expands the DNA targeting scope. *PLOS Biology* 18, e3000686.

Huang, Y., Sun, H., Yu, H., Li, S., Zheng, Q., and Xia, N. (2020a). Neutralizing antibodies against SARS-CoV-2: current understanding, challenge and perspective. *Antibody Therapeutics* 3, 285-299.

Huang, Y., Yang, C., Xu, X.-f., Xu, W., and Liu, S.-w. (2020b). Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacologica Sinica* 41, 1141-1149.

Huang, Y.-H., Su, J., Lei, Y., Brunetti, L., Gundry, M.C., Zhang, X., Jeong, M., Li, W., and Goodell, M.A. (2017). DNA epigenome editing using CRISPR-Cas SunTag-directed DNMT3A. *Genome Biology* 18, 176.

Hur, J.K., Kim, K., Been, K.W., Baek, G., Ye, S., Hur, J.W., Ryu, S.-M., Lee, Y.S., and Kim, J.-S. (2016). Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoproteins. *Nature Biotechnology* 34, 807-808.

Ihry, R.J., Worringer, K.A., Salick, M.R., Frias, E., Ho, D., Theriault, K., Kommineni, S., Chen, J., Sondey, M., Ye, C., *et al.* (2018). p53 inhibits CRISPR-Cas9 engineering in human pluripotent stem cells. *Nat Med* 24, 939-946.

Jackson, S.A., McKenzie, R.E., Fagerlund, R.D., Kieper, S.N., Fineran, P.C., and Brouns, S.J.J. (2017). CRISPR-Cas: Adapting to change. *Science* (Washington, DC, U S) 356, 40.

Jiang, F., and Doudna, J.A. (2017). CRISPR-Cas9 Structures and Mechanisms. *Annu Rev Biophys* 46, 505-529.

Jiang, F., Taylor, D.W., Chen, J.S., Kornfeld, J.E., Zhou, K., Thompson, A.J., Nogales, E., and Doudna, J.A. (2016). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science* 351, 867-871.

Jiang, F., Zhou, K., Ma, L., Gressel, S., and Doudna, J.A. (2015). STRUCTURAL BIOLOGY. A Cas9-guide RNA complex preorganized for target DNA recognition. *Science* 348, 1477-1481.

Jiang, W., Bikard, D., Cox, D., Zhang, F., and Marraffini, L.A. (2013). RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nature Biotechnology* 31, 233-239.

Jiang, Y.-P., Zhao, X.-X., Lv, H.-Q., and Wen, C.-P. (2021). Drug screening and development from the affinity of S protein of new coronavirus with ACE2. *European Journal of Clinical Microbiology & Infectious Diseases* 40, 715-723.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* 337, 816-821.

Jinek, M., Jiang, F., Taylor, D.W., Sternberg, S.H., Kaya, E., Ma, E., Anders, C., Hauer, M., Zhou, K., Lin, S., *et al.* (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science* 343, 1247997.

Jore, M.M., Lundgren, M., van Duijn, E., Bultema, J.B., Westra, E.R., Waghmare, S.P., Wiedenheft, B., Pul, Ü., Wurm, R., Wagner, R., *et al.* (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature Structural & Molecular Biology* 18, 529-536.

Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79, 926-935.

Jorgensen, W.L.C., J.; Madura, J. D.; Impey, R. W.; Klein, M. L. (1983). Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* 79, 926-935.

Josephs, E.A., Kocak, D.D., Fitzgibbon, C.J., McMenemy, J., Gersbach, C.A., and Marszalek, P.E. (2016). Structure and specificity of the RNA-guided endonuclease Cas9 during DNA interrogation, target binding and cleavage. *Nucleic Acids Res* 44, 2474.

Joung, I.S., and Cheatham, T.E., 3rd (2008). Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* 112, 9020-9041.

Joung, J., Ladha, A., Saito, M., Kim, N.-G., Woolley, A.E., Segel, M., Barretto, R.P.J., Ranu, A., Macrae, R.K., Faure, G., *et al.* (2020). Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing. *N Engl J Med* 383, 1492-1494.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K.F., von Haeseler, A., and Jermin, L.S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 14, 587-589.



Karvelis, T., Gasiunas, G., Young, J., Bigelyte, G., Silanskas, A., Cigan, M., and Siksnys, V. (2015). Rapid characterization of CRISPR-Cas9 protospacer adjacent motif sequence elements. *Genome Biology* 16, 253.

Kazlauskienė, M., Kostiuk, G., Venclovas, Č., Tamulaitis, G., and Siksnys, V. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. *Science* 357, 605.

Kazlauskienė, M., Tamulaitis, G., Kostiuk, G., Venclovas, Č., and Siksnys, V. (2016). Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. *Molecular Cell* 62, 295-306.

Kelley, C.P., Haerle, M.C., and Wang, E.T. (2022). Negative autoregulation mitigates collateral RNase activity of repeat-targeting CRISPR-Cas13d in mammalian cells. *Cell Reports* 40, 111226.

Khare, S., Azevedo, M., Parajuli, P., and Gokulan, K. (2021). Conformational Changes of the Receptor Binding Domain of SARS-CoV-2 Spike Protein and Prediction of a B-Cell Antigenic Epitope Using Structural Data. *Frontiers in Artificial Intelligence* 4.

Kiga, K., Tan, X.-E., Ibarra-Chávez, R., Watanabe, S., Aiba, Y., Sato'o, Y., Li, F.-Y., Sasahara, T., Cui, B., Kawauchi, M., *et al.* (2020). Development of CRISPR-Cas13a-based antimicrobials capable of sequence-specific killing of target bacteria. *Nat Commun* 11, 2934.

Kim, E., Koo, T., Park, S.W., Kim, D., Kim, K., Cho, H.-Y., Song, D.W., Lee, K.J., Jung, M.H., Kim, S., *et al.* (2017). In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat Commun* 8, 14500.

Kim, H., and Kim, J.-S. (2014). A guide to genome engineering with programmable nucleases. *Nature Reviews Genetics* 15, 321-334.

Kirschner, K.N., Yongye, A.B., Tschampel, S.M., González-Outeiriño, J., Daniels, C.R., Foley, B.L., and Woods, R.J. (2008). GLYCAM06: A generalizable biomolecular force field. *Carbohydrates. Journal of Computational Chemistry* 29, 622-655.

Kleinstiver, B.P., Pattanayak, V., Prew, M.S., Tsai, S.Q., Nguyen, N.T., Zheng, Z., and Joung, J.K. (2016). High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* 529, 490-495.

Klompe, S.E., Vo, P.L.H., Halpin-Healy, T.S., and Sternberg, S.H. (2019). Transposon-encoded CRISPR-Cas systems direct RNA-guided DNA integration. *Nature* 571, 219-225.

Kolesnik, M.V., Fedorova, I., Karneyeva, K.A., Artamonova, D.N., and Severinov, K.V. (2021). Type III CRISPR-Cas Systems: Deciphering the Most Complex Prokaryotic Immune System. *Biochemistry (Moscow)* 86, 1301-1314.

Kollman, S.M.P.A. (1992). Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* 13, 952-962.

Komor, A.C., Kim, Y.B., Packer, M.S., Zuris, J.A., and Liu, D.R. (2016). Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* 533, 420-424.

Konermann, S., Lotfy, P., Brideau, N.J., Oki, J., Shokhirev, M.N., and Hsu, P.D. (2018). Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* 173, 665-676.e614.

Koonin, E.V., Makarova, K.S., and Zhang, F. (2017). Diversity, classification and evolution of CRISPR-Cas systems. *Curr Opin Microbiol* 37, 67-78.

Koonin, E.V., and Wolf, Y.I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* 36, 6688-6719.

Koonin, E.V., and Zhang, F. (2017). Coupling immunity and programmed cell suicide in prokaryotes: Life-or-death choices. *BioEssays : news and reviews in molecular, cellular and developmental biology* 39, e201600186.

Ku, Z., Xie, X., Davidson, E., Ye, X., Su, H., Menachery, V.D., Li, Y., Yuan, Z., Zhang, X., Muruato, A.E., *et al.* (2021). Molecular determinants and mechanism for antibody cocktail preventing SARS-CoV-2 escape. *Nat Commun* 12, 469.

Kuscu, C., Arslan, S., Singh, R., Thorpe, J., and Adli, M. (2014). Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol* 32, 677-683.

Kushawah, G., Hernandez-Huertas, L., Abugattas-Nuñez del Prado, J., Martinez-Morales, J.R., DeVore, M.L., Hassan, H., Moreno-Sanchez, I., Tomas-Gallardo, L., Diaz-Moscoso, A., Monges, D.E., *et al.* (2020). CRISPR-Cas13d Induces Efficient mRNA Knockdown in Animal Embryos. *Developmental Cell* 54, 805-817.e807.

Lee, C.M., Cradick, T.J., and Bao, G. (2016a). The *Neisseria meningitidis* CRISPR-Cas9 System Enables Specific Genome Editing in Mammalian Cells. *Mol Ther* 24, 645-654.

Lee, Y.J., Hoynes-O'Connor, A., Leong, M.C., and Moon, T.S. (2016b). Programmable control of bacterial gene expression with the combined CRISPR and antisense RNA system. *Nucleic Acids Research* 44, 2462-2473.

Levy, A., Goren, M.G., Yosef, I., Auster, O., Manor, M., Amitai, G., Edgar, R., Qimron, U., and Sorek, R. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 520, 505-510.

Li, B., Yan, J., Zhang, Y., Li, W., Zeng, C., Zhao, W., Hou, X., Zhang, C., and Dong, Y. (2020). CRISPR-Cas12a Possesses Unconventional DNase Activity that Can Be Inactivated by Synthetic Oligonucleotides. *Mol Ther Nucleic Acids* 19, 1043-1052.

Li, P., Roberts, B.P., Chakravorty, D.K., and Merz, K.M., Jr. (2013). Rational Design of Particle Mesh Ewald Compatible Lennard-Jones Parameters for +2 Metal Cations in Explicit Solvent. *J Chem Theory Comput* 9, 2733-2748.

Li, R., Fang, L., Tan, S., Yu, M., Li, X., He, S., Wei, Y., Li, G., Jiang, J., and Wu, M. (2016). Type I CRISPR-Cas targets endogenous genes and regulates virulence to evade mammalian host immunity. *Cell Research* 26, 1273-1287.

Li, S.Y., Cheng, Q.X., Liu, J.K., Nie, X.Q., Zhao, G.P., and Wang, J. (2018a). CRISPR-Cas12a has both cis- and trans-cleavage activities on single-stranded DNA. *Cell Res* 28, 491-493.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22, 1658-1659.

Li, W., Moore, M.J., Vasilieva, N., Sui, J., Wong, S.K., Berne, M.A., Somasundaran, M., Sullivan, J.L., Luzuriaga, K., Greenough, T.C., *et al.* (2003). Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. *Nature* 426, 450-454.

Li, X., Wang, Y., Liu, Y., Yang, B., Wang, X., Wei, J., Lu, Z., Zhang, Y., Wu, J., Huang, X., *et al.* (2018b). Base editing with a Cpf1–cytidine deaminase fusion. *Nature Biotechnology* 36, 324-327.

Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., Chen, L., Liang, Y.X., and She, Q. (2015). Harnessing Type I and Type III CRISPR-Cas systems for genome editing. *Nucleic Acids Research* 44, e34-e34.

Lim, H., Baek, A., Kim, J., Kim, M.S., Liu, J., Nam, K.-Y., Yoon, J., and No, K.T. (2020). Hot spot profiles of SARS-CoV-2 and human ACE2 receptor protein interaction obtained by density functional tight binding fragment molecular orbital method. *Sci Rep-Uk* 10, 16862.

Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., and Shaw, D.E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 78, 1950-1958.

Liu, K., Tan, S., Niu, S., Wang, J., Wu, L., Sun, H., Zhang, Y., Pan, X., Qu, X., Du, P., *et al.* (2021). Cross-species recognition of SARS-CoV-2 to bat ACE2. *Proceedings of the National Academy of Sciences* 118, e2020216118.

Liu, K.I., Ramli, M.N.B., Woo, C.W.A., Wang, Y., Zhao, T., Zhang, X., Yim, G.R.D., Chong, B.Y., Gowher, A., Chua, M.Z.H., *et al.* (2016). A chemical-inducible CRISPR–Cas9 system for rapid control of genome editing. *Nature Chemical Biology* 12, 980-987.

Liu, T.Y., and Doudna, J.A. (2020). Chemistry of Class 1 CRISPR-Cas effectors: binding, editing, and regulation. *Journal of Biological Chemistry*.

Liu, Z., Dong, H., Cui, Y., Cong, L., and Zhang, D. (2020). Application of different types of CRISPR/Cas-based systems in bacteria. *Microbial Cell Factories* 19, 172.

Louwen, R., Horst-Kreft, D., de Boer, A.G., van der Graaf, L., de Knecht, G., Hamersma, M., Heikema, A.P., Timms, A.R., Jacobs, B.C., Wagenaar, J.A., *et al.* (2013). A novel link between *Campylobacter jejuni* bacteriophage defence, virulence and Guillain-Barre syndrome. *Eur J Clin Microbiol Infect Dis* 32, 207-226.

Lu, G., Hu, Y., Wang, Q., Qi, J., Gao, F., Li, Y., Zhang, Y., Zhang, W., Yuan, Y., Bao, J., *et al.* (2013). Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. *Nature* 500, 227-231.

Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N., *et al.* (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *The Lancet* 395, 565-574.

Luan, B., and Huynh, T. (2020). In Silico Antibody Mutagenesis for Optimizing Its Binding to Spike Protein of Severe Acute Respiratory Syndrome Coronavirus 2. *The Journal of Physical Chemistry Letters* 11, 9781-9787.

Ma, K., Cao, Q., Luo, S., Wang, Z., Liu, G., Lu, C., and Liu, Y. (2018). cas9 Enhances Bacterial Virulence by Repressing the regR Transcriptional Regulator in *Streptococcus agalactiae*. *Infection and Immunity* 86, e00552-00517.

Maier, J.A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K.E., and Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* 11, 3696-3713.

Maji, B., Moore, C.L., Zetsche, B., Volz, S.E., Zhang, F., Shoulders, M.D., and Choudhary, A. (2017). Multidimensional chemical control of CRISPR–Cas9. *Nature Chemical Biology* 13, 9-11.

Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B., and Koonin, E.V. (2002). A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic Acids Research* 30, 482-496.

Makarova, K.S., Grishin, N.V., Shabalina, S.A., Wolf, Y.I., and Koonin, E.V. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol Direct* 1, 7.

Makarova, K.S., and Koonin, E.V. (2015). Annotation and Classification of CRISPR-Cas Systems. *Methods Mol Biol* 1311, 47-75.

Makarova, K.S., Wolf, Y.I., Iranzo, J., Shmakov, S.A., Alkhnbashi, O.S., Brouns, S.J.J., Charpentier, E., Cheng, D., Haft, D.H., Horvath, P., *et al.* (2020). Evolutionary classification of CRISPR-Cas systems: a burst of class 2 and derived variants. *Nat Rev Microbiol* 18, 67-83.

Marino, N.D., Pinilla-Redondo, R., Csörgő, B., and Bondy-Denomy, J. (2020). Anti-CRISPR protein applications: natural brakes for CRISPR-Cas technologies. *Nature Methods* 17, 471-479.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science* 322, 1843-1845.

Mayo-Muñoz, D., He, F., Jørgensen, J.B., Madsen, P.K., Bhoobalan-Chitty, Y., and Peng, X. (2018). Anti-CRISPR-Based and CRISPR-Based Genome Editing of *Sulfolobus islandicus* Rod-Shaped Virus 2. *Viruses* 10, 695.

Meirson, T., Bomze, D., and Markel, G. (2021). Structural basis of SARS-CoV-2 spike protein induced by ACE2. *Bioinformatics* 37, 929-936.

Merkle, F.T., Ghosh, S., Kamitaki, N., Mitchell, J., Avior, Y., Mello, C., Kashin, S., Mekhoubad, S., Ilic, D., Charlton, M., *et al.* (2017). Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations. *Nature* 545, 229-233.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., and Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* 37, 1530-1534.

Mody, V., Ho, J., Wills, S., Mawri, A., Lawson, L., Ebert, M.C.C.J.C., Fortin, G.M., Rayalam, S., and Taval, S. (2021). Identification of 3-chymotrypsin like protease (3CLPro) inhibitors as potential anti-SARS-CoV-2 agents. *Communications Biology* 4, 93.

Mogila, I., Kazlauskienė, M., Valinskyte, S., Tamulaitiene, G., Tamulaitis, G., and Siksnys, V. (2019). Genetic Dissection of the Type III-A CRISPR-Cas System Csm Complex Reveals Roles of Individual Subunits. *Cell Reports* 26, 2753-2765.e2754.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Almendros, C. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology* 155, 733-740.

Mojica, F.J., Diez-Villasenor, C., Garcia-Martinez, J., and Soria, E. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol* 60, 174-182.

Mojica, F.J.M., Juez, G., and Rodriguez-Valera, F. (1993). Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified PstI sites. *Molecular Microbiology* 9, 613-621.

Moore, R.M., Harrison, A.O., McAllister, S.M., Polson, S.W., and Wommack, K.E. (2020). Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ* 8, e8584.

Mosterd, C., Rousseau, G.M., and Moineau, S. (2020). A short overview of the CRISPR-Cas adaptation stage. *Canadian Journal of Microbiology* 67, 1-12.

Müller, M., Lee, C.M., Gasiunas, G., Davis, T.H., Cradick, T.J., Siksnys, V., Bao, G., Cathomen, T., and Mussolino, C. (2016). *Streptococcus thermophilus* CRISPR-Cas9 Systems Enable Specific Editing of the Human Genome. *Molecular Therapy* 24, 636-644.

Murugan, K., Babu, K., Sundaresan, R., Rajan, R., and Sashital, D.G. (2017). The Revolution Continues: Newly Discovered Systems Expand the CRISPR-Cas Toolkit. *Mol Cell* 68, 15-25.

Mycroft-West, C.J., Su, D., Pagani, I., Rudd, T.R., Elli, S., Gandhi, N.S., Guimond, S.E., Miller, G.J., Meneghetti, M.C.Z., Nader, H.B., *et al.* (2020). Heparin Inhibits Cellular Invasion by SARS-CoV-2: Structural Dependence of the Interaction of the Spike S1 Receptor-Binding Domain with Heparin. *Thromb Haemost* 120, 1700-1715.

Nayeemul Bari, S.M., and Hatoum-Aslan, A. (2019). Chapter Seventeen - CRISPR–Cas10 assisted editing of virulent staphylococcal phages. In *Methods in Enzymology*, S. Bailey, ed. (Academic Press), pp. 385-409.

Nelles, David A., Fang, Mark Y., O'Connell, Mitchell R., Xu, Jia L., Markmiller, Sebastian J., Doudna, Jennifer A., and Yeo, Gene W. (2016). Programmable RNA Tracking in Live Cells with CRISPR/Cas9. *Cell* 165, 488-496.

Newsom, S., Parameshwaran, H.P., Martin, L., and Rajan, R. (2020). The CRISPR-Cas Mechanism for Adaptive Immunity and Alternate Bacterial Functions Fuels Diverse Biotechnologies. *Front Cell Infect Microbiol* 10, 619763.

Newsom, S., Parameshwaran, H.P., Martin, L., and Rajan, R. (2021). The CRISPR-Cas Mechanism for Adaptive Immunity and Alternate Bacterial Functions Fuels Diverse Biotechnologies. *Frontiers in Cellular and Infection Microbiology* 10, 898.

Nierzwicki, L., Arantes, P.R., Saha, A., and Palermo, G. (2021a). Establishing the allosteric mechanism in CRISPR-Cas9. *Wiley Interdiscip Rev Comput Mol Sci* 11.

Nierzwicki, Ł., Arantes, P.R., Saha, A., and Palermo, G. (2021b). Establishing the allosteric mechanism in CRISPR-Cas9. *WIREs Computational Molecular Science* 11, e1503.

Niewoehner, O., Garcia-Doval, C., Rostøl, J.T., Berk, C., Schwede, F., Bigler, L., Hall, J., Marraffini, L.A., and Jinek, M. (2017). Type III CRISPR–Cas systems produce cyclic oligoadenylate second messengers. *Nature* 548, 543-548.

Nihongaki, Y., Kawano, F., Nakajima, T., and Sato, M. (2015). Photoactivatable CRISPR-Cas9 for optogenetic genome editing. *Nature Biotechnology* 33, 755-760.

Nishimasu, H., Ran, F.A., Hsu, P.D., Konermann, S., Shehata, S.I., Dohmae, N., Ishitani, R., Zhang, F., and Nureki, O. (2014). Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* 156, 935-949.

Nowotny, M., Gaidamakov, S.A., Crouch, R.J., and Yang, W. (2005). Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. *Cell* 121, 1005-1016.

Nunez, J.K., Kranzusch, P.J., Noeske, J., Wright, A.V., Davies, C.W., and Doudna, J.A. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. *Nat Struct Mol Biol* 21, 528-534.

O'Connell, M.R. (2019). Molecular Mechanisms of RNA Targeting by Cas13-containing Type VI CRISPR–Cas Systems. *Journal of Molecular Biology* 431, 66-87.

Oh, Y., Lee, W.-j., Hur, J.K., Song, W.J., Lee, Y., Kim, H., Gwon, L.W., Kim, Y.-H., Park, Y.-H., Kim, C.H., *et al.* (2022). Expansion of the prime editing modality with Cas9 from *Francisella novicida*. *Genome Biology* 23, 92.

Özcan, A., Pausch, P., Linden, A., Wulf, A., Schühle, K., Heider, J., Urlaub, H., Heimerl, T., Bange, G., and Randau, L. (2019). Type IV CRISPR RNA processing and effector complex formation in *Aromatoleum aromaticum*. *Nature Microbiology* 4, 89-96.

Ozono, S., Zhang, Y., Ode, H., Sano, K., Tan, T.S., Imai, K., Miyoshi, K., Kishigami, S., Ueno, T., Iwatani, Y., *et al.* (2021). SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat Commun* 12, 848.

Palermo, G. (2019). Structure and Dynamics of the CRISPR-Cas9 Catalytic Complex. *J Chem Inf Model* 59, 2394-2406.

Palermo, G., Chen, J.S., Ricci, C.G., Rivalta, I., Jinek, M., Batista, V.S., Doudna, J.A., and McCammon, J.A. (2018a). Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Q Rev Biophys* 51, e91.

Palermo, G., Chen, J.S., Ricci, C.G., Rivalta, I., Jinek, M., Batista, V.S., Doudna, J.A., and McCammon, J.A. (2018b). Key role of the REC lobe during CRISPR-Cas9 activation by 'sensing', 'regulating', and 'locking' the catalytic HNH domain. *Q Rev Biophys* 51.

Palermo, G., Miao, Y., Walker, R.C., Jinek, M., and McCammon, J.A. (2016). Striking Plasticity of CRISPR-Cas9 and Key Role of Non-target DNA, as Revealed by Molecular Simulations. *ACS Cent Sci* 2, 756-763.

Palermo, G., Miao, Y., Walker, R.C., Jinek, M., and McCammon, J.A. (2017a). CRISPR-Cas9 conformational activation as elucidated from enhanced molecular simulations. *Proc Natl Acad Sci U S A* 114, 7260-7265.

Palermo, G., Ricci, C.G., Fernando, A., Basak, R., Jinek, M., Rivalta, I., Batista, V.S., and McCammon, J.A. (2017b). Protospacer Adjacent Motif-Induced Allostery Activates CRISPR-Cas9. *J Am Chem Soc* 139, 16028-16031.

Parameshwaran, H.P., Babu, K., Tran, C., Guan, K., Allen, A., Kathiresan, V., Qin, P.Z., and Rajan, R. (2021). The bridge helix of Cas12a imparts selectivity in cis-DNA cleavage and regulates trans-DNA cleavage. *FEBS Lett* 595, 892-912.

Paul, B., and Montoya, G. (2020). CRISPR-Cas12a: Functional overview and applications. *Biomed J* 43, 8-17.

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C.A., Li, Z., Cress, B.F., Knott, G.J., Jacobsen, S.E., Banfield, J.F., and Doudna, J.A. (2020). CRISPR-CasΦ from huge phages is a hypercompact genome editor. *Science* 369, 333.

Perez-Pinera, P., Kocak, D.D., Vockley, C.M., Adler, A.F., Kabadi, A.M., Polstein, L.R., Thakore, P.I., Glass, K.A., Ousterout, D.G., Leong, K.W., *et al.* (2013). RNA-guided gene activation by CRISPR-Cas9-based transcription factors. *Nature methods* 10, 973-976.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* 25, 1605-1612.

Pickar-Oliver, A., and Gersbach, C.A. (2019). The next generation of CRISPR-Cas technologies and applications. *Nature Reviews Molecular Cell Biology* 20, 490-507.

Pinilla-Redondo, R., Mayo-Muñoz, D., Russel, J., Garrett, R.A., Randau, L., Sørensen, S.J., and Shah, S.A. (2020). Type IV CRISPR–Cas systems are highly diverse and involved in competition between plasmids. *Nucleic Acids Research* 48, 2000-2012.

Plagens, A., Richter, H., Charpentier, E., and Randau, L. (2015). DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. *FEMS Microbiology Reviews* 39, 442-463.

Polstein, L.R., and Gersbach, C.A. (2015). A light-inducible CRISPR-Cas9 system for control of endogenous gene activation. *Nature Chemical Biology* 11, 198-200.

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., and Lim, W.A. (2013). Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* 152, 1173-1183.

Qiu, Y., Zhao, Y.-B., Wang, Q., Li, J.-Y., Zhou, Z.-J., Liao, C.-H., and Ge, X.-Y. (2020). Predicting the angiotensin converting enzyme 2 (ACE2) utilizing capability as the receptor of SARS-CoV-2. *Microbes and Infection* 22, 221-225.

Rahdar, M., McMahon, M.A., Prakash, T.P., Swayze, E.E., Bennett, C.F., and Cleveland, D.W. (2015). Synthetic CRISPR RNA-Cas9-guided genome editing in human cells. *Proc Natl Acad Sci U S A* 112, E7110-7117.

Raj, V.S., Mou, H., Smits, S.L., Dekkers, D.H.W., Müller, M.A., Dijkman, R., Muth, D., Demmers, J.A.A., Zaki, A., Fouchier, R.A.M., *et al.* (2013). Dipeptidyl peptidase 4 is a functional receptor for the emerging human coronavirus-EMC. *Nature* 495, 251-254.

Ran, F.A., Cong, L., Yan, W.X., Scott, D.A., Gootenberg, J.S., Kriz, A.J., Zetsche, B., Shalem, O., Wu, X., Makarova, K.S., *et al.* (2015). In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature* 520, 186-191.

Raper, A.T., Stephenson, A.A., and Suo, Z. (2018). Functional Insights Revealed by the Kinetic Mechanism of CRISPR/Cas9. *J Am Chem Soc* 140, 2971-2984.

Rath, D., Amlinger, L., Hoekzema, M., Devulapally, P.R., and Lundgren, M. (2014). Efficient programmable gene silencing by Cascade. *Nucleic Acids Research* 43, 237-246.

Ratner, H.K., Escalera-Maurer, A., Le Rhun, A., Jaggavarapu, S., Wozniak, J.E., Crispell, E.K., Charpentier, E., and Weiss, D.S. (2019). Catalytically Active Cas9 Mediates Transcriptional Interference to Facilitate Bacterial Virulence. *Molecular cell* 75, 498-510.e495.

Rezaei, S., Sefidbakht, Y., and Uskoković, V. (2022). Comparative molecular dynamics study of the receptor-binding domains in SARS-CoV-2 and SARS-CoV and the effects of mutations on the binding affinity. *Journal of Biomolecular Structure and Dynamics* 40, 4662-4681.

Ricci, C.G., Chen, J.S., Miao, Y., Jinek, M., Doudna, J.A., McCammon, J.A., and Palermo, G. (2019). Deciphering Off-Target Effects in CRISPR-Cas9 through Accelerated Molecular Dynamics. *ACS Cent Sci* 5, 651-662.

Roe, D.R., and Cheatham, T.E., 3rd (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* 9, 3084-3095.

Roe, M.K., Junod, N.A., Young, A.R., Beachboard, D.C., and Stobart, C.C. (2021). Targeting novel structural and functional features of coronavirus protease nsp5 (3CLpro, Mpro) in the age of COVID-19. *Journal of General Virology* 102.

Romiti, M. (2006). Entrez Nucleotide and Entrez Protein FAQs. In Entrez Sequences Help Manual [Internet] (Bethesda (MD): National Center for Biotechnology Information (US)).

Rousseau, B.A., Hou, Z., Gramelspacher, M.J., and Zhang, Y. (2018). Programmable RNA Cleavage and Recognition by a Natural CRISPR-Cas9 System from *Neisseria meningitidis*. *Mol Cell* 69, 906-914 e904.

Russel, J., Pinilla-Redondo, R., Mayo-Muñoz, D., Shah, S.A., and Sørensen, S.J. (2020). CRISPRCasTyper: Automated Identification, Annotation, and Classification of CRISPR-Cas Loci. *Crispr j* 3, 462-469.

Saha, C., Horst-Kreft, D., Kross, I., van der Spek, P.J., Louwen, R., and van Baarlen, P. (2020a). *Campylobacter jejuni* Cas9 Modulates the Transcriptome in Caco-2 Intestinal Epithelial Cells. *11*, 1193.

Saha, C., Horst-Kreft, D., Kross, I., van der Spek, P.J., Louwen, R., and van Baarlen, P. (2020b). *Campylobacter jejuni* Cas9 Modulates the Transcriptome in Caco-2 Intestinal Epithelial Cells. *Genes (Basel)* 11.

Saha, C., Mohanraju, P., Stubbs, A., Dugar, G., Hoogstrate, Y., Kremers, G.J., van Cappellen, W.A., Horst-Kreft, D., Laffeber, C., Lebbink, J.H.G., *et al.* (2020c). Guide-free Cas9 from pathogenic *Campylobacter jejuni* bacteria causes severe damage to DNA. *Sci Adv* 6, eaaz4849.

Salomon-Ferrer R, C.D., Walker RC. (2013). An overview of the amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science* 3, 198-210.

Sampson, T.R., Saroj, S.D., Llewellyn, A.C., Tzeng, Y.-L., and Weiss, D.S. (2013). A CRISPR/Cas system mediates bacterial innate immune evasion and virulence. *Nature (London, U K)* 497, 254-257.

Sampson, T.R., and Weiss, D.S. (2013). Degeneration of a CRISPR/Cas system and its regulatory target during the evolution of a pathogen. *RNA Biol* 10, 1618-1622.

Sansbury, B.M., Hewes, A.M., and Kmiec, E.B. (2019). Understanding the diversity of genetic outcomes from CRISPR-Cas generated homology-directed repair. *Communications Biology* 2, 458.

Sarkar, P., and Sumbly, P. (2017). Regulatory gene mutation: a driving force behind group a *Streptococcus* strain- and serotype-specific variation. *Molecular Microbiology* 103, 576-589.

Schmidt, F., Cherepkova, M.Y., and Platt, R.J. (2018). Transcriptional recording by CRISPR spacer acquisition from RNA. *Nature* 562, 380-385.

Schneider, C.A., Rasband, W.S., and Eliceiri, K.W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9, 671-675.

Seyran, M., Takayama, K., Uversky, V.N., Lundstrom, K., Palù, G., Sherchan, S.P., Attrish, D., Rezaei, N., Aljabali, A.A.A., Ghosh, S., *et al.* (2021). The structural basis of accelerated host cell entry by SARS-CoV-2†. *The FEBS journal* 288, 5010-5020.

Shabbir, M.A.B., Tang, Y., Xu, Z., Lin, M., Cheng, G., Dai, M., Wang, X., Liu, Z., Yuan, Z., and Hao, H. (2018). The Involvement of the Cas9 Gene in Virulence of *Campylobacter jejuni*. *Frontiers in cellular and infection microbiology* 8, 285-285.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., and Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences of the United States of America* 117, 11727-11734.



Shipman, S.L., Nivala, J., Macklis, J.D., and Church, G.M. (2017). CRISPR-Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature* (London, U K) *547*, 345-349.

Shmakov, S.A., Sitnik, V., Makarova, K.S., Wolf, Y.I., Severinov, K.V., and Koonin, E.V. (2017). The CRISPR Spacer Space Is Dominated by Sequences from Species-Specific Mobilomes. *mBio* *8*, e01397-01317.

Sievers, F., and Higgins, D.G. (2014). Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* *1079*, 105-116.

Slaymaker, I.M., Gao, L., Zetsche, B., Scott, D.A., Yan, W.X., and Zhang, F. (2016). Rationally engineered Cas9 nucleases with improved specificity. *Science* *351*, 84-88.

Smith, E.E., Buckley, D.G., Wu, Z., Saenphimmachak, C., Hoffman, L.R., D'Argenio, D.A., Miller, S.I., Ramsey, B.W., Speert, D.P., Moskowitz, S.M., *et al.* (2006). Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proceedings of the National Academy of Sciences* *103*, 8487.

Spencer, B.L., Deng, L., Patras, K.A., Burcham, Z.M., Sanches, G.F., Nagao, P.E., and Doran, K.S. (2019). Cas9 Contributes to Group B Streptococcal Colonization and Disease. *Frontiers in Microbiology* *10*, 1930.

Starr, T., Greaney, A., Hilton, S., King, N., Veessler, D., and Bloom, J. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell* *182*, 1295-1310.

Stephenson, A.A., Raper, A.T., and Suo, Z. (2018). Bidirectional Degradation of DNA Cleavage Products Catalyzed by CRISPR/Cas9. *J Am Chem Soc* *140*, 3743-3750.

Sternberg, S.H., LaFrance, B., Kaplan, M., and Doudna, J.A. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* *527*, 110-113.

Sternberg, S.H., Redding, S., Jinek, M., Greene, E.C., and Doudna, J.A. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* *507*, 62-67.

Strutt, S.C., Torrez, R.M., Kaya, E., Negrete, O.A., and Doudna, J.A. (2018). RNA-dependent RNA targeting by CRISPR-Cas9. *Elife* *7*.

Sundaresan, R., Parameshwaran, H.P., Yogesha, S.D., Keilbarth, M.W., and Rajan, R. (2017). RNA-Independent DNA Cleavage Activities of Cas9 and Cas12a. *Cell Rep* *21*, 3728-3739.

Szczelkun, M.D., Tikhomirova, M.S., Sinkunas, T., Gasiunas, G., Karvelis, T., Pschera, P., Siksnys, V., and Seidel, R. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc Natl Acad Sci U S A* *111*, 9798-9803.

Tadić, V., Josipović, G., Zoldoš, V., and Vojta, A. (2019). CRISPR/Cas9-based epigenome editing: An overview of dCas9-based tools with special emphasis on off-target activity. *Methods* *164-165*, 109-119.

Tamulaitis, G., Venclovas, Č., and Siksnys, V. (2017). Type III CRISPR-Cas Immunity: Major Differences Brushed Aside. *Trends in Microbiology* *25*, 49-61.

Tang, B., Gong, T., Zhou, X., Lu, M., Zeng, J., Peng, X., Wang, S., and Li, Y. (2019). Deletion of *cas3* gene in *Streptococcus mutans* affects biofilm formation and increases fluoride sensitivity. *Archives of Oral Biology* *99*, 190-197.

Tang, H., Yuan, H., Du, W., Li, G., Xue, D., and Huang, Q. (2021). Active-Site Models of *Streptococcus pyogenes* Cas9 in DNA Cleavage State. *Frontiers in Molecular Biosciences* *8*, 235.

Tang, W., Hu, J.H., and Liu, D.R. (2017). Aptazyme-embedded guide RNAs enable ligand-responsive genome editing and transcriptional activation. *Nat Commun* 8, 15939.

Taylor, D.W., Zhu, Y., Staals, R.H.J., Kornfeld, J.E., Shinkai, A., van der Oost, J., Nogales, E., and Doudna, J.A. (2015). Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. *Science* 348, 581.

Taylor, H.N., Laderman, E., Armbrust, M., Hallmark, T., Keiser, D., Bondy-Denomy, J., and Jackson, R.N. (2021). Positioning Diverse Type IV Structures and Functions Within Class 1 CRISPR-Cas Systems. *Frontiers in Microbiology* 12.

Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., Doolabh, D., Pillay, S., San, E.J., Msomi, N., *et al.* (2021). Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* 592, 438-443.

Teng, S., Sobitan, A., Rhoades, R., Liu, D., and Tang, Q. (2021). Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. *Briefings in Bioinformatics* 22, 1239-1253.

Toor, H.G., Banerjee, D.I., Lipsa Rath, S., and Darji, S.A. (2021). Computational drug re-purposing targeting the spike glycoprotein of SARS-CoV-2 as an effective strategy to neutralize COVID-19. *Eur J Pharmacol* 890, 173720.

Trott, O., and Olson, A.J. (2010). AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem* 31, 455-461.

Van Orden, M.J., Klein, P., Babu, K., Najar, F.Z., and Rajan, R. (2017). Conserved DNA motifs in the type II-A CRISPR leader region. *PeerJ* 5, e3161.

Van Orden, M.J., Newsom, S., and Rajan, R. (2020). CRISPR type II-A subgroups exhibit phylogenetically distinct mechanisms for prespacer insertion. *J Biol Chem* 295, 10956-10968.

Vazquez Reyes, C., Tangprasertchai, N.S., Yogesha, S.D., Nguyen, R.H., Zhang, X., Rajan, R., and Qin, P.Z. (2017). Nucleic Acid-Dependent Conformational Changes in CRISPR-Cas9 Revealed by Site-Directed Spin Labeling. *Cell Biochem Biophys* 75, 203-210.

Vega, L.A., Malke, H., and McIver, K.S. (2016). Virulence-Related Transcriptional Regulators of *Streptococcus pyogenes* (University of Oklahoma Health Sciences Center, Oklahoma City (OK)).

Vercoe, R.B., Chang, J.T., Dy, R.L., Taylor, C., Gristwood, T., Clulow, J.S., Richter, C., Przybilski, R., Pitman, A.R., and Fineran, P.C. (2013). Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet* 9, e1003454-e1003454.

Verma, J., and Subbarao, N. (2021). A comparative study of human betacoronavirus spike proteins: structure, function and therapeutics. *Archives of Virology* 166, 697-714.

Viswanathan, P., Murphy, K., Julien, B., Garza, A.G., and Kroos, L. (2007). Regulation of dev, an Operon That Includes Genes Essential for *Myxococcus xanthus* Development and CRISPR-Associated Genes and Repeats. *Journal of Bacteriology* 189, 3738.

Waldrip, Z.J., Byrum, S.D., Storey, A.J., Gao, J., Byrd, A.K., Mackintosh, S.G., Wahls, W.P., Taverna, S.D., Raney, K.D., and Tackett, A.J. (2014). A CRISPR-based approach for proteomic analysis of a single genomic locus. *Epigenetics* 9, 1207-1211.

Wallace, R.A., Black, W.P., Yang, X., and Yang, Z. (2014). A CRISPR with Roles in *Myxococcus xanthus* Development and Exopolysaccharide Production. *Journal of Bacteriology* 196, 4036-4043.

Walls, A.C., Park, Y.-J., Tortorici, M.A., Wall, A., McGuire, A.T., and Velesler, D. (2020). Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 181, 281-292.e286.

Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A., and Case, D.A. (2004). Development and testing of a general amber force field. *Journal of Computational Chemistry* 25, 1157-1174.

Wang, L., and Xiang, Y. (2020). Spike Glycoprotein-Mediated Entry of SARS Coronaviruses. In *Viruses*.

Wang, M.-Y., Zhao, R., Gao, L.-J., Gao, X.-F., Wang, D.-P., and Cao, J.-M. (2020a). SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Frontiers in Cellular and Infection Microbiology* 10.

Wang, Y., Liu, M., and Gao, J. (2020b). Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. *Proceedings of the National Academy of Sciences* 117, 13967-13974.

Wang, Y., Wang, M., Zheng, T., Hou, Y., Zhang, P., Tang, T., Wei, J., and Du, Q. (2020c). Specificity profiling of CRISPR system reveals greatly enhanced off-target gene editing. *Sci Rep-Uk* 10, 2269.

Wang, Y., Yin, X., Zhou, Z., Hu, S., Li, S., Liu, M., Wang, X., Xiao, Y., Shi, D., Bi, D., *et al.* (2019). Cas9 regulated gene expression and pathogenicity in *Riemerella anatipestifer*. *Microbial Pathogenesis* 136, 103706.

Warwicker, J. (2021). A model for pH coupling of the SARS-CoV-2 spike protein open/closed equilibrium. *Brief Bioinform* 22, 1499-1507.

Watson, B.N.J., Staals, R.H.J., and Fineran, P.C. (2018). CRISPR-Cas-Mediated Phage Resistance Enhances Horizontal Gene Transfer by Transduction. *mBio* 9, e02406-02417.

Wei, J., Hou, L., Liu, J., Wang, Z., Gao, S., Qi, T., Gao, S., Sun, S., and Wang, Y. (2022). Closely related type II-C Cas9 orthologs recognize diverse PAMs. *eLife* 11, e77825.

White, J.M., Delos, S.E., Brecher, M., and Schornberg, K. (2008). Structures and Mechanisms of Viral Membrane Fusion Proteins: Multiple Variations on a Common Theme. *Critical Reviews in Biochemistry and Molecular Biology* 43, 189-219.

Wilson, T.E., Topper, L.M., and Palmbo, P.L. (2003). Non-homologous end-joining: bacteria join the chromosome breakdance. *Trends in Biochemical Sciences* 28, 62-66.

Wimmer, F., and Beisel, C.L. (2020). CRISPR-Cas Systems and the Paradox of Self-Targeting Spacers. *Frontiers in Microbiology* 10, 3078.

Wintjens, R., Bifani, A.M., and Bifani, P. (2020). Impact of glycan cloud on the B-cell epitope prediction of SARS-CoV-2 Spike protein. *npj Vaccines* 5, 81.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.-L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 367, 1260-1263.

Wright, A.V., Liu, J.-J., Knott, G.J., Doxzen, K.W., Nogales, E., and Doudna, J.A. (2017). Structures of the CRISPR genome integration complex. *Science*, eaao0679.

Xu, C., Wang, Y., Liu, C., Zhang, C., Han, W., Hong, X., Wang, Y., Hong, Q., Wang, S., Zhao, Q., *et al.* Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM. *Science Advances* 7, eabe5575.

Xu, K., Ren, C., Liu, Z., Zhang, T., Zhang, T., Li, D., Wang, L., Yan, Q., Guo, L., Shen, J., *et al.* (2015). Efficient genome engineering in eukaryotes using Cas9 from *Streptococcus thermophilus*. *Cellular and Molecular Life Sciences* 72, 383-399.

Xu, R., Qin, R., Xie, H., Li, J., Liu, X., Zhu, M., Sun, Y., Yu, Y., Lu, P., and Wei, P. (2022). Genome editing with type II-C CRISPR-Cas9 systems from *Neisseria meningitidis* in rice. *Plant biotechnology journal* 20, 350-359.

Xu, X., and Qi, L.S. (2019). A CRISPR–dCas Toolbox for Genetic Engineering and Synthetic Biology. *Journal of Molecular Biology* 431, 34-47.

Yan, M.-Y., Yan, H.-Q., Ren, G.-X., Zhao, J.-P., Guo, X.-P., and Sun, Y.-C. (2017). CRISPR-Cas12a-Assisted Recombineering in Bacteria. *Appl Environ Microbiol* 83, e00947-00917.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* 367, 1444-1448.

Yang, S., Li, S., and Li, X.J. (2018). Shortening the Half-Life of Cas9 Maintains Its Gene Editing Ability and Reduces Neuronal Toxicity. *Cell Rep* 25, 2653-2659 e2653.

You, L., Ma, J., Wang, J., Artamonova, D., Wang, M., Liu, L., Xiang, H., Severinov, K., Zhang, X., and Wang, Y. (2019). Structure Studies of the CRISPR-Csm Complex Reveal Mechanism of Co-transcriptional Interference. *Cell* 176, 239-253.e216.

Young, J.K., Gasior, S.L., Jones, S., Wang, L., Navarro, P., Vickroy, B., and Barrangou, R. (2019). The repurposing of type I-E CRISPR-Cascade for gene activation in plants. *Communications Biology* 2, 383.

Yu, Y., Wu, X., Guan, N., Shao, J., Li, H., Chen, Y., Ping, Y., Li, D., and Ye, H. (2020). Engineering a far-red light–activated split-Cas9 system for remote-controlled genome editing of internal organs and tumors. *Science Advances* 6, eabb1777.

Zebec, Z., Manica, A., Zhang, J., White, M.F., and Schleper, C. (2014). CRISPR-mediated targeted mRNA degradation in the archaeon *Sulfolobus solfataricus*. *Nucleic Acids Research* 42, 5280-5288.

Zetsche, B., Gootenberg, J.S., Abudayyeh, O.O., Slaymaker, I.M., Makarova, K.S., Essletzbichler, P., Volz, S.E., Joung, J., van der Oost, J., Regev, A., *et al.* (2015a). Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* 163, 759-771.

Zetsche, B., Volz, S.E., and Zhang, F. (2015b). A split-Cas9 architecture for inducible genome editing and transcription modulation. *Nature Biotechnology* 33, 139-142.

Zhang, J., Cai, Y., Lavine, C.L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M.L., Rits-Volloch, S., *et al.* (2022). Structural and functional impact by SARS-CoV-2 Omicron spike mutations. *Cell Reports* 39, 110729.

Zhang, J., Xiao, T., Cai, Y., Lavine, C.L., Peng, H., Zhu, H., Anand, K., Tong, P., Gautam, A., Mayer, M.L., *et al.* (2021). Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. *Science* 374, 1353-1360.

Zhang, X., Wang, J., Cheng, Q., Zheng, X., Zhao, G., and Wang, J. (2017). Multiplex gene regulation by CRISPR-ddCpf1. *Cell Discovery* 3, 17018.

Zhang, Y., Heidrich, N., Ampattu, B.J., Gunderson, C.W., Seifert, H.S., Schoen, C., Vogel, J., and Sontheimer, E.J. (2013). Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Mol Cell* 50, 488-503.

Zhao, X., Chen, D., Szabla, R., Zheng, M., Li, G., Du, P., Zheng, S., Li, X., Song, C., Li, R., *et al.* (2020). Broad and Differential Animal Angiotensin-Converting Enzyme 2 Receptor Usage by SARS-CoV-2. *Journal of Virology* 94, e00940-00920.

Zheng, Y., Li, J., Wang, B., Han, J., Hao, Y., Wang, S., Ma, X., Yang, S., Ma, L., Yi, L., *et al.* (2020). Endogenous Type I CRISPR-Cas: From Foreign DNA Defense to Prokaryotic Engineering. *Front Bioeng Biotechnol* 8, 62.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., *et al.* (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* 579, 270-273.

Zhou, Z., Qiu, Y., and Ge, X. (2021). The taxonomy, host range and pathogenicity of coronaviruses and other viruses in the Nidovirales order. *Animal Diseases* 1, 5.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., *et al.* (2020). A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* 382, 727-733.

Zuccaro, M.V., Xu, J., Mitchell, C., Marin, D., Zimmerman, R., Rana, B., Weinstein, E., King, R.T., Palmerola, K.L., Smith, M.E., *et al.* (2020). Allele-Specific Chromosome Removal after Cas9 Cleavage in Human Embryos. *Cell* 183, 1650-1664 e1615.

Zuo, Z., and Liu, J. (2016). Cas9-catalyzed DNA Cleavage Generates Staggered Ends: Evidence from Molecular Dynamics Simulations. *Sci Rep* 5, 37584.

Zuo, Z., and Liu, J. (2020). Allosteric regulation of CRISPR-Cas9 for DNA-targeting and cleavage. *Curr Opin Struct Biol* 62, 166-174.

| >II-A_WP_010922251.1_Streptococcus_pyogenes_SpyCas9                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| MDKKYSIGLDIGTNSVGWAVITDEYKVPSKKFKVLGNTDRHSIKKNLIGALLFDSGETA<br>EATRLKRTARRRYTRRKNRICYLQEIFSNEMAKVDDSSFFHRLEESFLVEEDKKHERHP<br>IFGNIVDEVAYHEKYPTIYHLRKKLVDSTDKADLRILIYLAHAMIKFRGHFLIEGDLNPD<br>NSDVDKLFQILVQTYNQLFEENPINASGVDAKAILSARLSKSRRENLIAQLPGEKKNG<br>LFGNLIALSLGLTPNFKSNFDLAEDAKLQLSKDTYDDDLNLLAQIGDQYADFLAAKN<br>LSDAILLSDILRVNTEITKAPLSASMIKRYDEHHQDLTLLKALVRQQLPEKYKEIFFDQS<br>KNGYAGYIDGGASQEEFYKFIKPILEKMDGTEELLVKLNREDLLRKQRTFDNGSIPHQI<br>HLGELHAILRRQEDFYFPLKDNREKIEKILTFRIPYYVGPLARGNSRFAWMTRKSEETI<br>TPWNFEEVVDKGASAQSFIERMTNFDKNLPNEKVLPKHSLLYEYFTVYNELTKVKYVT<br>EGMRKPAFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEISGVEDRFNA<br>SLGTYHDLLKIIKDKDFLDNEENEDILEDIVLTTLTFEDREMIEERLKYAHLFDDKVMK<br>QLKRRRYTGWGRLSRKLINGIRDKQSGKTILDFLKSDGFANRNFQMQLIHDDSLTFKED<br>IQKAQVSGQGDSLHEHIANLAGSPAIKKGILQTVKVVDLVKVMGRHKPENIVIMARE<br>NQTTQKGQKNSRERMKRIEEGKELGSQILKEHPVENTQLQNEKLYLYLQNGRDMY<br>VDQELDINRLSDYDVHDIVPQSFLKDDSIDNKVLTRSDKNRGKSDNVPSEEVVKKMK<br>NYWRQLLNAKLITQRKFDNLTKAERGGLSELDKAGFIKRQLVETRQITKHVAQILDSR<br>MNTKYDENDKLIREVKVITLKSCLVSDFRKDFQFYKVINNYHHAHDAYLNAVVGTA<br>LIKYPKLESEFVYGDYKVYDVRKMIKSEQEIGKATAKYFFYSNIMNFFKTEITLANGE<br>IRKRPLIETNGETGEIVWDKGRDFATVRKVLSPQVNVKKTEVQTGGFSKESILPKR<br>NSDKLIARKKDWDPKKYGGFDSPTVAYSVLVAKVEKGKSKKLKSVKELLGITIMERS<br>SFEKNPIDFLEAKGYKEVKKDLIILPKYSLFELENGRKRMLASAGELQKGNELALPSK<br>YVNFYLYASHYEKLKGSPEDEQKQLFVEQHKHYLDEIIEQISEFSKRVLADANLDKV<br>LSAYNKHDKPIREQAENIIHLFTLTNLGAPAAFKYFDTTIDRKRYTSTKEVLDTLIHQ<br>SITGLYETRIDLSQLGGD |
| >II-A_WP_024704047.1_Streptococcus_thermophilus_1_Sth1Cas9                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| MSDLVLGLDIGIGSVGVGILNKVTGEIIHKNSRIFPAAQAENNLVRRNTRQGRRLTRRK<br>KHRRVRLNRLFEESSLITDFTKISINLNPYQLRVKGLTDELSNEELFIALKNMVKHRGIS<br>YLDDASDDGNSSIGDYAQIVKENSQKLETCTPGQIQLERYQTYGQLRGDFTVEKD GK<br>KHRLINVFPTSAYRSEALRILQTQQEFNPQITDEFINRYLEILTGKRKYHGPNGEKS<br>TDYGRYRTSGETLDNIFGILIGKCTFYPDEFRAAKASYTAQEFNLLNDLNNLTVPTETK<br>KLSKEQKNQIINYVKNEKAMGPAKLFKYIAKLLSCDVADIKGYRIDKSGKAEIHTFEAYR<br>KMKLTLETLDIEQMDRETLDKLAYVLTNTEREGIQEALHEFADGSFSQKQVDELVQF<br>RKANSSIFGKGWHNFSVKLMMELIPELYETSEEQMTILTRLGKQKTTSSSNKTKYIDE<br>KLLTEEIYNPVVAKSVRQAIKIVNAAIKEYGDFDNIVIMARETNEDDEKKAIQKIQKANK<br>DEKDAAMLKAANQYNGKAELPHSVFHGHKQLATKIRLWHQQGERCLYTGKTISIHDLI<br>NNSNQFEVDHILPLSITFDDSLANKVLVYATANQEKQRTQPYQALDSMDDAWSFREL<br>KAFVRESKTLNKKKEYLLTEEDISKFDVRKKFIERNLVDTRYASRVVLNALQEHFRAH<br>KIDTKVSVVRGQFTSQLRRHWGIEKTRDTYHHHAVDALIAASSQLNLWKKQKNTLVS<br>YSEDQLLDIETGELISDDEYKESVFKAPYQHFDVTLKSKEFEDSILFSYQVDSKFNRI<br>SDATYATRQAKVGKDKADETYVLGKIKDIYTQDGYDAFMKIYKKDKSKFLMYRHDPQ                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |

TFEKVIEPILENYPNKGINEKGKEVPCNPFLKYKEEHGYIRKYSKKGNGPEIKSLKYYD  
SKLGNHIDITPKDSNNKVVLQSVSPWRADVFNKTTGKYEILGLKYADLQFEKGTGTY  
KISQEKYNDIKKKEGVDSSEFKFTLYKNDLLL VKDTETKEQQLFRFLSRTMPKQKH  
VELKPYDKQKFEGGEALIKVLGNVANSQGCKKGLGKSNISIIYKVRTDVLGNQHIIKNE  
GDKPKLDF

>II-A\_WP\_024703962.1\_Streptococcus\_thermophilus\_2\_Sth2Cas9

MTKPYSIGLDIGTNSVGWAVITDNYKVPSKKMKVLGNTSKKYIKKNLLGVLLFDSGITA  
EGRRLKRTARRRYTRRRNRILYLQEIFSTEMATLDDAFFQRLDDSFVLPDDKRDSKYP  
IFGNLVEEKVYHDEFPTIYHLRKYLADSTKKADLRLVYLALAHMIKYRGHFLIEGEFNSK  
NNDIQKNFQDFLDTYNAIFESDLSLENSKQLEEIVKDKISKLEKKDRILKLPGEKNSGI  
FSEFLKLIVGNQADFRKCFNLDEKASLHFSKESYDEDLETLLGYIGDDYSDVFLKAKKL  
YDAILLSGFLTVDNETEAPLSSAMIKRYNEHKEDLALLKEYIRNLSLKTYNEVFKDDTK  
NGYAGYIDGKTNQEDFYVYLKNLLAEFEGADYFLEKIDREDFLRKQRTFDNGSIPYQI  
HLQEMRAILDKQAKFYFPLAKNKERIEKILTFRIPYYVGPLARGNSDFAWSIRKRNEKIT  
PWNFEDVIDKESSAEAFINRMTSFDLYLPEEKVLPKHSLLYETFNVYNELTKVRFIAES  
MRDYQFLDSKQKKDIVRLYFKDKRKVTDKDII EYLHAIYGYDGI ELKGIEKQFNSSLSTY  
HDLLNIINDKEFLDDSSNEAII EIIHTLTIFEDREMIKQRLSKFENIFDKSVLKKLSRRHY  
TGWGKLSAKLINGIRDEKSGNTILDYLIDDGISNRNFMQLIHDDALSFKKKIQKAQIIGD  
EDKGNIEVVKSLPGSPAIIKKGILQSIKIVDELVKVMGGRKPESIVVEMARENQYTNQG  
KSNSQQRLKRLEKSLKELGSKILKENIPAKLSKIDNNALQNDRLYLYYLQNGKDMYTG  
DDLIDRLSNYDIDHIIPQAFLKDNSIDNKVLVSSASNRGKSDDFPSLEVVKRKTFWY  
QLLKS KLISQRKFDNLTKAERGGLLPEDKAGFIQRQLVETRQITKHVARLLDEKFNNKK  
DENNRVVRTVKIITLKSTLVSQFRKDFELYKVREINDFHHAHDAYLNAVIASALLKKYPK  
LEPEFVYGDYPKYNSFRERKSATEKVYFYSNIMNIFKKSISLADGRVIERPLIEVNEET  
GESVWNKESDLATVRRVLSYPQVNVVKKVVEEQNHGLDRGKPKGLFNANLSSKPKPN  
SNENLVGAKEYLDPKKYGGYAGISNSFAVLVKGTIEKGAKKKITNVLEFQGISILDRINY  
RKDKLNFLEKGYKDIELIIELPKYSLFELSDGSRRMLASILSTNNKRGEIHKGNQIFLS  
QKFVKLLYHAKRISNTINENHRKYVENHKKFEFEELFYIILEFNENYVGAKKNGKLLNSA  
FQSWQNH SIDELCSSFIGPTG SERKGLFELTSRGSAADFEFLGVKIPRYRDYTPSSLL  
KDATLIHQSVTGLYETRIDLAKLGEG

>II-A\_HCG2514184.1\_Staphylococcus\_aureus\_SauCas9

MKRNYILGLDIGITSVGYGIIDYETRDVIDAGVRLFKEANVENNEGRRSKRGARRLKRR  
RRHRIQRVKLLFDYNLLTDHSELGINPYEARVKGLSQKLSEEEFSAALLHLAKRRG  
VHNVNEVEEDTGNELSTKEQISRNSKALEEKYVAELQLERLKKDGEVRGSINRFKTS  
YVKEAKQLLKVKAYHQLDQSFIDTYIDLLETRRTYYEGPGEGSPFGWKDIKEWYEM  
LMGHCTYFPEELRSVKYAYNADLYNALNDLNNLVITRDENEKLEYEKFQIIENVFKQK  
KKPTLKQIAKEILVNEEDIKGYRVSTGKPEFTNLKVYHDIKDITARKEI IENAELLDQIAK  
ILTIYQSSEDIQEELTNLNSLTQEEIEQISNLKGYTGTHNLSLKAINLILDELWHTNDNQ  
IAIFNRLKLVPPKVDLSQQKEIPTTLVDDFILSPVVKRSFIQSIKVINAIIKKYGLPNDIIEL  
AREKNSKDAQKMINEMQKRNRQTNERIEEII RTTGKENAKYLIEKIKLHDMQEGKCLY  
SLEAIPLEDLLNPNFNYEVDHIIPRSVSFDNSFNKVLVKQEENSKKGNRTPFYLLSS  
DSKISYETFKKHILNLA KGKGRISKTKEYLLEERDINRFSVQKDFINRLVDTRYATRG  
LMNLLRSYFRVNNLDVKVKSINGGFTSFLRRKWKFKKERNKGYKHH AEDALI ANADF  
IFKEWKKLDKAKKVMENQMFE EKQAESMPEIETE QEYKEIFITPHQIKHIKDFKDYKYS

HRVDKKPNRELINDTLYSTRKDDKGNTLIVNNLNGLYDKDNDKLKKLINKSPEKLLMY  
HHPDQTYQKLKLIMEQYGDEKNPLYKYEEETGNLYTKYSSKDNGPVIKKIKYYGNKLN  
AHLDTDDYPNSRNKVVKLSLKPYRFDVYLDNGVYKFVTVKNLDVIKKENYYEVNSKC  
YEEAKLKKISNQAEFIASFYNNDLIKINGELYRVIGVNNDLLNRIEVNMIDITYREYLEN  
MNDKRPPRIIKTIASKTQSIKKYSTDILGNLYEVKSKKHHPQIIKKG

>II-A\_WP\_003043819.1\_Streptococcus\_canis\_ScaCas9

MEKKYSIGLDIGTNSVGWAVITDDYKVPSKKFKVLGNTNRKSIKKNLMGALLFDSGET  
AEATRLKRTARRRYTRRKNRIRYLQEIFANEMAKLDDSFQRLSEESFLVEEDKKNERH  
PIFGNLADEVAYHRNYPTIYHLRKKLADSPEKADLRILIYALAHIIKFRGHFLIEGKLNAE  
NSDVAKLFYQLIQTYNQLFEESPLDEIEVDAKGILSARLSKSKRLEKLIAVFPNEKKNGL  
FGNIIALALGLTPNFKSNFDLTEDAKLQLSKDITYDDDLDELLGQIGDQYADLFSAAKNL  
SDAILLSDILRSNSEVTKAPLSASMVKRYDEHHQDLALLKTLVRQQFPEKYAEIFKDDT  
KNGYAGYVGIGIKHRKRTTKLATQEEFYKFIKPILEKMDGAEELLAKLNRDDLRLKQRT  
FDNGSIPHQIHLKELHAILRRQEEFYFPLKENREKIEKILTFRIPIYYVGPLARGNSRFAW  
LTRKSEEAITPWNFEEVVDKGASAQSFIERMNTNFDEQLPNKKVLPKHSLLYEYFTVYN  
ELTKVKYVTERMRKPEFLSGEQKKAIVDLLFKTNRKVTVKQLKEDYFKKIECFDSVEII  
GVEDRFNASLGTYHDLLKIKDKDFLDNEENEDILEDIVLTLTLFEDREMIEERLKTYAH  
LFDDKVMKQLKRRHYTGWGRLSRKMINGIRDKQSGKTILDFLKSDGFSNRNFMQLIH  
DDSLTFKEEIEKAQVSGQGDSLHEQIADLAGSPAIIKKGILQTVKIVDELVKVMGHKPEN  
IVIMARENQTTTKGLQQSRERKKRIEEGIKELESQILKENPVENTQLQNEKLYLYYLQ  
NGRDMYVDQELDINRLSDYDVDHIVPQSFIKDDSIDNKVLTRSVENRGKSDNPSEEV  
VKKMKNYWRQLLNAKLITQRKFDNLTAKAERGGLSEADKAGFIKRLVETRQITKHVAR  
ILDSRMNTKRDKNDKPIREVKVITLKSCLVSDFRKDFQLYKVRDINNYHHAHDAYLNAV  
VGTALIKKYPKLESEFVYGDYKVYDVRKMIKSEQEIGKATAKRFFYSNIMNFFKTEVK  
LANGEIRKRPLIETNGETGEVVWNKEKDFATVRKVLAMPQVNIVKKTEVQTGGFSKE  
SILSKRESAKLIPRKKGWDTRKYGGFGSPTVAYSILVAKVEKGKAKKLKSVKVLVGIT  
IMEKGSYEKDPIGFLEAKGYKDIKKELIFKLPKYSLFELENGRRRMLASATELQKANEL  
VLPQHLVRLLYYTQNISATTGSNNLGYIEQHREEFKEIFEKIIDFSEKYILKNKVNLSLKS  
SFDEQFAVSDSILLSNSFVSLLKYTSFGASGGFTFLDLVDVKQGRLRYQTVTEVLDTLI  
YQSITGLYETRTDLSQLGGD

>II-A\_WP\_107392933.1\_Staphylococcus\_auricularis\_SauriCas9

MQENQQKQNYILGLDIGITSVGYGLIDSKTREVIDAGVRLFPEADSENNNSNRRSKRGA  
RRLKRRRIHRLNRVKDLLADYQ MIDLNNVPKSTD PYTIRVKGLREPLTKEEFAIALLHIA  
KRRGLHNISVSMGDEEQDNELSTKQQLQKNAQQLQDKYVCELQLERLTNINKVRGE  
KNRFKTEDFVKEVKQLCETQRQYHNIDDQFIQQYIDL VSTRREYFEGPGNGSPY GWD  
GDLLKWYEKLMGRCTYFPEELRSVKYAYSADL FNALNDLNNLVVTRDDNPKLEY YEK  
YHIENVFQKQKNPTLKQIAKEIGVQDYDIRGYRITKSGKPQFTSFKLYHDLKNIFEQAK  
YLEDVEMLDEIAKILTIYQDEISIKKALDQLPELLTESEKSQIAQLTGYTGTHRLSLKCIHI  
VIDELWESPENQMEIFTRLNLKPKKVEMSEIDSIPTTLVDEFILSPVVKRAFIQSIIKVINA  
VINRFGLPEDIIIELAREKNSKDRRK FINKLQKQNEATRKKIEQLLAKYGN TNAKYMIEKI  
KLHDMQEGKCLYSLEAIPLEDLLSNPTHYEVDHIIPRSVSFDNSLNNKVLVKQSENSK  
KGNRTPYQYLSSNESKISYNQFKQHILNLSKAKDRISKKKRDM LLEERDINKFEVQKE  
FINRNLVDTRYATRELSNLLKTYFSTHDYAVKVKTINGGFTNHLRKVWDFKKHRNHGY  
KHHAEDALVIANADFLFKTHKALRRTDKILEQPGLEVNDTTVKVDTEEKYQELFETPK



QVKNIKQFRDFKYSHRVDKKPNRQLINDTLYSTREIDGETYVVQTLKDLYAKDNEKVK  
KLFTERPQKILMYQHDPKTFEKLMTILNQYAEAKNPLAAYYEDKGEYVTKYAKKGNP  
AIHKIKYIDKKLGSYLDVSNKYPETQNKLVKLSLKSFDFDIYKCEQGYKMVSIGYLDVLK  
KDNYYYIPKDKYEAQKQKKIKESDLFVGSFYNDLIMYEDELFRVIGVNSDINNLEVEL  
NMVDITYKDFCEVNNVTGEKRIKKTIGKRVVLEIKYTTDILGNLYKTPLPKPQLIFKRG  
EL

>II-A\_WP\_124975533.1\_Ligilactobacillus\_salitolerans

MTQNISLGLDIGVSSVGFSILDYKAGEIIELGADLFKASDAEKNQTRRVQRGGRRRLGR  
RKQNRRIDAVKLFDYGLINDSIFGGFEITKENYYLPFNHNQNPYELRVKGQKSKLTK  
EELSFALYNLVKRRGISYTLDDAEAEQEQDSSYSKSLEINKNSLEKKSVAEIQLARLNEN  
HKVRGQVNSNDDTLLNVFPNSAFLQEAKSIIQTQRQFYPNILTADFEKYCNILTRKRE  
YFEGPGSDKSRTDYGIYKTDGRTLNNLFEELIGHDQINPKELRAPNNSYTAQLFNLLN  
DLNNLTITTYEDGKISSADKQEITTLKNATGNVQMLATIKKVAHCTDSDIRGYRTDPLN  
DKPDISSLASYRKVKKEFAKEGIDVDQWPTEFFDQLCFILTLNTEDEGEIRKRIETGLAP  
KYKFLNKRIVQTIIDLKDSFKTSSKNKWHRFSLKTMNQLIPDMLQRPVEQMTLLSELGL  
IKNDHKKFKDNSYLPYKEIAEGIFNPVVAKSRESLKIVNYFIKKYGTLDYIVVEMPRDK  
NEEEQKKFIQNLQKENAKVKDMAITTFSSQLSNPLDLESTLAKSGSGKLLTKIRLWYE  
QDGKCLYSGKTITANDLLQNPNMFEIDHIIPQSVSYDDSLANKTLCYAGMNQAKSQRT  
PYEFLITKDTKGAQTWEQYKAMVKANRTFSNKSNAKAKANLLFKDSLNSIETRKRFA  
RNLVDTRYASRVVLNNLQEFFSEKETGTKVSVVRGKFTSTLRKKWHLNKTRETFFHHH  
AVDASIIATTPFLKIWNRRGGSIFAKKIAENTIDIQTGEILNNTKFNELALSLPFKNFTDQFE  
NIESKIKFSHQVDKMMNRRLSDDTIRSVRTTQIDGETKSFDYMTWKTKDIYTTEGFKN  
FKEIYDKPNPNKFLMKQIDPSSFSKLEKIMSNYPAFEELLDQSSKVKRVDVSPFELYRR  
EHGDIKKDSKKNKGPVIRRLKYYDKKVGTKIDITPSSAKNKKAILISLPWRTDVFFNN  
DTHEYEIMGIKYSDLNVKGTGIIKKQRYQAIKNKENISNNSEFKFSLYENNRIKVLNTE  
TNESVEMLFLSRIKENQTLVELKVIDSDKPHDITLPIYGKVKEGGRIQKRLIPKKYNN  
VCRLWKVNTDIVGNTFYIEKEGDIPKNILV

>II-A\_WP\_164905699.1\_Limosilactobacillus\_mucosae

MDSSKQGLRRYHLGLDIGTNSIGWTAIDHFSLLRVKGKNAIGVRTFKEGETAADRR  
GFRTARRRLSRRRWRLQFLDEFFAPYLAEVDPNFLARLKQSDISAKDPAKNQEFMGK  
LLFPDQEAASDNGGYPTLIQMRKEMPAEKAADFPVFNIYQLRYALMNEKRKFDLREIY  
LAVHHIVKYRGNFLNNTSVDKFKADQIDFNKSFETLNDLFRQVQQIAPFQIDLTQVDEL  
GKLLLDTKQKKLDRQKAAAKLLPLIEDGTDEIVKAHQQTAREISKLILGYKAKAGLILQ  
NNLDQTLDMSTENSDDQLMQIAEELDDYQMEINQLSLLYSQIMLNEIVPNGETVSAS  
MLNRYKHHRRQLKELKTYGAARDKKIREQLDHLAIEYIGQVPKDSKFDFTKDLKKLM  
DKSDLGQKIKGEIEAGDYLPKQRTSANGVIPHQMHQQELDQIIENQKEYYPWLAEPN  
PVEKNIKNAKYKLDQLVSFRVPYVVGPMITLADQQKTSKASFVAWRKETGKITPWN  
FDEKIDREKSANNFIKRMNTKDTYLLSEDVLPASSLLYQKFTVLNELNNIRVNGRKLQP  
ALKQAIYTDLFQKRKTVNLKALTNFLTVHEPTLTCPKIEGLANGRSFNSSLGTIDMSKI  
LGDRVDDLNYRDDIEKIIEWSTVFEDSEIYGVLQEIQWLTNEQREKLVRKRYQGWG  
RLSKLLTGITDENGQRIDLMWDTNKNFMQIVHQPVFEEQITELNQQLLQDADKAPL  
EVVDDILADAYTSPQNKKAIRQVVKVADVVKAVGNAPASISIEFARNEDRNKELKNS  
RRRSIEKLYHETAELVKQTNLLEELGNVKDLNDRYLYFTQAGRDMYDGTIPINIDEIS  
THYEIDHILPQSFLKDNSLDNRVLVKKSENAAKSDRVPAKLYGAKMRAFWKQLKESQ

LISNRKYLNLMTDPEQVDKYQMQGFVHRQLVETRQIIKLAANIFGSLYPDSAVIETRAE  
LTKQLREKFSLYKVRDVNDYHHAVDAYLTAFAGHYLYQRYPKLRPMFVYGDYAKIYA  
DDLKRLRSVNLFDHLEKDQYNSGTAVNPETKEKIVDWDWLSDYVSKIQHFKYMLISKA  
IYTQRGQLYGATILPAAQVDKISAPIGIKNDKPVDIYGAYTNEVAAYMAIVRVNGKKPKY  
KVVSVPVRMLQRLKMAEQEGHEQYMQTLHDILSRNFTSLKKNRKTGLKEAVVKDFD  
VLVPCVRYNQLVIDGDVKFTLGSAAYQHNALQVLSDHSLETLANNFEYVRKHPDEA  
DERMTMLYDDILDQVNHYFTLYDKNKFREKLNAGRDYKYLKLPFSSQFENNKKTAIGKS  
EAIINILRGLHANATITNLKCLGMSTPFGQLQSQSGIELSENAKLVIYQSPTGLFSREVLL  
KDL

>II-A\_WP\_204709078.1\_Fusobacterium\_mortiferum

MGYRLGLDIGIASVGWCILDENDTIMKAGVRLFPEGKSDENQKRREKRASRRLRRR  
QFRIERVRELLRYEIIDNYEYDFYTNETTPYHIRKKGLFEKLTKKELAIALFHLAKRRG  
AGEFNIEVKEDSKEEESTKSILAENEKLLKLGKYICEIQLERFEDKTEEVGRGNKNNFRT  
SDYIKEARKIIDTQREFYKNLDKCFEEYIGIIEGRRKYFEGPGKNSPYSWKNQEEWM  
EGLMGNCTYFPEELRIVKNCYTAEKFNLLNDFNNLRIKRAENDRLTKEEKLKLLNLFKE  
TKTEKSITLKKIASILNVNEEDISGYRIKKDGSPEFTFLKSYIEVNKIFRCDDAEILDNISKI  
ATYYQDIETRKEYREELKDILAITEEIIENLAQKIKFTGSHSLSKKAIKLIEDDLLETSKN  
QMELFNEKGLIPYKMDFKDEKIKYIPKDYVNNWILSPTVKRSMIQTINVVNEILKEHGTP  
EEIVIELAREKNSDDKRKTLIDIQKKNEKENNEIKELLENRKLDKRYFELLKYWKRQDGI  
CLYSGSAIGINEILENPLAFEIDHIIPRSISFDDSDQDNKVLVRRIVENQRKGQRSPYEYFS  
SKMTSRSYEEFKAQVLEMFKNKKISLKKRDLLLLLEEDISKYSKNFIARNLVDTRYATKE  
FMSLLRVFFRDKNKKVKIKAIRGSFTSQIRKEWKMTKMREVSHNHHAQDAYIILIAEKT  
LNKWKWIKNYDENLESSKIYIKREDILSDDELKRLFDYKYSLPKINYKDYLVSFVDDKKP  
NRQLSDETLYSTRIIQEKDKKGKNIIESEYIIGKISNIYDKKSNVIKYFSDEKKHSELLIYQ  
HDKQTFDKMLEIFNEYGGKGAKINPFTKYFEEFGKIRKYSKKNNGPEISDLKYRVSQL  
GNHLDITHKYKESKNRVVMLSIPYRADMYFNGKEYKFISVRYLMLRDKGKCYIIDSKE  
YKKEKNKKGIDENYDFIFSLYTGDIIEERKDGSIKVKFKGVNNDERKCIQVNDNIERNY  
GSYIELIRDIVAKKEEAEEKGSRYNFEEDLSLIIGTTLTKEDEGEKILKSTPISSKQKIVTIG  
GDIKKLRKIYTNILGKEYDGKENFEETIKK

>II-A\_WP\_106190294.1\_Alkalibacterium\_olivapovliticus

MHEYSKQILIIHYCTPFHDKLFNIFIIMSGGFKMTNLVLGLDIGISSVGWGIIDKDTNEVV  
DAGVRLFEEATRANANEDRRNFRGSRRLKRRRKHRLERAKDFLDHGYSSSENIGSVD  
PYRVRYKGLHEELSKKELTAALYHLVKRRGTTIDTPQEDEKTSSELSTKEQLKRNAK  
KLEDKHVIEIQYEKLKNNEVVRNHTNRFKTSDYINEANALLEKQMEFHSEIDEEFQKGII  
DLINNRRQYYDGPSSKKSPTPYGQYFYDEHGILQHESMIDKMRGRGTYFTDQFRIPK  
KAYTAELFNLLNDLNNLRFPSDESGEVEGLTEKHKQYLIENFAEKGGKVTLNIAKLIG  
LFNELDIKGARLDLKTNNPIFTEFVGLKKLKKDLKDITLPADFFNNHDVLDQIAEILTAEK  
SITRREEQLSEVFNTHYDDPLNDVIAALVNDTSFKEYHALSKKAMDILPELWSTSKN  
QMQLFTDHGLGKSRLEQLQKGSKIQFDDEAILSTVARRAHREAIKVTNAVVRDKHGEL  
SHVVVEMAREKNSDEAKQSYRNQKQKQGDFFKKMAKLLVKKELKDLKLNKGKQMLAL  
KLMDSDQDYKCTYTGKQILPHDIVKEHYQFEIDHIIPLSYSFDDSSQQNKVLCYRQENQD  
KGQLTPFQYFQSGKAKRSFEEFKAECMNLFKSRKLSKSKLNYLLEQRDIQHDPEVQK  
NFINRNLVDTRYAMRSFSSNLRTFYMNNEVDTKVLSIRGSFTSALRRRARLHKDRDE  
SHAHHAIDALIVAAIGTTPILKQFKAVDNFREGIAVNTETGEILGEEEEIFSSPTLKFHWL

RNMESEIKYSHKVDRKPNRTMTNQTLYSTREKDGEKYLVAKEKNIYEMDKKGFDAIK  
KKIDKNPDNFLMAQHDPKTWKLKVMEEHKHAENPFQDYFKEHKYILKDGKVPVKS  
LKYLD SKLGIHVDITNKYPESKR VVLLSRKSVRIDVYKNDEGKYKYLGV PYNWFKKIG  
NEYVLDLDKYYGKDGLAAPYKNIDESYEFESFYKNDRISYDKIEKIINPETGEKESQL  
VQYDKLFRGDANPRQNKIEIYDISFKNEKRQFPAPSSLNNIQKYGVDVLGNEFLIDNEK  
FQSKI KSI

>II-A\_WP\_013252099.1\_Olsenella\_uli

MNLRNTTDEYNI GLDMGTNSVGWAVTDKAGAI AHFKGKPTLGSRLFDAAQTAAGARI  
HRSQRRRYVRRRWRLDLLQSLFQDEVSKVDSEFFCRLRQSRLYVG DREKGHADYT  
WPIFNDSDFNEVDYYKRFP TIYHLRKWLMETDEKADIRLIYLATHNIVKHRGNFLRENN  
KSLSSRDANPAQAVKSLRLALNDWC SERDYEC AWKGEEDDGAIDEAILNV LADGEK  
GRSDQAKEIAKLTVDCGDAKQSKQCNEAIAKAIVGLKVEMKNIFGEFPAEKTSYIYLS  
SDEDVEAVRESCPD DSGELFQALCKLYSAYILQELLSYADPSGQTISANMVAKYKQY  
GRDLRLLKELTRTYAPQKYDLFFRGETYAGSSDYV VDEARGYTRYNLGTTKANYDDF  
QKSVKELFKGTDALQDERYQEMMVAFGEQKFLRRLKTS DN GSIYYQLHLEELHTILE  
NQGKHYPFLLSEEDKIESLVSFRIPIYYVGPLTRKNAAIGHDGKERFAWSSRKDGM EH  
AVITPWNWEEIIDRDKSAEDFIMRMTGDCTYLQGE PVLPKCSLLYEEFCALNEFN GAH  
WTVDGDDERRFDAAQRERIMCDLFRRTSVTYNKVSDWLERKYNFTGARVRGGQG  
EHGFESKLGSIYFFCKDVFKTDGLDEADYTMIEQIILWSTLFEDRDILKRKIEEAYGGD  
RLTSEQIKVICKKRFTGWGRLSKRLLTDIKVNTDTGRKSIMDVLREGNPNNGQRSRT  
MVFMEVLRDEDLGFQKAIDGLNRDYFREHN NGLMDVNDLPGSPAIRRSINQAVRIVD  
EIASIARKEPTNIFIEVTREEDDPRKKGKRTKRRYDAIKSALEKFKKDDPSLWKELCEK  
APGDMDERLSLYFMQRGKCMYSGRPIKIEELHTGKYEVDHIIPRSYIKDDSFENKVLV  
YREENQHKT DALLIDSSIRQKMGGYWRMLHDAKLVGDKKFRNLFRNHIDDKAMKGF  
VARQIVETSQMVKL VQTLFEARYPNTRIVPVKAGISHDLREAAGLVKCREANDFHHAH  
DAFLACRIGLFIQKRHPSVYDNPIGLARIMRDYARMQAREYKKSHCLAHTQGFIVNSF  
LSSGFDEETGEIFKDDWDAKAEVEGIRRVLNC RQC FISRM PVEGDTRYGGRYWGETI  
YSPKGKHKHEIPLKQGLDPNLYGGYS DLLSYFLIYEASDVSSGERVLRFIGIPT SIAAR  
MANQKRLEEYVSRLALQENLEHVRMVRKIMKKQLIEVDNCRLLVTAADAVTNATELAF  
SLDEM KIISTCAKCINGADPSAIDKIGNAQDSFFDNLKRRGSDRCALLFSQLKLDKLSQ  
NFYMISSDDKSKILISILDRIRAKSNSVDLSLVGGAKHAAQMKFDFNAKLNNPASSFVLI  
DQSVTGMFERRTRIGL

>II-A\_WP\_076614067.1\_Companilactobacillus\_allii

MNRKTTKYNVGLDIGTASVGWATTGNNYNLLKAKKRNLWGVRLFNTAETAADRRMN  
RSIRRRYRRRRNRNLNWLDEIFSSSELFKTDPGFLNRMKYSWVSKNDKSRT RDNYNLF  
DKDFNDQTYEEYPTIFHLRKRLIENPEKADIRLVYLAIHNILKYRGNFTYEHQKFDVSR  
MNDGLE YTLKELNQALDQFGLSFPNDTDFKLIGDILVKKDWN PSSKVSRIIKELNPTKD  
MKQFYTYVIKLLVGNKADLT KLFNIESNELSPISFSSNSIENDLATAEEVLSDEQYNIILL  
ANSIYSTIVLNNILNGKTYISFAQVEKYTEHHEDLMKLKNIWRNDEDTAAVKKARNAYE  
KYLNNGKYTIQEFYKDIGKYLEEKDDDDSKNALEKIDNNKYLLKQRTSDNGVIPFQLNE  
AELIKIIDNQSQYYPFLKDNKDKILSLINFRIPYYVGPLQSKDKIQSKDKIQSKDKSGFA  
WMARKENGPIRPWNFDEKVDREKSSNNFIRMTSTD TYLIGEPVVPKNSLIYQKYEV  
LSELNNVKIVSTGEGSENQERLRVEVKQRIFNELFKKYNTVSAKRLKDWLIKESYYS A  
PEIHGLSDKTKFVSSLSSYRKLSKIFGNDFVDNVKNQDQLEQII EWQTVFEDREILKLK

LNKSNQYDEKQINQLVAIRYQGWGRFSNKLLTQLFVNTKIGNEHEPSNHSIIDLLWQT  
KSNLMEILRDDKYNFESQIKELNIEDSSDKKPLELVNDLHGSPALKRGIWQAISIVQELS  
EFMGHAPEHIFIEFTRDDQDSSITKSRYNSLKKRYQDIKQMVTDLAPTLKESLFPTKDL  
EDLMKDKRNSLSNQRLMLYFSQMGRSLYSDAEIDITRLFTSDYQVDHILPQSYIKDDS  
LENKALVKASENQKQDDLLLSKDIIANNLTRWEYLKAGLMGPKKFANLTRTVVTDR  
QKEGFINRQLVQTSQMVKNVANILDSIYPDTQVIETRASLGMGFRDSFSNLNKKTWH  
YEHPEFVKNRNVNDFHHAQDAYISTIVGTYQLKKYPRDNMRLVFNAYSKFFEDVKKK  
TRQERGKIPAYSSNGFIIGSMFNGKTQVNKNGEIWDQQIKDSISKTFKFKQYNITKQN  
YINDGALYKQTI LNKNKELIPLKKDLDPHIYGGYTGDITSYSVLIDVDGKKKLISIPVRIA  
REITAKRINIKDWISNKVKHKKEIQILIDVVPVGQLVKSGDKGLISLPSGTEIANANQLILD  
YKETALLSLEHSTLDNYRFILSGDNEDILQSIYSDLIFKIQKLYPLYSSSESKRFNDNLDE  
FNNCSIYDQFNIIEQILNLLHANSTCANLNFGNIKSTRLGRRSNGYEFSDSDFIYKSPTG  
LYESIIHID

>II-A\_WP\_220751152.1\_Apilactobacillus\_sp.

MKNYHLGLDIGTSSVGFVAIDNNNIIHKHGKNIIGARLFNEGQTAEERRTKRGARRR  
YSRRRWARMHLLDNIFEPYLNVDENFLQRMHESDMSPKDENKKYYGSLLPDITDST  
YYASNQTIYHLNRMLMNSNEKADIREIYLALKHIVKYRGNFLDSTPVSSFEISDLHLEDI  
VYKVNELYEKL SINFKINVDNVNKISAILLDNKITNINKQKQIKLLQVHSDMQDKTLQKA  
VDKLNKDIA TEITKLVLGYKTDIKLLHFENDDQFKLCLSDENSEDQLASVMEVLNETQ  
GTILTELKTLYSRVRLNQIIPNGKSLSQSMIDKYNQHKEQLRILKDSVLNVLP SNSDLKR  
NIKILYAMYVGNFDSEEFKNLKAVEDCLSNSINDSNAKTEIKRLFKNERHVISKTAFY  
NALLSLIGNPVLYDSLKKIVNNNLDGMDIKYFKDAIKQYLNDDKSSKLSKLIFNMESKVT  
KNSKYVNSSDMQAILDNPISHPWGNNEAMKATIENTIRQDINNNDYLPKQRTNENGNL  
PHQVHQLEMDKIIDKQKQYYPFLAELNPNQKRQHI AKYKLDELVTFRVPYVVGPMITP  
EDQTKTSGANFAWMKRKAAGEITPWNFDEKVDRTKTANAFIKRMTVKDSYLLSEDLV  
PDNSLIYQEFKVLNELNMVKVNSHRLTTAQKQAIFNDL FKQHKNITIKRLQTYLMSAFH  
YPSTPNITGLADEKHFNNSLSTYVDFKNILNPDVFENEKIKNDLEKIVEWSTVFEDRNIF  
ADKLKELDWLTDEQRKQFVNKRYSGWGRLSKRLLTEIVNENGERIIDTLWNSQVNF  
QAVSNPSIKKQIDKINGKQLKSLGMESILDNAFTSPQNKKAIRQTIKVVIEDIERAMDGQ  
KPSSVSIEFTRSPKDSSEISKTRFKNIDNLFKSLASEVSKQLKDELKDTKKRMGKLTQ  
YYLYFTQLGKDIYTGE PINIDDPNYHIDHIIPRA FVKDNSLSNRVLTNHKTNEDKGNNT  
PLQGLNISENVKQSWHKLAKMELIPKRKLTNLCTTLDNVN KYARRGFIQRQLVETSQV  
IKLVANILGDRYNGDTKIIEVKAKMNSEIRKTFNLYKVRAVNDYHHALDAYLTAAIGRYL  
YQRYPKLRNYFVYDEHKQMDIDDLKDLKTFDFLHEITNGKKSEIRDNVKEEVILNRDSL  
IEKLNDIYKYKFMLVNHEVTTRKGALFDASLYSPKAGKKHLIPLKNSRNP SVYGGHSG  
EKSAYMALIEMTDKQQNKTYKFIGIPMRYANKLKRLSNTDEQAYNEILN NIAKQNISKQ  
TIAFRVLLNKVMYRQLIVDGDVKYTLGSAAYKYN AKQLVLSVDSMKT LGDKDFISSLSE  
DELSQRYINVYKDILEKLNKYMPIYDTNNFRNKLNEGLEK FVKLPNYDKKTNSKVAKG  
KYMILINILES LHNNPTNIKLNELGMSTPLGKMQIKTGIPVSPNARIYYQSPTGIFSHYV  
VKDL

>II-A\_WP\_125715884.1\_Agrilactobacillus\_yilanensis

MMKDYSVGLDIGTGSVGFVAMDQQYRLMRAKGKNIIGVRLFEP SKTAEKRRRRFRTR  
RRLSRRRWRLSLLND CFAPELQNQGDPNFLHRLKYSWVHPKDTLNQQNYSAAPLFN  
DSGQDKAFYNKYPTIYHLRRALMNDTKQHDLREVYLAIHHIVKYRGNFLLGDGQIDQE

AIFDIHSFLEAMQDFCQPILSDESTLIDVDEARLVITILFDKQLSRAKRVTEGVKVFDGQ  
KNKALKTVIKLSLTGIVGNQFNLATMLQKEGLNPKDSKWKIKFTESDLEAKLETINSGTI  
LSDSEKEFLTAIQKAYNGLTLKLLDDAHWSWSEAMIKLYEAHKGDWKLKIKELRTSENK  
AEVNQAYLNLQADNDDVRRKANKYFTGLIKKSTTLTKEKQNELLSKIENEIFLPVQRSK  
ANVSIPHQLHLNELIKIENQKQYYPFLAEVDKKGRNKLQQLLKFRVPYYPVGPLVEPKT  
VDGDSSNHWMTRLEAGPITPWNIEEKVDTDQAGLDFIKRMTGTDTYLIGEPALPKNAL  
IYQSYNVLQELNNVRINVNQRLEVAQKQEIYENIFKGQKSVSVIDIKDYIKTKYGKDVNV  
SGLSDTKGGHFNSKLTYYHYFLNIFDRAYVEDHDNQETLETIIELTIFEDRKILKRQLS  
KLSVLTkdQVESLSKQHLTGWGRLSKLLTTKFVQLSTKAGTDKKSILETLYETKMNL  
MEIISDSRDRYGARAWLEEENAIEDKETSLSKSVINELAGPKNIKRGIQSFIYLDEITRVL  
GRPPKNVYLEVARNTQASELTNSRLAYLKTVYGNKSLSAEFKALSTQLNSVEDAKMN  
AALRDDRLYLYYLQQGKDMYTGEPIINIDLISSNYQIDHIIPRAYGIDCLDNRVLVKSELN  
ARKADSPTYLPELICSRCNWWRSKTAGLLSEQKFKNLTRKSDDFSDLQKERFIARSL  
VETRQITRNVATLIGEHFNDETkvRAIRSEITTDmrHYLAIYKNRDLNDYHHAFDALLFT  
TAGLFMDNIGFFFKGQLSDYAGHIYNRYTKQALAQARANAKDKHQQRIRPLGFVVGs  
MASSNKQFQVNQATGELVWTMADEAYLRKVM DYKKILVTkMVYDQTGELYKQTRIP  
RDGKEYGKRKIYgKvAFDKRKDLALYGGFDSLQTAYTALVFSKkkfKLVNVQRLWLS  
EIKADANYLTKMVQQSLNDPKAKIILDHIPYEQLLIKDQAKVTIASENELHNfQQLYLSK  
TDSKDLLILLNARSEEMAEARLKQYSHSQDAKQRFTELLTHIEQALKYyKLQTHQTAL  
LKLQDKVIEFTTLEFEKQQEVVQRVLQALHANAIrSNLEILGLSESWGRLNAMVNLSP  
SDKFVYTSYTGlyTSTVSVQELADQAKLE

>II-A\_WP\_154548804.1\_Lactobacillus\_porci

MAKPKNYTIGLDIGTNSVGWTVTDDQNNILRIKSRKAIGARLFEEGKVAAERRSFRTT  
RRRLSRRKWRIKLLLEELFDEEISKVDPNFFARSHESWISPKDKRKRYSAIVFPTPEED  
KKFHDEYPTIYHLRDRMLMKDDQKHDIReIYVAIHhIMKYRGNFLHDESIETYRSGMSSL  
GGGDERTILSTKILDDLNNAFADNLATQGISVNSASADQINEIVISGHLNADSQKAIACL  
LLPTTLSSDDKKQEKKTKKLVTDAAKNVASAWLGYKANFAAILNIQEVDKDKKKVYNF  
ALQGADEEGKVETLESILSQDQIEVDQIIRIRHAIVLSEIIPVGLTLSEAMIEKYDQHAKD  
LKNLKTvirNLKDRKNAIKLQAIYDLYVNNRHADAAMKLAGVKKRTELLEPEEFNKS  
IEKLLDDSPVSAEIKKRIAekDFLPLQRSNSNGVIPNQLHQIELDEIiKKQSEYYPFLAET  
NEHASVDVARKAPTKLDQLLTFRVPYYPVGPLITKEEQEKQNGHSFAWMIRKNPKDHE  
KITPWNFDEKVDrmASATQFIKRMttKDTYLLGEDVLPSSSLKYQLFTVLNELNNLRIN  
HKKLTrDEKEQVIEDLFKKYKTVKTDKfKNYwQKGhAGLDVkiQGMADPEKFNSTMS  
TYIDLRRIFGDQLNDDQQRQNDFEKIIeYSTIFEDRKILGEKLradFTWLSEEQIKQLCRL  
RLQGWGRLSLkLLTRLTDADGQNVLERLLNSNDNFMQIVKDPVIKAKIEQKNQKLLRA  
DQSENATESILDEAYTSPQNKKAIRQVMRVVHDIVKACGGNLPakFAIEFARSQDDDP  
KRTVARADRLQKIYAQISDEVVSQGVKDQLNGMKSLKDRYYLYFMQGGGRDAYTGEP  
LNIDRLSDYDIDHIMPQSFikDDSLDNrvLVTRAVNNQKSDNVPVLLYGNHLVADLGIT  
REMWDRWQKSGMISKrKLANLLTDPATLSEYRAEGFINRQLVETSQVIKLTATILQSE  
FPDSKIIeVPAKYNSILRKQFDLYKSREVNDYHHaIDAYLSTIVGNyLYQVYPNLRRLFV  
YGEFKKFShGEEESSHQAVRRAKSMNFLDdLLRGKHGDDIYCRSTGEIVFSRKDIISK  
LKQAYSfKQMLVTQEVFTKKSALFDQTVYPSPARDSKTRSGlipRKQGMDTELYGGY  
SGNKDAFFALVDVEKKKGHTLRIVGIPVRDFESLKNASNYSEKLMEIiKSQVLFDRNTG  
KSIRGILNVKIIIDKILYRQPIEDDSYMLGSSKYRYSLKQISLSQRSMKYILDYIDDPDF  
NKHVAVHIEQQDEKQCllSVYDEILEKMDKYLPLFDIRSFQRKLHDGREAFIALSSTEE

ALAINKILRGLHANAEISKLPGLFGQAALGALVSTSGIKISDDAVFIYQSPTGLFERRV  
KVSDLLK

>II-A\_WP\_006907934.1\_Facklamia\_hominis

MYSVGLDIGIASCGYSIINDVNGNIVELGTLLFKSRNSDNEERRTNRGSRRRLINRRRT  
RLDDGKKLLAEKGFKQDSKCLKHVCPYELRVKGLHEKLTQGEIYKVVLHILKKRGISYLD  
ESDENSSGIDGASYKDQVNLNLNLQKELTPGQIQLKRLKENGRIKSGVTFENG YQLN  
VFSVSAYANELNKILTKQSEYYSEIDQIWIDRFTKNGYRKEAGLVYRKRPPYYDGPNE  
KNSIYGRWANFKEDGHPKENIFDQLIGLDSVDSTQVRASASTLTAQKYNLLNDLNNL  
TIIEDDRKFTTDEKVSIIINYLMTQDITRFGPKELAKFYQLEVDQIRGWRIDNKG NPEIHS  
LKVYRDWKKIFNTSSIDLTSIQTWVDELANVITL NTEIPSTLATLKIKIPDIDQQILNFVE  
NNFYELKKKTSKNSWHSFSLKTLNLIPELLNTSEEQNTILERLNLKFDLKNKYAEKNQI  
PVKELLEQIYNPTVTKSVHQSIKLFNEIIEKYGKNNISYVTIEMPRDKNDQDQKDTIKRIN  
KLNSERKNDSEKYFLERSGWD TIRFESALKKPRFAEKLLFYEEQKGKCAYSQQDILAE  
SLLNESCEVDHIIPLSISFDDSIHNKVLVTASSNQEKGQRTPYMAFIDGAKLGQSWDDF  
VAWVNSCGFHRKKVQNLLYEENIFNPDVQRGFVSRNLNDTRYASRIVLNAIQSFFYK  
QKDLTKVKV VNGSFTHTLRKKWGSSLDKNRDTYHHHAVDATLCAVTPFINIVKYEFFE  
EDGIKYMVDLTTGEKVAYS DYKKMKFYDKRAYVPIWDGFIKQLTPANLYPKIKFSHQV  
DSKWNRKISDATIYSTRKIVETKINKKG VAKETVKDIVIDKIKDIYTPEGWRKFKEKQDK  
LLMKNFDEKTYNKLIIKIAKEYPEKKEVQTSNGKVSLVEVSPFKMYCEEQNVP AIRKYS  
KKGNGPYIRSLKFYDKELGKHINITKDSYGNDIERTIRNKKIILKNQNSWR TDVYYNSV  
NNCYELLGIKYCHLKFYRGKYGIPKSIYDELKKEDEISSESVFKFSLYRHDRVRFEKDG  
EYIEALFGSKNESNKNYLELSPIDKNKWLSKENVKFFGEVAASGRFIKGLKPGIKVIKQ  
STDYMGHVYAIHQEKLK FVLSDA

>II-A\_WP\_195217860.1\_Turicibacter\_sanguinis

MGYTLGLDIGVASVGFLINNESGEVLEAGARLFSSANPDNNQTRRSNRGARRLNR  
RKVHRLERIRDLFENYDILQGDIQEFTPYHLRV RGLNEKLELPELYAALYNLAKHRGIS  
YLEDVDVSELKGNNPLEKNIDLLKELYPCQIQLQRYETYGQVRGMIEFQSSSGEIDYL  
VNIFPTPAYKAEAETIIRKQMEFYSQLNEDFLKEYITILTDKREYFVGPGNDKSRTDYGI  
YRTDGTNLTNLFDILVGKCSIYS DERRGARASYTAQLFDLNDMNNLT VAGEKLTFEQ  
KQQIYNEIMSSKSVNVIKFIMKVTHCKKDDIKGYRVDSKQKPEFHSFEGYRKL RNHLLL  
TLDFQLDDVSREDRDKIAEVLTLNTEYMQIINQLKEQVNILSEELCEEIASFRKSCGSLF  
SKWHLSLVKAMREIEEELWQTSYNQMQLFTQNGVFKSKSERFKNLKYIPTELLLEEIY  
NPVVKGSIRESIKVINAILKKYGTLD AIVIEMPRETNEKEEKQKIEKMQKANANQK KSAI  
AKAKEEYSFGKEAYRQQNNDLKLRLWYEQKQRCLYSGKTITIYDLVNNPHLFDIDHII  
PKSISYDDSLNNKVL CYASENRQKGQTTPYKYFMRKSNSDWNYEQYKSYIMTSLYID  
GKGEMPRSKVDYLLFEEDLNKYEVRRQQFINRNLDTRYASKMVLNTLQSFMKENYE  
HTKVAVVRGKFTAQLRRKWKIVKNREESHSHHAIDALIISVPHLNLWKKAKNQVLSE  
KVIYDEQTGEVLNQDYDSL VYQEPFSKYFIQLEQVLTNCKYSYKVDRKPNRQVADATI  
YSTRTG VVKDSKGDKSFKIDVDSEEHYVVAKIKDIYQDNDAKRFMDRYKKDKTQFLM  
YHHDPETFKILEEVMKTYPTSKNPFADYAKEYG TIRKYSKKNNGPEIRSIKYYDSRLG  
QHIELQPGQSKSKKVVLQSLNPWRS DLYYYKKEQRYIVAGLKYADLSFQKGTGEYGI  
SYEKYQYILQLEGIH LTIREIKDLIENNQEPQEYSFCFSLYKNDFIQWQDEENILQQYRF  
LSRNLSSKNRIEVKPKNSKYEKQADGLKTIRKDIECIQKVSVDVLGYQHII NFEKLKLT  
K

>II-A\_WP\_132252121.1\_Natranaerovirga\_pectinivora

MKKNERYYVGLDIGTNSVGWAVTDDKYKLYKCNGHKMWGVRLFDEAETAESRRVS  
RTSRRRLQRRNARIDLFQELMAEEICKVDPGFYQRLEDSKFYMEDKTVVEKHTLFIGD  
KYNDKDYYSEFPTIFHLRAALIKGEKEYDIRLIYLAIHHILKNRGHFLYQGDKFDTS  
DESIHEVFDNNDVDVTNGIEVTVELVSKIKDVILNERLTKTDKKKKLKYLFGKSKQLESIF  
GLAIGTKESLEKVFNGPEYKELENDVVKISFSDKVYDEVRLHYEEVLEDRLVIDNLKKI  
YDTVLLNGIKKAELSLSESKVELYQKHKEELTTLKKIHKAYSQEEYKRLFSEKDSNSTNY  
KNYIGDGDKKCSREDFYKTLKALLEKVEDSPVKEYILEEISLDKYLPLQRVKENGVI  
QIHQEELKLILDNASGYLDFLNQTDGKLSVKDKIMKIMTFRIPYYVGPINTYHAGKKN  
FAWAVKNSDEKITPWNFDDIIDLEASHDNFIRKMTNKCTYLRGKDVIPKSSLLYSEYTL  
LNELNNIRCNGEKVSVSTRNKMIDSLFKDTDKKGKVTVKKILEFLKCEGECEGNATITG  
IDDEVKADLKSYPDFKSVLGEKFNYEMVEDIHKWITLYANEHNSIKKRIKDNADKLTSS  
QINKICKLRYKEWGRLSKEFLTEIICDEFTDYSTGEVGNVINAMRNTTNNLMQLLSNKY  
DYIKQIKEINDLLINPNEEISHEMLDELVSPGVKRMIVQSILIMEEIKKIMSREPEKIFIE  
TVRSNKANKKRTDTRKKKLLELYSSCKDEVIDWPKEIQGRSDGQLRSKKLYLYLQ  
GKCLYTGEKINLDKLMSGEDYDIDHIYPRSKTKDSDFDNLVLVVRVINNEKSDNYPINP  
AIQAKMKKTWDYLYQKGFISTRKYERLVRKEEFGTNELADFIARQLIETSQSTKAVAEI  
MQRLYLDSHIYVKAENVSAFRNDMKFIKVRDINDLHHAKDAYLNIVVGNVFDTKFTRS  
PANFVKEAGFREYNLSRMYDFLIKRDGYIAWDGRNGDSKKMVDYQMKSNDRVTRR  
PTEQKGQFLDLTLHKKGDSKAHSYMGSKTGDLRLADVTKYGGYTSIKIAYLIPYSCDII  
NKKGARKSISRLTGPVYLANNVKSVGEFTYIFSQVPTKSGETIENFKIINVKLRI  
KYNIFYVVGRTGENFCADNAIQLLLGEDEERYIKEIFKFINNQKENKDLKAEKFS  
DITKENNYKLYDALVNKLNSNIYVQSTNNKYDVFNDAMRNKFCDLVVEEQIKMLLNL  
LNLLTNKTSTYDFKPLGFGIGRRKLGFNLADVKEFKLINQSITGLFENSLDLLSQL

>II-B\_WP\_003038941.1\_Francisella\_novicida\_FnoCas9

MNFKILPIAIDLGVKNTGVFSAFYQKGTSLERLDNKNKGVYELSKDSYTLLMNNRTAR  
RHQRRGIDRKQLVKRLFCLIWTEQLNLEWDKDTQQAISFLFNRRGFSFITDGYSPEYL  
NIVPEQVKAILMDIFDDYNGEDDLSYLKLATEQESKISEIYNKLMQKILEFKLMKLCTDI  
KDDKVSTKTLKEITSYEFELLADYLANYESLKTQKFSYTDKQGNLKELSYHHDKYNI  
QEFLKRHATINDRILDTLTDDLDIWNFNFEKFDKNEEKLNQEDKDHIQAHLLHFV  
FAVNKIKSEMASGGRHRSQYFQEITNVLDENNHQEGYLKNFCENLHNKKYSNLSVKN  
LVNLIGNLSNLELKPLRKYFNDKIHAKADHWDEQKFTETYCHWILGEWRVGVKDQDK  
KDGAKYSYKDLCLNELKQKVTKAGLVDFLLELDPCRTIPPYLDNNNRKPPKCQSLILNP  
KFLDNQYPNWQQYLQELKKLQSIQNYLDSFETDLKVLKSSKDQPYFVEYKSSNQQA  
SGQRDYKDLDAFILQFIFDRVKASDELLNEIYFQAKKLKQKASSELEKLESSKKLDEVI  
ANSQLSQILKSQHTNGIFEQGTFLHLVCKYYKQRQRARDSRLYIMPEYRYDKKLHKY  
NNTGRFDDDNQLLTYCNHKPRQKRYQLLNDLAGVLQVSPNFKDKIGSDDDLFISKW  
LVEHIRGFKKACEDSLKIQKDNRGLLNHKINIARNTKGKCEKEIFNLICKIEGSEDKKGN  
YKHGLAYELGVLLFGEPNEASKPEFDRKIKKFNSIYSFAQIQQIAFAERKGNANTCAVC  
SADNAHRMQQIKITEPVEDNNDKILSAKAQRLPAIPTRIVDGAVKKMATILAKNIVDDN  
WQNIKQVLSAKHQLHIPITESNAFEFEPALADVKGKSLKDRRKALERISPENIFKDN  
NRIKEFAKGISAYSGANLTDGDFDGAKEELDHIIPRSHKKYGTLNDEANLICVTRGDNK  
NKGNRIFCLRDADNYKLKQFETDDLEIEKKIADTIWDANKKDFKFGNYRSFINLTPQ  
EQKAFRHALFLADENPIKQAVIRAINNRNRTFVNGTQRYFAEVLANNIYLRAKKENLNT

DKISFDYFGIPTIGNGRGIAEIRQLYEKVDSDIQAYAKGDKPQASYSHLIDAMLAF CIAA  
 DEHRNDGSIGLEIDKNYSLYPLDKNTGEVFTKDIFSQIKITDNEFSDKKLVKKAIEGFN  
 THRQMTRDGIYAENYLPILIHKELNEVRKGYTWKNSEEIKIFKGKKYDIQQLNNLVYCL  
 KFVDKPISIDIQISTLEELRNILTTNNIAATAEYYYINLKTQKLHEYYIENYNTALGYKKYS  
 KEMEFLRSLAYRSERVKIKSIDDVKQVLDKDSNFIIGKITLPFKKEWQRLYREWQNTTI  
 KDDYEFLKSFFNVKSITKLHKKVRKDFSLPISTNEGKFLVKRKTWDNNFIYQILNDSDS  
 RADGTPKPFIPAFDISKNEIVEAIDSFTSKNIFWLPKNIELQKVDNKNIFAIDTSKWFEVE  
 TPSDLRDIGIATIQQYKIDNNSRPKVRVKLDYVIDDDSKINYFMNHSLLKSRYPDKVLEIL  
 KQSTIIEFESSGFNKTIKEMLGMLAGIYNETSNN

>II-B\_WP\_014548420.1\_Francisella\_hispaniensis

MNIKILPIAIDLGAKN TGVSFAFYQKGTSLES LDNKN GK VYELSKDSY TLLMNNRTARR  
 HQR RGIDRKQLVKRLFKLIWTKQLNLEWNKDTQQTISFLLNRRGFSFITDGY SPEYLN I  
 APEAVKAILMSILDDYNGEDDLDSYLQSATENDSKIDELYNKLLQKALEFKLRKLCIDIK  
 EDKVTTKTLKELSNTEFKLLANYLV DYDRILRTQKFSYTDKQGNLRELNYHHDKYNI  
 QEFLKRNIINDVILEKLTDDLDIWNLNFDKFD FEKNLEKLENQEDKDY LQTHLHHFVFA  
 VN KIKSEMASGGRHRSQYFQEITNVLVENNHQEGYLKNFCENLHNKKYSNLSVKNLV  
 NLIGNLSNLELKPLRKYFNDKIHAKADYWDEQKFAETYSDWILGEWRVGAKDKDKKD  
 GAKYSYKNLCDELKQKVGINQDGIINQTKSDFVD FLELDPCRTIPPYLDN NNRKPPK  
 CQSLILNPKFLDNKYPNWQQYLQELKKLQTVQNYLGNFEIDLKDLKSSKEQPYFVKYK  
 SLNQQIASGQRDYKDLDARVLQFIFDRVKASDELLLNEIYSHAKKLKQNTSSELEKLEL  
 CKKLDEVIANSQLSQILKSQHVNGIFEQGTFLHLVCKYKQRRQARDSRLYIMPEYRY  
 DKKLDKYNN TGRFDDDNQLLTYCNHKPRQKRYQLLNDLAGVLQVSPNFLKDKIGSDD  
 DLFISKWLVEHIRGFKKACEDSLKIQKDN RGLLNHKN IARNTKGKCEKEIFNLICKIEGS  
 EDKKGNYKHGLAYELGVLLFGEPNQASKLEFDRKIKKFNSIYSFAQIQQIAFAERKGN  
 ANTCAVCSADNAHRMQQIKVAKPVEGNKDNIILSAKAQRLPAIPTRIVDGAVKKIATILA  
 RNIVDDNWDNIKQALSNNQQLHVP IITESNAFEFEPALADIKGKSLKDKRKKALERIDP  
 ENVFKDKNNRIKEFAKGISAYS DANLANGDFDGAKEELDHIIPRSHKKYGTLNDEANLI  
 CVTREDNQNRGNKAVFLYDLKPNYKLKQFN TDDLEIEKKIADTIWDASKQDFKFGNY  
 RSFINLTPQE QIAFRHALFLADKNPIKKAVIRAIDNRNRTFVNGTQRYFAEVLANNIYLR  
 AKKENLATNRITFDYFGIETTNSNGRG IADVRKLYEKVDSDIQAYAKGDKPQDSYSHLI  
 DAMIAFCVAAD EHKNEGSIGLKMGNQYGLFPTPD KYDENTGEIISWHNDDIFRKIKVS  
 DNQFSDKKLVRRKATEGFNTHRQMTRDGIYGESYLPILIHKNLNEVRKGYNWENSEEI  
 KIFKGKKYDIQQLNNLVYCLKFVDKPISIDIQIT TLEELRNILETNNISTTA EYYYINLKTQ  
 KLHEYYIENYNTALGYKKYTKEMEFLRSLAYRTERVKIKSIDDVSMILAKDSNFKAGKIE  
 LPFKTEWQRLYLEWQNTTIKDNHEFLKS YFSVKNATKQHKKVRKDFSLPISTNEGKLL  
 VKRKTWDNNFIYQILNDSDSRADGTPKPFIPAFDISKNEIVETIISFTSKNIFWLPKNIEL  
 QKVDNKSIFAIDTSRWFEVETPKDLIEIGVSTIQYKIDNNSRPKVRVKLDYVMDDDSKIN  
 YFMNHSLLKSRYPDKVLEILKQSTIIEFESSGFNKTIKEMLGMTLAGIYNETSNN

>II-B\_WP\_211490026.1\_Francisella\_philomiragia

MNIKILPIAIDLGAKN TGVSFAFYDKGLKLENLNNKAGKVYELSKDSY TLLMNNRTAKR  
 HQR RGIDRKQLVKRLFRLIWNEKLNLEWDKNTQQAISFLLNRRGFSFLEEKHDASYFE  
 KFP IEAIELLPQQLRLDKSSDEKYDLNSKIFETLENDTFKDV FVEIENNIKQVKKNIVKAN  
 IINIIKSAIDDKVSKGKIESTKQNKFDELKKVSSWIVEDIFDSEIFAYIDNVDNYSKNIFNY  
 LNSLSNDQLEALRKSLATIDLNEYKNSVWNFNISN FNLDNEKLEKLEAGDISAHLHL



GYALQNIKKELDEGSRHRSKYFEEVASVLEQNHKEEYLRSFQCNLHAGKYINLNVKK  
LVNLIGNLSNLELKPLRKYFNDKAHTQADYWDEQKFAQTYCRWILGEWRVGSKDKD  
KKDGAKYSYKELCNDLKREVGTQNIKSSIEFLADLDPCRTIPPYQDNNNRKPPKCQ  
SLILNPKFLDTKYPEWQSWLDKLKSIQSVKEYLAIFEDDLKDLKTGKGQPYFIETKSSN  
KQIASGQRDKKDLARVLQFIFDRVKASDDLLEIYSQAKVKQNASSEQDKQKAFN  
NLKAVLLDSHLSNELKPEGHFDKLSDLKNDIFEQGTFLHLVCKYYKQQRQARDSRLYI  
MPEYRHDKKLDKYNNTGRFDDDNQLLTYCNHKPRQKRYQLLNDIAGVLQVSPNLLK  
DKIGSDDELQIIEWLKNFKVASYCKAAVDMQKQYRGTLKNVAQTAIFKQRLEIHKNNK  
ASDEEKALVEKYKNAKALTSDENKLVKLVENIAKSSQKIGENLGLGDKQIAKFNSIYSF  
AQIQQIAFAERNGNANTCAVCSADNAHRMQIISALSQSKCDTQTSNVEASTSSAILDK  
TNISHNKVTELVEVKTNDTPNDFDKAQSSLGAKAQRLPAIPTRIVDGAVKKIATILARNI  
VDDNWHNIKQALYDKQQLHIPITESNAFEFEPSLAEIKGTKKKDRISSQELFAKKENRI  
KNFSKGISAYSDTILADGDYDGSKEELDHIIPRSHKKYGTLNDEANLICVTRDDNQKRG  
NKAVYLSDLKPNYKSKQFGTTDNSQIEKYIAEKIWDAAEKQDFKFGNYRSFINLTPEEQI  
AFRHALFLADENPIKQAVIKaidNRNRTFVNGTQRYFAEVLANNIYLRAKKENLDTNKIS  
FDYFGVPTTDSGRGIADIRKLYEVIDSDIQAYAKGDKPQDSYSHLIDAMLAFCVADE  
HKNDGSIGLEMDNQGGFLPTPDKYDENTGEILSWCDDDFRKKIVADSEFGESKLAR  
RKAYTVETNIRQEIQSRGDFKKSIPYKIHIDSLAERFFPLLQISEDKWKYGFTLDNSVD  
IPKKDIELIQEFLTKSKEYKDYTVWIVNKTQAQEFLINIGYVGADKQQEKIAKLLDKLSYT  
TIKKEITSILAKANKTPDTVGDALDVIDELVFKTGKKNKLINFVKSEILLPVYNDWIKLKK  
AMQKADMNQNLHDFLSMYFITTNFKNHQKVRKNFSLPVVSTIGTIRIRKSWDGS  
YQVTGDESLAKYGFDTVRPKTILSRNSVPIKYNGKEKVVLIEPLSWINLDIDNLTDS  
SIISAQIKNRSAGQVSIKVELSSLDNICLKEDSWLGNIVYAGDKDKPNIIGDKFYIKED  
FYWLKDFSLMFDGSKTKASLSIDKIGNSYILEFVSSKGKKEKQWLKNETPNS

>II-B\_WP\_011212792.1\_Legionella\_pneumophila

MESSQILSPIGIDLGGKFTGVCLSHLEAFAELPNHANTKYSVILIDHNNFQLSQAQRRRA  
TRHRVRNKKRNQFVKRVALQLFQHILSRDLNAKEETALCHYLNNRGYTYVDTDLDEYI  
KDETTINLLKELLPSESEHNfidwFLQMQSSSEFRKILVSKVEEKDDEKELKNAVKNIK  
NFITGFEKNSVEGHRHRKVYFENIKSDITKDNQLDSIKKKIPSVCLSNLLGHLSNLQWK  
NLHRYLAKNPQKFDEQTFGNEFLRMLKNFRHLKGSQESLAVRNLIQQLEQSQDYISIL  
EKTPPEITIPPYEARTNTGMEKDQSLLLNPEKLNLYPNWRNLIPGIIDAHFPLEKDLEH  
TKLRDRKRIISPSKQDEKRDSYILQRYLDLNKKIDKFKIKKQLSFLGQKGQLPANLIETQ  
KEMETHFNSSLVSVLIQIASAYNKEREDAAQGIWFDNAFSLCELSNINPPRKQKILPLL  
VGAILSEDFINNKKDWAKFKIFWNTHKIGRTSLKSKCKEIEEARKNSGNAFKIDYEEAL  
NHPEHSNNKALIKIIQTIPDIIQAIQSHLGHNDSQALIYHNPFSLSQLYTILETKRDGFHK  
NCVAVTCENYWRSQKTEIDPEISYASRLPADSVRPFDGVLARMMQRLAYEIAMAKWE  
QIKHIPDNSSLLIPIYLEQNRFEFEESFKKIKGSSSDKTLEQAIEKQNIQWEEKFQRIINA  
SMNICPYKGASIGGQGEIDHIYPRSLSKKHFGVIFNSEVNLIYCSSLQGNREKKEEHYLL  
EHLSPLYLKHQFGTDNVSDIKNFISQNVANIKKYISFHLLTPEQQKAARHALFLDYDDE  
AFKTITKFLMSQQKARVNGTQKFLGKQIMEFLSTLADSKQLQLEFSIKQITAEVHDHR  
ELLSKQEPKLVKSRQQSFPSHAIDATLTMSIGLKEFPQFSQELDNSWFINHLMPEVH  
LNPVRSKEKYNKPNISSTPLFKDSLyaERFIPVWVKGETFAIGFSEKDLFEIKPSNKEK  
LFTLLKTYSTKNPGESLQELQAKSKAKWLYFPINKTLALEFLHHYFHKEIVTPDDTTVC  
HFINSLRYYYTKKESITVKILKEPMPLSVKFESSKKNVLGSFKHTIALPATKDWERLFNH  
PNFLALKANPAPNPKEFNEFIRKYFLSDNNPNSDIPNNGHNIKPKQKHAVRKVFSPLPVI

PGNAGTMMRIRRKDNKGQPLYQLQTIDDTSPSMGIQINEDRLVKQEVLMDAYKTRNLS  
TIDGINNSEGGAYATFDNWLTLPVSTFKPEIHKLEMKPHSKTRRYIRITQSLADFIKTIDE  
ALMIKPSDSIDDPLNMPNEIVCKNKLFGNELKPRDGKMKIVSTGKIVTYEFESDSTPQ  
WIQTLYVTQLKKQP

>II-B\_WP\_135767763.1\_Leptospira\_fletcheri

MILKSPIAIDLGAKNTGVVYTTYPPHSPKQGGIHGATLILDNLKLSQVDRRAKRHQRRNY  
SRMRLARRLVYLIVRSHYKVPENDLTPEVKVELQSLLTRRGYTFQSVVTESNEHGQT  
AWAEIEDYFPKMQIQDKSLPIQTILDSFFSSENIRLEYLEPIKEAETKDDKNHLKMIREQ  
KEAFEAEQNGARHRIRYFEEIEEDLPRKIKLLECLKKNKIDEKRFLNLIQNIISNLQIKPL  
RDYFNDKNMKEKEIWDPKKLHSSIAGWIRVWHADSKDAEKRNIKKLLSLENGRLK  
AIRGNSTYSTLKHIINFLEEVDPKDTIPPYEDQNNRKPCKQNLWLDPGILDYFPGW  
SGSLQKLLAKEEDSPFEQLVTKDFQINETTEEGKSLTLIKLYSRAHSDKELDVRAILQ  
RFLDRSANLDPWSIRLQRKASKEIKITGSLTKQELKKASDAYERMKETIGETDAKQFLD  
LSGHYYREVEDIKKGWFWSPDSDDKHKEIPKSIVHEENDEVSLFYKCDTKTGSKQKN  
LSSLVGNVLNVRFSAELFRSFLTFTWTEKFDGRSSLNSLAKSIEEARLSTDEFGVEYSIL  
LKKIEKGEIDRKSDIFKAYQNTTKSSGRLEQILKEELGFQAISKSIFYDNPFSAIQLYNIIS  
VDRSGFSKVCRCQEEHWRSTLIDTDRGLRAWASKQVSDTGRPFDGQIAMLLDRI  
ALEVVRLKKAIEQYLSNPSKIESVEIPLLEENSFSFRYELAELKKLAKKKREQLGKS  
VEEFDKRLNTKQDRIKEASREICPYTGKRISKGEIDHILPRSQTQKQFGTVFNAEVNLI  
YVSREGNQAKGKRDYSLDDLDEKYLKKIFGTKDLTFIKKTISETVDRLVSSRKNLSILVF  
DRLSMEEQDCLRHALFEKPLRPKVFEELSGQLKARVNGTQLYLAKKIRNLLRKDQDLP  
KIKFPQFSFFEPRGIDLPYMRRELAQKFPNLEKQNPQPTGSHIVDAAMVMAYAVETD  
SGLSYSGSKSGSERFDSELLSKYLPNDIRFLRIKSKEKTNSRYPHKKPLFKDGVFGER  
FLPILVLENGRLRGYTLNLSLELSPQDSKKFFEVLRVLLFKRKPVSHDYPALSSFAKK  
DKRKYIVLPVNKAKALEVVSQKEHSLLASILKSISYTVFRTPVVSFLFDQKKELIPKIRGK  
KVDWGEKIEKEFRIKIEVKLGDVKITSGFIQYPGKSEWEKLLLSDSLGAIKNKSLLTK  
WFDSRENLSKLDGFFKDTSEPKAKHKRRSNHYALSVPKSPSGRFRIRRNSFDDYKIY  
QLYSVEGNAFNAFEINTETKENPILLDVYKNSNSVTYADKDYIYPTSGETIAFSEYREIV  
FKPEERKNLLSTGIISLHIAPGTADRRYIRIVISKDFFSKIFKVKDSFWELALNESIPKENI  
EKLSLILDKIPSEKRPGMPSDAPKLNVTKIAVNSVELSYTSPSWKSWLSSYNDIS

>II-B\_WP\_087108494.1\_Parendozoaicomonas\_haliclonae

MRYISPISIDFGAKYTGIIHNMHYLEGEDPRGSVQNTGCVLVMDDDNLQLSQVGRTQK  
RHQSRNYKRQDLARRLLIVILKDCFDQDVWQWPVKNREFLFGLINRRGFNRFDDDSL  
IEIEFCDLSVFESVLPDIAGYEGLEEWLQALAENAAECRLALQRLGRTREFLKTLPFIW  
KEEKKSIETHKHIIHFLETSINASEEGHLHRRKYLRNIAADIILEQHGRLASLFQTNSGI  
TPDQFANLVGHISNLQLRVLRRYFNDQNAKPQNGQTDWREDHFSKQFWRWIKSW  
HCDEQTKSRWELIKFKKKGGSVFKMLMLTVQPELTIPPYEDQNNRRPPKDKTLFLNP  
VALNKAYPLWREIVKRLVAAHPELDSHASTIAHMDAALDQDCILLQRMMLDRTSGLDL  
YKIRALAMGRPDRGNQLKVASENLRLRLNQHAEPFIQSIARPYYLQINEARQGLWVEE  
DSLLLCLNRTQPQRKARQLPTLLGRIVGDLGSHGQEHQSLKQQWSQRGVIKGARTPK  
GWCKFAAEQAQKDFGNGLKETIRSYRQLQARIKAGDTEEKLLPEAKKLLALEAALPDVA  
RFIAQGLGLAAENSARFANPFSLAQIYNLLESDPSGFSHVSKPSAHENLWRMQYEQS  
VDGKDCARASVLSADSIRPFDGMLSRILEKQAYSIAKRKLKQIESHEMTSAKVRVPILIE  
ENRFQFAEDLQTIKKNTRKKKDAQKSCESDFIRFEDKEQRICKASKGMCAYTGEDLG

RYGEFDHIIPRSYTQKRTGAIFNSEANLIFVSSRGNQQKRDQLYTLAELDDRYLTSQF  
 GSTNRQQIAQAVQDAVTPFLTKEGELIGFHNGLQEQRMIHALFVPGLRDQVLQLLA  
 GQRKAKVNGTQAWLAKRIIHHIRQGASATIEFDIHKVPTDSVRFRCRGLASRFPAMAK  
 QSPQPVASHAIDAALVSGSLLMEPAIARSLAPVDTTTQAEWLKGLLPETIDIQILRKK  
 PLYRKDAPASSPLFKAGLYGERFLPLLVMNDELRVGFTLPNSAKVLDHADNLYALLRP  
 VLVCKDSHELPSTLAELQESLALGQRQILNINRQKAFAYLHEFGQSPTHNDNCLAIEAL  
 LYCVQKKKIDAHYINKDKKTFKKDEDILADKFFIFETHKNNRPIFPRTVARIEKKDLTLPF  
 KQAWVELCDTDELKNFLGTKQTLPNFWRNFGSSLFPSIHNEHDHGHKKTRKVYSLPL  
 KISEGSYRVRRHTPEGASVWQRMVSGDFGAAGFAKIDGKTQWNKVVEIPHFAKSKN  
 LHSLNRSLEKNPQSITYFDEWRS�CALADALPNGVIKVACAPGTVGRPYIRFEMLESQ  
 FIKKINAASGSKYKSSWELPMEIPNVQLAGFDIFSCPLLKPRTETKKPNKHKARFLRL  
 GKTVLIEYFASSKANKALQQAYEQGEPECI

>II-B\_WP\_008864843.1\_Parasutterella\_excrementihominis

MGKTHIIGVGLDLGGTYTGTFITSHPSDEAEHRDHSSAFTVVNSEKLSFSSKSRTAVR  
 HRVRSYKGFDLRRRLLLLVAEYQLLQKKQTLAPEERENLRIALSGYLKRRGYARTEAE  
 TDTSVLES�DPSVFSSAPSFTNFFNDSEPLNIQWEAIANS PETTKALNKELSGQKEAD  
 FKKYIKTSFPEYSAKEILANYVEGRRAILDASKYIANLQSLGHKHSKYLSLILQDMKR  
 DSRITRLSEAFGSTDNLWRIIGNISNLQERAVRWYFNDAKFEQGGQEQLDAVKLKNVLV  
 RALKYLRSDDKIEWSASQKQIIQSLEQSGDVLDVLAGLDPDRITIPPYEDQNNRRPPED  
 QTLYLNPKALSSEYGEKWKSWANKFAGAYPLLTEDLTEILKNTDRKSRIKIRSDVLPD  
 SDYRLAYILQRAFDRSIALDECSIRRTAEDFENGVIKNEKLEDVLSGHQLEEFLEFAN  
 RYYQETAKAKNGLWFPENALLERADLHPPMKNKILNVIVGQALGVSPAEGTDFIEEIW  
 NSKVKGSTVRSICNAIENERKTYGPYFSEDYKFVK TALKEGKTEKELSKKFAAVIKVL  
 KMVSEVVPFIGKELRLSDEAQSKFDNLYSLAQLYNLIETERNGFSKVSLAAHLENAWR  
 MTMTDGSQAQCCRLPADCVRPFDGFIRKAIDRNSWEVAKRIAEEVKKSVDFTNGTVKI  
 PVAIEANSFNFTASLTDLKYIQLKEQKLKKKLEDIQRNEENQEKRWLSKEERIRADSHG  
 ICAYTGRPLDDVGEIDHIIPRSLTLKKSESIYNSEVNLI FVSAQGNQEKKNNIYLLSNLAK  
 NYLAAVFGTSDLSQITNEIESTVLQLKAAGRLGYFDLLSEKERACARHALFLNSDSEA  
 RRAVIDVLGSRRKASVNGTQAWFVRSIFSKVRQALAAWTQETGNELIFDAISVPAADS  
 SEMRKRFAEYRPEFRKPKVQPVASHSIDAMCIYLAACSDPFKTKRMGSQLAIEPINF  
 DNFTGSCQVIQNTPRNFSKTNIANSPIFKETIYAERFLDIIVSRGEIFIGYPSNMPFEE  
 KPNRISIGGKDPFSILSVLGAYLDKAPSSEKEKLTIRVVKNKAFELFSKVAGSKFTAEE  
 DKA AKILEALHFVTVKQDVAATVSDLIKSKKELSKDSIENLAKQKGCLKKVEYSSKEFK  
 FKGSLIIPAAVEWGKVLWNVFKENTAEELKDENALRKALEAAWPSSFGTRNLHSAK  
 RVFSLPVVATQSGAVRIRKTAFGDFVYQSQDTNNLYSSFPVKNGKLDWSSPIIHPAL  
 QNRNLTAYGYRFVDHRSISMSEFREVYNKDDLMRIELAAGTSSRRYLRVEMPGEK  
 FLAWFGENSISLGSSFKFSVSEVFDNKIYTENA EFTKFLPKPREDNKHNGTIFFELVGP  
 RVIFNYIVGGAASSLKEIFSEAGKERS

>II-B\_WP\_143066515.1\_Ruminobacter\_amylophilus

MSNTYYSNIAIDMGGKYTGFI SYTSNDLPDSENVKAAIEMPDDGGSVINYTVKNRTAVR  
 HHLRANNRYKKARKLIFTILADTIKRNLTQSEQEAISSMLRRGYTRLETEIDLEQLKEC  
 PVDLFFQCCPGLFTGDVPLYDQFINNCNDDTVVYVRESLSSIKETVSNLSDKTEKSMY  
 GSALKTMVDFVDKFIDQKEFGHKHRTEYLANIRKDIQKDSRLNDLIKISGEDRLYNCI  
 GNISNLQLRALRWYFNDRGMSGKAVFDKKRFKDVWYRAYKYFHYTAGSNQKDNAE

KVDIIKKKNEFLKRIEQTENITDLLLAQDPVFTIPPYEDQNNRRPPVDQTLTLLSPQSLDA  
HYPGKWELWVQKFM DYSGSYSSL SDELEQIVKL TDRSSRLQSAGGKQSDHIAS YT  
EQKLIHSYYLQRLLDITYNRGNPESQIRKWAQNPNSCSATNAVIKNILGESDIDSSFFEL  
ANNYYREVSLSKKGLWSIVEHPLLEISGIHPPMKSKMLEELVAGVLNIRKGFNLERFKE  
VWNSTVKGKSTVRSLCSYIEKTRKKYGNCFKQEYKNVCRYEELKNKASEVHFSSDIKI  
DKEVVKVYEYTKNLSKFIGEKLELSTDQQSKFANPFSLAQLYTILETDVHGFSGNCKA  
VCLENTCRMQMSDKGNALCCRLPAESTRPF DGSLGNILDRQAYEIAKIKIAEIKELSAL  
KNTKVKLSVLVEENSFEFSSSLSTIKKSLKAKQLREFALKAEKTEREKWKSKDERLKN  
DSHNICAYTGMALSDDLSENDHIVSRSETKESYGTIYNSEFNLRVSRRG NQLKGNNI  
YTLDNLNPKYLEAVFGTSDTAITTRIRETLNRLLISNP NLDIMTEEERQCCRHALFTS  
RTGVPFNQIIQAISKQYRARVNGTQAWFVKNLIAKIDEGLEWKSINNITVEYNAYKIDA  
GSTSEIRHDIGKLDARYEKKDIQPITSHIDA VCLLGNASTQQEISEEISGNNEISVLSD  
VKELIKLIPDRFDIIRISSLNFAEKNNPESKQLFKDTIYAEHFLCIMEKDDTVKVGDFDGS  
NSIQIVKGGKQLLSCLANYLESPKQNETTGFR TYKVCKTAAFKL FERFYRGLVAEGSD  
EELAYLALMSLRYTTLHQDVFSLLYDVKSKKFAAKKDILKNLDIKISFPSVLKLTDKAS  
LVLP SFNEWAHIASIFEDYLGGKDEQESGYDKLKQYLG TILNKTINKNHAKSKRVYSLP  
VVASPSGGVRIKRLTHTHGDIYQLVAANTPVTSISKGFASADNGDVLWDEAVMVDSY  
TTKNLTVINKKISGDKNFVSMDEQRLVYESPDERVYMT PATDIRRGITIEQKFEKFVYCI  
DQQNIYKSFHDLPT EIKTDPKIFMERLNIDICGKPRGNL KILSIGNIIKYSYMVESSNRVM  
NQVYNNPIDKGTIFKSN

>II-B\_WP\_215641693.1\_Sutterella\_wadsworthensis

MTQSERRFSCSIGIDMGAKYTG VFYALFDREELPTNLNSKAMTLVMPETGPRYVQAQ  
RTAVRHRLRGQKRYTLARKLAFLVVD DMIKKQEKRLTDEEWKRGREALSGLLKRRG  
YSRPNADGEDLTPLENVRADVFAAHPAFSTYFSEVRSLAEQWEEFTANISNVEKFLG  
DSNIPADKEFVDFAVAEG LIDKTEKKAYLSALSTMRANANVLTGLRQMGHKPRSEYFK  
AIEADLKKDSRLAKISEAFGGADRLARLLGNLSNLQLRAERWYFNAPDIMKDRGWEP  
DRFKKTLVRAFKFFHPAKDQNKQHLELIKQIENSEDIIETLCTLDPNRTIPPYEDQNNR  
RPPLDQTLTLLSPEKLTRQYGEIWKTWSARL TSDEPTLAPAAEILERSTD RKSRVAVNG  
HEPQPTLAYQLSYALQRAFDRSKALDPYALRALAAGSKSNKLTSARTALENCIGGQN  
VKTFLDCARRYREADDAKVGLWFDNADGLLERSDLHPPMKKKILPLL VANILQTDET  
TGQKFLDEIWRKQIKGRETVASRCARIETVRKSFGGGFNIAYN TAQYREV NKLPRNA  
QDKELLTIRDRVAETADFIAANLGLSDEQKRKFANPFSLAQFYTLIETEVS GFSATTLA  
VHLENAWRMTIKDAVINGETVRAAQCSRLPAETARPFDGLVRR LVDRQAW EIAKRVS  
TDIQSKVDFSNGIVDVSIFVEENKFEFSASVADLKKNKRVKDKMLSEAEKLETRWLIKN  
ERIKKASRETCPYTGDRLAEGGEIDHILPRSLIKDARGIVFNAEPNLIYASSRGNQLKK  
NQRYSLSDLKANYRNEIFKTSNIAAITAEIEDVVT KLQQTHRLKFFDLLNEHEQDCVRH  
ALFLDDGSEARDAVLELLATQRRTRVNGTQIWMIKNLANKIREELQNWCKTTNNRLH  
FQAAATNVSDAKNLR LKLAQNQPDFEKPDIQPIASHSIDALCSFAVG SADAERDQNGF  
DYLDGKTVLGLYPQSCEVIHLQAKPQEEKSHFDSVAIFKEGIYAEQFLPIFTLNEKIWIG  
YETLNAKGERCGAIEVSGKQPKELLEMLAPFFNKPVGDL SAHATYRILKKPAYEFLAK  
AALQPLSAEEKRLAALLDALRYCTSRKSLMSLFMAANGKSLKKREDVLKPKLFQLKVE  
LKGEKSFKLNGSLTLPVKQDWL RICDSPELADAFGKPCSADELTSKLARIWKRPVMR  
DLAHAPVRREFSLPAIDNPSSGGFRIRRTNLF GNELYQVHAINAKKYRGFASAGSNVD  
WSKGILFNELQHENLTECGGRFITSADVTPMSEWRKVVEDNLSIWIAPGTEGRRYV

RVETTFIQASHWFEQSVENWAITSPLSLPASFKVDKPAEFQKAVGTELSELLGQPRD  
KIFIESIVNTKRIRFWYTVKNSNKKMNESYNNVSKS

>II-B\_WP\_158222317.1\_Tamilnaduibacter\_salinus

MTKLVSPIAVDQGARYTGIYSLSYEPGVDPFLESKNAEVVVDNDKLTLSMAPRRAG  
RHLRRGNQRFKLARRLTLVLREAFGVDPDQLDQREREFLLGLLKRRGFTFLGEEID  
EEAISGLDPDLVKRLLRNCPPFESLESLSWQSLSSDPGMCREFLQLSTGSDDSLDLETL  
DKRWLKKWLKDNVEEASDKETIERLTSGVLALRDGCRAVIRAKWDGHHPRHIYLENI  
AQDLKHQPEGNALADRIGLSTGELSHLIGNISNLPLRPLRRYFNDKRFSRNADRWYPE  
RMARQLTREWKGWQPDGVSKDIVQRSRELLKGLSNAAQDPVRFWKYTDPLQTIPPY  
ERHSNRRPPKQCQSLLDVLTALGERWPEWERWIAPLHKANPELSEGIDGFSPSISGPN  
EPRLAFLQRVLDIRSTLFDPYGLRQLAVESRSHRAAQGKETLFRDLGDQHARGFLSFV  
TAYHREREQARSGFWDNQREGRLLRCHTNPPHKGKILHLLLNRLLDGDIEDVEGFR  
DEITKTKIGRSTLKGIAQRASDVQKRYGNALNATLKEADLDKATYKRLNDLQALIEKAG  
PEIGRIAGCPENRRNRLGNVFILAQLFNLLFGDPSGFSSTCRSCSEENAWRSRSVPV  
PGEDREGWRMIPLTADSVRPFDPGVVRRRAVDAQARAMARLKKQELKKRDLPAQSFSVS  
IPVLEQNRFSFSEDLAAIKQRIAPSARHRNRQKAAREAEQQISRWRGKLDRIKRDAS  
HGICPYSGRDLNQVNGQIDHILPRAQSRKLDGTTWNHEANLIYCSTSANVDKGETLK  
NLEDLDPGYLEAVFPGQSLTSIRQWIADRVKPLLARSGRMPSFVELERDQSIALRHAL  
FDEDLRPQVMELLRGQARATRVNGTQAWLARRFVEELRRSVPELTLSFSVNRIQPED  
VGLMRESLSEIRPEFLKAPDQPLSSHAIDAMVYAAASSLPRMESELHVPAEPLAELH  
DRLARLLPERIDFTDVSARKPVYSLPRADSQTLFKDTYYAERFLPLLVSDDGALRAGFT  
PENSVEVKPARKRSESTPAEIAISRDLGFLRLRGARWSGDLSSLFASDASSDQWVAL  
SVDRPLALARLQQVGRGEDELTDTDALLGALKYLTNKKELRTAVVGTGSALMSWEKL  
NEEIEKKTTLSSVSIKGAGLKS DGAKLSLPARSHWLRFAARLNEEGVLGSNPGWSDAE  
WRRFFQTVFPSAKRDRQHQQVRKVWVSLPVVPGSPGLVRHQRRTRSGQTVFQVAS  
LHEFKHMGVVLRRERKAEGVFPGYLRWSTNLAPLAGVLQSPSQRMIRFDQWYPLDRP  
SVCQSADILSVSAAWGSGLRRSVEVTVSQELACRVFDVNSAWNLPANLKLHNNASE  
LWPIEAIGWPRNNKVELRAITPSGVVRLRYQKDKGGKPIQQEELERAIRREE

>II-B\_WP\_066592711.1\_Turicimonas\_muris

MESSKYIFPVGIDMGAKYTGVMFLSHLDGVLPTAADATAMTIVNSDKIHYSMKHRTAV  
RHRVRSKKRFLVARLLNLIVDHQLMKAGFVLSAQEKSQLTEALSGYLKRRGYSRIET  
EIDLDSLNDVDPIVFAMCPGFEDFFSSDIELLTQWNSLTADLSEVNKINSNKKIPDVKTF  
KKYVKDNFPDYSSKESLERYLEARKTIEAARQISELTSLGHLHREKYLDNILHDMGRD  
SRLKKIGEAFGGICRFWNLIGNLSNLQLRALRWYFNDPWIEKVKPVETENSQPLNEQK  
LKDIWVRAYKYFHPDPENKARISDVIRALSKADDIIDCLCELDCKSTIPPYEDQNNRRP  
PLDYTLYLNPALKTEVYGDSWRSWASRLVNADSDLQDNLDVILQNRDRKSRRTTQD  
SLDFRLSYILQRALDRSKVNDIFEIRKILSSPTLYSEAATRNLFFFKTLEEQNFPGFLDFA  
RRYFYEVAKAQKGLWNEKESILERSDLHPKQKKKVLDRLVANVLRVDEETAKRFIADV  
WNGKVKGNSTVKSICTRIEQLRKDRGSFAFNLEYQSALTKAALGNEKLDSEEKEYVKL  
KQDVSLTAAFIGEQLGLSDEQRTKFANPFLAQLYTIETERSGFTKITTAALLENAWRL  
TFKNGAAQCGPLPADSVRPFDPGYMRRVLERNAREVGKRLIEKIKRDTQDQGAIEVPI  
LVEENKFNFASLLELKRLELSDSLKKKQEKIEKAGEAQKKKWLSKEERIKDDAKGIC  
AYTGAPLNSEGEVDHIIPRSKTIKNYSSIFNSEVNLIFVSRQGNAACKENLYFLKQLNSA  
YLTAVFGTSDFSTIESKIESTVQQLIENKRLGYLDLLSQEERNVCVRHALFLNPGSDARK

EVLSVLGARRKTVVNGSQAWFIRTLRSYLENGLQNWLTETGNKLSFDAFSIPAVTVSS  
LRSQLGELNSMFKEDIQSVSSHSDAMCVYAAACGDASAECVLHGDALLADSHSM  
RAAQLLKLYPETCQVVQISPKSFSQKDRINNSAIFKDTIYAERFLNIYKKKEEVFVGFD  
NDKKGSNEKESLTQVKGKNPHLLLQILEPFFEKRNESKDEGTYRIDKTKAFQLMGKA  
GCALLSNDESRLSLLNQLHYTTLLKQEVLTTELVDTASKIKLKSKYQSNFKREFLLKLE  
LRDRERQFSAKGNIVYPSYKRWQELSEELSNFLTACPEGGLEEVNARLSSLWTTSEK  
SSVRHKKVKRSMPLASQSGAVRIRRKDFRDGYVYQSIDANQLYYGFSKKDGKIL  
FNSPLPHPRLTTKNLKIYEYQYQPVNECVKMDEWKTVEKEGVIVEHAPGTKARRYIR  
ATLPFDLYLLWDKELQKAKDKSSLTLP SKVIRETKDFFEEGIYKNEELAQQFFPKPRENN  
KKFTMNIELVGSTVKFSYVTA

>II-B\_WP\_011139431.1\_Wolinella\_succinogenes

MLVSPISVDLGGKNTGFFSFTDSLNSQSGTVIYDESFVLSQVGRRSKRHSKRNNLR  
NKLVKRFLLLILQEHHGLSIDVLPDEIRGLFNKRGYTYAGFELDEKKKDALESDDLKEFL  
SEKLQSIDRDSVEDFLNQIASNAESFKDYKKGF EAVFASATHSPNKKLELKDELKSE  
YGENAKELLAGLRVTKEILDEFDKQENQGNLPRAKYFEELGEYIATNEKVKSFFDSNS  
LKLTDMTKLIGNISNYQLKELRRYFNDKEMEKGDIWIPNKLHKITERFVRSWHPKND  
DRQRRAEMLMKDLKSKEIMELLTTTEPVMTIPPYDDMNNRGAVKQCQTLRLNEEYLDKH  
LPNWRDIAKRLNHGKFNDLADSTVKGYS EDDLHRLLDTSKEIDIYELRGKKPNELL  
VKTLGQSDANRLYGFAQNYIELIRQKVRAGIWWPVKNKDDSLNLEDNSNMLKRCNH  
NPPHKKNQIHNLVAGILGVKLDEAKFAEF EKELWSAKVGNKKLSAYCKNIEELRKTHG  
NTFKIDIEELRKKDPAELSKEEKAKLRLTDDVILNEWSQKIANFFDIDDKHRQRFNLF  
SMAQLHTVIDTPRSGFSSTCKRCTAENRFRSETAFYNDETGEFHKKATATCQRLPAD  
TQRPFSGKIERYIDKLGYELAKIKAKELEGMEAKEIKVPIILEQNAFEYEE SLRKS KTGS  
NDRVINSKKDRDGKKLAKAKENAEDRLKDKDKRIKAFSSGICPYCGDTIGDDGEIDHIL  
PRSHTLKIYGT VFNPEGNLIYVHQKCNQAKADSIYKLS DIKAGVSAQWIEEQVANIKGY  
KTFSVL SAEQQKAFRYALFLQNDNEAYKKVVDWLRTDQSARVNGTQKYLAKKIQEKL  
TKMLPNKHL SFEFILADATEVSELRRQYARQNPLLAKEKQAPSSHAIDAVMAFVARY  
QKVFKDGT PPNADEVAKLAMLD SWNPASNEPLTKGLSTNQKIEKMIKSGDYGQKNM  
REVFGKSIFGENAIGERYKPIVVQEGGYI GYPATVKKGYELKNCKVVT SKNDIAKLEK  
IKNQDLISLKENQYIKIFSINKQTISELSNRYFNMNYKNLVERDKEIVGLLEFIVENC RYY  
TKKVDVKFAPKYIHETKYPFYDDWRRFDEAWRYLQENQNKTS SKDRFVIDKSSLNEY  
YQPDKNEYKLDVDTQPIWDDFCRWYFLDRYKTANDKKSIRIKARKTFSLLAESGVQG  
KVFRARKKIPTGYAYQALPMDNNVIAGDYANILLEANSKTL SLVPKSGISIEKQLDKKLD  
VIKKT DVRGLAIDNNSFFNADFDTHGIRLIVENTS VKVGNFPISAIKSAKRMIFRALFE  
KEKGKRKKKTTSFKESGPVQDYLKVFLKKIVKIQLRTDGSISNIVVRKNAADFTLSFRS  
EHIQKLLK

>II-C\_WP\_002864485.1\_Campylobacter\_jejuni\_CjeCas9

MARILAFDIGISSIGWAFSENDELKDCGVRIFTKVENPKTGESLALPRRLARSARKRLA  
RRKARLNHLKHLIANEFKLN YEDYQSFDESLAKAYKGLISPYELRFRALNELL SKQDF  
ARVILHIAKRRGYDDIKNSDDKEKGAILKAIQNEEKLAN YQSVGEYLYKEYFQKFEN  
SKEFTNVRNKKESYERCIAQSFLKDELKLIFKKQREFGFSFSKKFEEEVLSVAFYKRAL  
KDFSHLVGNCSFFTDEKRAPKNSPLAFMFVALTRIINLLNNLKNTEGILYTKDDL NALL  
NEVLKNGTLTYKQTKKLLGLSDDYEFKGEKGT YFIEFKKYKEFIKALGEHNLSQDDL N  
EIAKDITLIKDEIKLKKALAKYDLNQNQIDSLSKLEFKDHLNISFKALKLV TPLMLEGKKY

DEACNELNLKVAINEDKKDFLPAFNETYYYKDEVTNPVVLRAIKEYRKVLNALLKKYGV  
HKINIELAREVGKNHSQRAKIEKEQNENYKAKKDAELECEKLGKINSKNILKLRLFKE  
QKEFCAYSGEKIKISDLQDEKMLEIDHIYPYSRSFDDSYMKNKVLVFTKQNQEKLNQTP  
FEAFGNDSAKWQKIEVLAKNLPKKQKRILDKNYKDKEQKNFKDRNLNDTRYIARLVL  
NYTKDYLDLPLSDDENTKLNDTQKGSKVHVEAKSGMLTSALRHTWGFSKDRNNH  
LHHAIDAVIIAYANNSIVKAFSDFKKEQESNSAELYAKKISELDYKNKRKFFEPFSGFRQ  
KVLDKIDEIFVSKPERKKPSGALHEETFRKEEEFYQSYGGKEGVLKALELGKIRKVNG  
KIVKNGDMFRVDIFKHKKTNKFYAVPIYTMDFALKVLPNKAVARSKKGEIKDWILMDEN  
YEFCSLYKDSLILIQTKDMQEPEFVYYNAFTSSTVSLIVSKHDNKFETLSKNQKILFKN  
ANEKEVIAKSIGIQNLKVFEKYIVSALGEVTKAEFRQREDFKK

>II-C\_WP\_033016936.1\_Geobacillus\_stearothermophilus\_GoeCas9

MRYKIGLDIGITSVGWAVMNLDIPRIEDLGVRIFDRAENPQTGESLALPRRLARSARRR  
LRRRKHRLERIRRLVIREGILTKEELDKLFEEKHEIDVWQLRVEALDRKLNNDELARVL  
LHLAKRRGFKSNRKSERSNKENSTMLKHIEENRAILSSYRTVGEMIVKDPKFALHKRN  
KGENYNTIARDDLEREIRLIFSKQREFGNMSCTEEFENEYITIWASQRPVASKDDIEK  
KVGFCFTFEPKEKRAPKATYTFQSFIWEHINKLRLISPSGARGLTDEERRLLYEQAFQ  
KNKITYHDIRTLLHLPDDTYFKGIVYDRGESRKQENENIRFLELDAYHQIRKAVDKVYGK  
GKSSSFLPIDFDTFGYALTFLKDDADIHSYLRNEYEQNGKRMPNLANKVYDNELIEELL  
NLSFTKFGHLSLKALRSILPYMEQGEVYSSACERAGYTFTGPKKKQKTMLLPNIPPIAN  
PVVMRALTQARKVVNAIIKKYGSPVSIHIELARDLSQTFDERRKTKKEQDENRKKNET  
AIRQLMEYGLTLNPTGHDIVKFKLWSEQNGRCAYSLQPIEIERLLEPGYVEVDHVIPYS  
RSLDDSYTNKVLVLTRENREKGNRIPAEYLGVGTERWQQFETFVLTNKQFSKKKRDR  
LLRLHYDENEETEFKNRNLNDTRYISRFFANFIREHLKFAESDDKQKVYTVNGRVTAH  
LRSRWEFNKNREESDLHHAVDVAVACTTPSDIAKVTAFYQRREQNKELAKKTEPHF  
PQPWPHFADELRLARLSKHPKESIKALNLGNYDDQKLESLQPVFVSRMPKRSVTGAA  
HQETLRRYVGIDERSGKIQTVVKTKLSEIKLDASGHFPMYGKESDPRTYEAIRQRLL  
HNNDPKKAFQEPLYKPKKNGEPGPVIRTVKIIDTKNQVIPLNDGKTVAYNSNIVRVDF  
EKDGKYYCVPVYTMDIMKGILPNKAIEPNKPYPYSEWKEMTEDYTFRFSLYPNDLIRIELP  
REKTVKTAAGEEINVKDVVYYKTIDSANGGLELISHDHRFSLRGVGSRTLKRFEKYQ  
VDVLGNIYKVRGEKRVGLASSAHSKPGKTIRPLQSTRD

>II-C\_WP\_003343632.1\_Brevibacillus\_laterosporus\_BlatCas9

MAYTMGIDVGIASCGWAIVDLERQRIIDIGVRTFEKAENPKNGEALAVPRREARSSRR  
RLRRKKHRIERLKYMFVRNGLAVDIQHLEQTLRSQNEIDVWQLRVDGLDRMLTQKEW  
LRVLIHLAQRRGFQSNRKTGSSSEDGQVLNVNTENDRLMEEKDYRTVAEMMVKDEK  
FSDHKRNKNGNYHGVSRSALLVEIHTLFESQRQHNFASKDFELEYINIWSSQRPV  
ATKDQIEKMIGTCTFLPKEKRAPKASWHFQYFMLLQTINHIRITNAQGTRPLNKEEIEQ  
VVMALTKSKVSYHDIRKVLDLSEEYQFVGLDYGKEDEKKKVESKETIILDDYHKLN  
KIFNEVELAKGETWEAVDYDTVAYALTFFKDDDEDIRDYLNKYKDSKNRLVKNVANKE  
YTNELIEKVSTLSFRKVGHLNLSLKALRKIIPFLEQGMYTDKACQAAGFNFGGSKKKRSV  
VLPVIDQISNPVVRNALTQTRKVINALIKKYGSPETIHETARELSKRFDERKNITKDYKE  
NRDKNERAKKHLSELGIINPTGLDIVKYKLWCEQQGRCMYSNQPISFERLRESGYTEV  
DHIIPYSRSMNDSYNNRVLVMTRENREKSNQTPYEYMGNDTQRWDEFEQRVTTNP  
QIKKEKRQNLKLGFTNRRELEMLERNLNDTRHITKYLSHFISTNLEFSPSDKKKKVVN  
TSGRITSHLRSRWGLEKNREQNDLHAMDAVVIAVTSDFSIIQVVTNYYKRKERRELN

GDDKFPLPWKFFFREEVIARLSPNPKEQIEALPNHFYSEDELADLQPIFVSRMPKRSITG  
EAHQAQFRRVVGKTKEGKNITAKKTALVDISYDKNGDFNMYGRETPATYEAIKERYL  
EFGGNAKKAFFSTDLHKPKKDGTGKPLIKSVRIMENKTLVHPVNKGKGVVYNSSIVRTD  
VFQRKEKYLLPVYVTDVTKGKLPNKVIAAKKGYDDWIEVDDSFTFLYSLYPNDLIFIR  
QNPKKKISLKKRIESHSISDSKEVQEILTYKGVDSSTAIEFITHDGSYYAKGVGVQNL  
DCFEKYQVDILGNYSKVKGEKRLELETSDDNNHKGKDVNSIKSTSR

>II-C\_WP\_018356570.1\_Pasteurella\_pneumotropica\_PpnCas9

MQNNPLNYILGLDLGIASIGWAVVEIDEESSPIRLIDVGVRTFERAEVAKTGESLALS  
RLARSSRRRIKRRARERLKKAKRLLKAEKILHSIDEKLPINWQLRVKGLKEKLERQEWA  
AVLLHLSKHRGYLSQRKNEGKSDNKELGALLSGIASNHQMLQSSEYRTPAEIAVKKF  
QVEEGHIRNQRGSYTHTFSRDLAEMELLFQRQAELGNSYTSTTLLENLTALLMWQ  
KPALAGDAILKMLGKCTFEPSEYKAAKNSYSAERFVWLTCLNNLRILENGTERALNDN  
ERFALLEQPYEKSCLTYAQVRAMLALSDNAIFKGVRYLGEDKKTVESKTTLIEMKFIH  
QIRKTLGSAELKKEWNEKGNLDEIGTAFSLYKTDDDICRYLEGKLPERVLNALLE  
NLNFDKFIQLSLKALHQLPLMLQGQRYDEAVSAIYGDHYGKKSTETTRLLPTIPADEIR  
NPVVLRTLTQARKVINAVVRLYGSPARIHIETAREVGKSYQDRKKLEKQKQEDNRKQR  
ESAVKKFKEMFPHFVGEPPKGDILKMRLYELQQAKCLYSGKSLELHRLLEKGYVEVD  
HALPFSRTWDDSFNNKVLVLANENQNKGNLTPYEWLDGKNNSERWQHFVVRVQTS  
GFSYAKKQRILNHLKDEKGFIERNLNDTRYVARFLCNFIADNMLLVGKGKRNVFASNG  
QITALLRHRWGLQKVREQNDRHHALDAVVVACSTVAMQQKITRFVRYNEGNVFSGE  
RIDRETGEIPLHFPSPWAFFKENVEIRIFSENPKLELENRLPDYPQYNHEWVQPLFVS  
RMPTRKMTGQGHMETVKSARKLNEGLSVLKVPLTQLKLSDLERMVNRDREIALYESL  
KARLEQFGNDPAKAFAPFYKKGALVKAVRLEQTQKSGVLVRDGNVADNASMVR  
VDVFTKGGKYFLVPIYTWQVAKGILPNRAATQGKDENDWDIMDEMATFQFSLCQNDL  
IKLVTKKKTIFGYFNGLNRATSNINIKEHDLDSKSGKGLGIYLEVGVKLALSLEKYQVDEL  
GKNIRPCRPTKRQHVR

>II-C\_WP\_014574210.1\_Neisseria\_meningitidis\_1\_Nme1Cas9

MAAFKPNSINYILGLDIGIASVGWAMVEIDEEENPIRLIDLGVVRFERAEVPKTGDSLA  
MARRLARSVRRLTRRRRAHRLLRTRRLKREGVLQAANFDENGLIKSLPNTPWQLRAA  
ALDRKLTPLEWSAVLLHLIKHRGYLSQRKNEGETADKELGALLKGVAGNAHALQTGD  
FRTPAELALNKFEEKESGHIRNQRSYSHTFSRKDLQAEILLLFEKQKEFGNPHVSGGL  
KEGIETLLMTQRPALSGDAVQKMLGHCTFEPAPKAAKNTYTAERFIWLTCLNNLRIL  
EQGSRPLTDTERATLMDEPYRKSCLTYAQARKLLGLEDTAFFKGLRYGKDNAEAST  
LMEKAYHAISRALEKEGLKDKKSPLNLSPELQDEIGTAFSLFKTDEDITGRLKDRIQP  
EILEALLKHISFDKFVQISLKLARRIVPLMEQGKRYDEACAEIYGDHYGKKNTEEKIYLP  
PIPADEIRNPVVLRLALSQARKVINGVVRRYGSPARIHIETAREVGKSFKDRKEIEKRQE  
ENRKDREKAAAKFREYFPNFVGEPPKSKDILKLRLYEQQHGKCLYSGKEINLGLRNEK  
GYVEIDHALPFSRTWDDSFNNKVLVLGSENQNKGNQTPYEYFNGKDNSREWQEFKA  
RVETSRFPRSKKQRILLQKFDEDEGFKERNLNDTRYVNRFLCQFVADRMRLTGKGGK  
RVFASNGQITNLLRGFWGLRKVRAENDRHHALDAVVVACSTVAMQQKITRFVRYKE  
MNAFDGKTIDKETGEVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTLEKL  
RTLLAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGQGHMETVKSARKLDEGVSVLRV  
PLTQLKLDLEKMVNREREPKLYEALKARLEAHKDDPAKAFAPFYKYDKAGNRTQQ  
VKAVRVEQVQKTGVWVRNHNGIADNATMVRVDVFEKGDYKYLVPYISWQVAKGILP



DRAVVQ GKDEEDWQLIDDSFNFKFSLHPNDLVEVITKKARMFGYFASCHRTGNINI  
RIHDL DHKIGKNGILEGIGVKTALS FQKYQIDELGKEIRPCRLKKRPPVR

>II-C\_WP\_002230835.1\_Neisseria\_meningitidis\_2\_Nme2Cas9

MAAFKPNPINYILGLDIGIASVGWAMVEIDEEENPIRLIDLGV RVFERAEV PKTGDSL A  
MARRLARSVRRLTRRRAHRLLRARLLKREGVLQAADF DENGLIKSLPNTPWQLRAA  
ALDRKLT PLEWSAVLLHLIKHRGYLSQRKNEGETADKELGALLKGVANNAHALQTGD  
FRTPAELALNKFEKESGHIRNQRGDYSHTFSRKDLQAE LILLFEKQKEFGNPHVSGGL  
KEGIETLLMTQRPALSGDAVQKMLGHCTFEPAPKAAKNTYTAERFIWLT KLNLRIL  
EQGSERPLTDTERATLMDEPYRKS KLYAQARKLLGLEDTAFFKGLRYGKDNAEAST  
LMEKAYHAISRALEKEGLKDKKSPLNLSSELQDEIGTAFSLFKTDEDITGRLKDRVQ  
PEILEALLKHISFDKFVQISLKALRRIVPLMEQGKRYDEACAEIYGDHYGKKNTEEKIYL  
PIPADEIRNPVVLRLALSQARKVINGVVRRYGSPARIHETAREVGKSFKDRKEIEKRQ  
EENRKDREKAAAKFREYFPNFVGEPKSKDILKLRLYEQQH GKCLYSGKEINLVRLNEK  
GYVEIDHALPFSRTWDDSFNNKVLVLGSENQNKGNQTPYEYFNGKDNSREWQEFKA  
RVETSRFPRSKKQRILLQKFDEDDGFKECNLNDTRYVNRFLCQFVADHILLTGKGKRR  
VFASNGQITNLLRGFWGLRKVRAENDRRHALLDAVVVACSTVAMQQKITRFVRYKEM  
NAFDGKTIDKETGKVLHQKTHFPQPWEFFAQEVMIRVFGKPDGKPEFEEADTPEKLR  
TLAEKLSSRPEAVHEYVTPLFVSRAPNRKMSGAHKDTLRS AKRFVKHNEKISVKRV  
WLTEIKLADLENMVNYKNGREIELYEALKARLEAYGGNAKQAFDPKDNPFYKKGQQL  
VKAVRVEKTQESGVLLNKKNAYTIADNGDMVRVDVFCKVDKKGKNQYFIVPIYAWQV  
AENILPDIDCKGYRIDDSYTFCSLHKYDLIAFQKDEKSKVEFAYYINCDSSNGRFYLA  
WHDKGSKEQQFRISTQNLVLIQKYQVNELGKEIRPCRLKKRPPVR

>II-C\_WP\_106659749.1\_Confluentibacter\_flavum

MKRILGLDLGTTSIGFAYIKEGENEKELSSIEKIGVRVNPLTTDEQTNFEKGKPLTTNAD  
RTLKR GARRNLD RYQLRREDLIEILKKASLISDD SILTENGKNTTFETWRLRAKAVNEKI  
EKHEFARVLLAINKKRGYKSSRRTKNEEDGQAVDGMGIAKLLYEENLTPGQLSFQILQ  
NGKTL LPDFYRSDLQAEFNKVWNFQKQFYSEIFTDEFYKELQGKGQRATSALFWITC  
KFNIAEIKDLED SLKTEKTIKFNKRQQKKLQAYKWRSDAISKQLPKEEVAYVLTEINNDL  
NKSSGYLGAISDRSKKLYFNKETVGQHLYRQLQKNQHTRLKNQVFYRQDYLDDEFEII  
WETQAKFHSELTKELKEEIRDIVIFYQRKLKSQKGLISFCEFESREIVKNGREKTVGLR  
VASKSSSMFQEFKIWQILSNLEFRNKT TKEKLSLDLEAKEILFEELNLKGNLSASKIIE LL  
GCESKDWELNYKTIEGNRTNKSLEYAYLKILEIEGYDVKDLLKV KSNKDELELDDLNPV  
ATKIKNMVNKIFDTLEISTEILDFDADLDGKDFEKQASYQLWHLLYSAEDDSKKYSEDD  
ILTYGNDNIGLKKQLCAKFGFKPEHAKILVNVD FQDDYGNLSSKAMRKIYPYIKEHQYS  
DACEQAGYRHSKYSLSKEELKKRELKPHLELLKKNLSLRNPVVEKILNQMVNVNTLID  
NEKKKDPNFKFDEIRIELARELKKNAKERAEMTTNINA AKMRHEKIVKRLQSEFGVAN  
PTRNDIIRYKLYEELKNNGYKDLYTNTYIPREKLFTKEIDIEHIIPQSKAFDDSF SNTICY  
REENLKKGNRTAIDYIQSEYGGEKLEEYLN RITALNKNWGKDKQEEGISKAKFQKLQK  
RESEIGEGFAERDLRDSQYIAKKA SMLFEITPSVVSTSGSITDR LREDWDLINVMKEL  
NLPKYRMLGLAEIQRKFGKQVEVITDWT KRNDHRHHAMDALAVAF TKRNHIQYLN Y  
LNARKNENHKLHANIIAIEKKETELVENGN GSKKRLFKKPLPNFRQVAKEHLENVLVSH  
KAKNKVVTKNKNKTKSKKGEKVKIELTPRGQLHKETVYGKYHYNSKEEKVGT KFNK  
ETIAQVSNPLYRKL LQRLEENNNNDPKKAFGGKNALVKNLIYLDKIKTEVLPEKVKLVW  
LEEDFSIRKDITPENFKDLKTIEKVLDDGVKRVLKKRLSDFGNDPKKAFSDLGKNPIWL

NEEKGISIKRVTISGVKNAEHLHYKKDHLGNFILDDDGHKIPVGYVSTGNNHHIAIYRD  
ENGELLDYVVSFSDAVKRKSYGLPIVPKDTNKLWVKVFLGIDNQSILATLPRQGLQL  
EFSMKQNECFVFPNEAMDFNPKEVDLHNPYKKLISP NFLFIVQTISKVMYGNNAVRD  
YVFRHHLDTTKDKSSALKGVSHQIIKSLSHLHGIVKVRINHLGDIVSVGEY

>II-C\_WP\_125042402.1\_Bacteroides\_faecalis

MKRILGLDLGTSSIGWALVDEAENGNEKSSIVKLGVRVNPLTADEL TNFEKGKSITTNA  
DRTLKRSMRRNLQRYKLRREVLIQILKEHKFITDETILSENGNRTTFETYRLRAKAATE  
EISLEEFARVLLMINKKRGYKSSRKANGTEEGTLIDGMDTALKLYNEDLTPGELCLQYL  
NEGKKTLPDFYRSDLQNECDKIWKQKEFYPEILDTLKEELRGKKRDAVWAICAKYFI  
WKETQTEWNNNEEGKSEQIEKEHKLEGIYSKRKREEAKKENLQWRVNGLKKKLSLEQ  
LVIIHQEINTQINASSGYLGAISDRSKELYFNKQTVGQYQMEILNRNSNASLRNMVFYR  
QDYLDEFNTLWEKQAEFHKELTEELKKEIRDIVIFYQRRLKSQKGLISCFEFESRQIEV  
EIDGKKKVKT VGSRVIPRSSPLFQEFKIWQILNDIKVTVGEKKSKRREKSEQNDSRCL  
NLEEKELLAKELFVRDKMTKSEVLKLLFDNPQELDLNFKTIDGNKTGAAIFRAYSKIEM  
SGHEPVD FKNSKEKIVEHIRAVFASFNWNTDILNFPNDEELDKQSYKWLHLLYSFEG  
DNTPTGNGCLIQKITELYGFEKEYATVLANVT FQDDYGSLSAKAIHKILPHMKEGNQY  
DVACTYAGYKYSKSSLSKEEIANKILKDKLELLPKNSLRNPVVEKILNQMVNVINAIIDIY  
GKPDEIRVELARELKKNAKEREDLT KSIAETTKAHEEYKLLQNEFGLKNVSRTDILRY  
KLYKELESRGHKTLYSNTYIPREKLSKEFDIEHIIPQARLFDDSF SNKTL EARNINIEKG  
NKTAYDFIKEKFGETGTENSLDCYLN NIEELFKSGKISKTKYNKLKMTQQDIPDEFIER  
DLRNTQYIAKKALSMLHEISYRVVATTGSITDKLRQDWQLVDVMKELNWDKYKALGL  
VEYFQDRDGRQIGRIKDWTKRNDHRHHAMDALTVAFTKDVFIQYYNNKNASLNPN TN  
EYAIKNKYFQDGKAI APIPLNEFRAEARKHLENTLISIKAKNKVVT SNMNETKKRGGTK  
KKIQQT PRGQLHLETIYGSNKRYLVREEKVNASFDVDKIETVSKLVYRNALLKRLQEN  
DNDPKKAFTGKNLLDKKPIWLDKEQT KKVPEKVKT VVFETIYSVRKEISPD LKIDKVIDV  
GIRKILADRLSEYGNNAKKA FSNLDENPIWLNKEKGISIKRVAVLGISNAQSLHVKKDK  
NGKVFLDENGKNIPVDFVNTGNNHHVAVYRKPVIDKKGQIVIDEDGNPQYELEE VVVP  
FFDAVTKANLGQPVIDKDYKTSEGWQFLFSMKQNEYFVFPNEKTGFNPKEIDLLDAD  
NYGVISP NFLFRVQKF SHKNYVFRHHLETTIKDTSVLLKGITWTD FRSSKGLDSIVKVRV  
NHIGQIVSVGEY

>II-C\_WP\_240456008.1\_Sphingomonas\_pseudosanguinis

MTDMMNERGMAWRLGLDLGTNSLGWIALQLEGGKPAGIMAIGSRIISSESAGRDPQ  
SKQSLAVNRRDKRAMRRRRDRFKRRQAALLRYLVKDGLFPADQAARQALET RD PYE  
LRARALDERLEPSEIARALFHLNQRRGFKSNRKTDRTGDKDDESGKVAFGIAGLKDKI  
AAAEAAAGAEPGSWTFGKFLCERRAAADDL NALPPTRARAHAAAPPENKNDVYDFYP  
SRDII EA EFAIWEAQAAHHPDLLTDEV RHRLYEIIFYQRPLKAPKIGLCTLVNGETRLA  
KAHPLFQQRRLL EINALQLIRTGEAPRRLMPDERDRILMKFQGKPSLTFVALRKELKL  
DSTVRFNKESENRPKLLGDEMAARMGDKKRFGTHWRHLSVDQQQWEVIQRRQNLES  
DADEAA FVAWLAERYSLTSEQAAAVASVELPMGFGRIGLTATTKLIEALTNGRTEGGE  
VLVYSEAVDAAGLGHHSDDRRSDTRYSDAKGQPILPY YGLPLEHHLLPGTGDPSPDPED  
MRVGRLTNPTVHIGLNQLARVVNALIRRFGRPTEIALELARDLKL SDEEKQRRNKENN  
RNRIEAEKRSIKLREIGKPDTGENRARLKLWEELGEDCLDRRCVYTGERIGIEKLFSDA  
VEIDHILPFSITLDDSNANKIVCMRDANRRKGNRSPHDAWGHGEDWPDIAARASRLP  
ANKRWRFDPDALARFTEERDFQPRQLVDTQYLGRLAREYLTSLYPERGGQSSKVWV

SPGQMTELVRRKLGLNPMLGDDNILGSVNQPKNRLDHRHHSIDAFVVAVIDRAMLAEI  
ARISGETGRDGTETKICIPDPWPNFDELHAALRGIVVSHRPDHGTASKKGLKPGHDQ  
TAGRLHNDTAYGFTGTTDAKGKPIVRRIALTDLKSADLNAIRDPHLRQALRDHTAH  
QDGKAFEERIKAFPPFTLKKVYGGIRHVRVTEALSVIPIRDASGRAYKAYKGDSNYRY  
DVWEMPDGKWITRWQAADGTERSSIVSMFDAHQPREQDRPHPAARKVLSVHQNDL  
LAIEPPGEARRLGRVVKFGQDGRLYLADPKEGGALKARDADKEDAFRYDSLFPKQK  
LWKARQVRIDELGRVLDPGFARKSVRRTRPRTAPVD

>II-C\_WP\_115309099.1\_Kingella\_potus

MNKTYRYILGLDLGIASVGWALVLVDEEENPVGLLDCGVRVFERAEPPKTGESLSAG  
RRVARSIRRLIRRRRAHRLRLRRTLKQSGILTAADFGSDGLPLDLPIDAWQLRVQGLD  
RKLGNKEWAAVLLHLVKHRGYLSQRKSEMKTDDKELGRLLAGVAENSKLLAEASDE  
YRTAAELALKRFAQANGHMRNKGGDYSHTFNRQDLQVELHLLFARQRELGNPFATV  
DLERQADNLLMTQRSALQGEAVLKMLGKCTFEPSEYKAAKNTYSAERFVWLTCLNNL  
RILHNGEERALDEGERKMLLDEPYKKAKLTYAQVRKLLGLPESAVFKGLRYGKDSGD  
LKAESSTTLMEMRAWHAVRKALEKAGLKTWQSLAGKPELLDAVGTAFLSYKTDEDI  
LNYLTGKVASKTVLAIRNFWGGGVLDCEDSGIEYLADEDMTQYLASGVLPVNVLQALL  
EGLNFDEFIHLCLKTLKILPLMEQGMRYDEACRVVYGNHYGLKKKEESRLPHIQAD  
DIRNPVVFRTLTQARKVINAIIRRYGSPQRVHIETARELGKSFKDRKEIEKRQEENRCD  
RERAAAHFKELFPYFVGEPPKGGDILKLRLYEQQQKCLYSGGEIDLRLNEKGYVEV  
DHALPFSRTWDDSFNNKVLVLGSENQKKRNQTPYEYLDGANNSPRWREFQARVAG  
CRFPYAKKQRIQTAKLDEQGFIERNLNDTRYIARFLCNFIGDNLHLEGAGKRRVFASN  
GQITSLLRGLWGLRKVREENDRHHALDAVVVACSTVSMQQKITKAMQRRETTFETVDT  
ETGEIKSRIPQPWDDFFRREVMIRVFSDDPVRELAEKLDSRPEAVHEHVPLFVSRMP  
MRKMSGQGHELETVRSPKRLDEKISTVKKPLSMLKSKDVEKIVGYPDREPALYAALLE  
RLAAHKDDPAKAFAEPFYKPAKGGGQSLVRSVRVEDVQKSGVMIRKDRSGRAQGI  
ADNATMVRVDVYGGGKNYLVPVYAWQVAEGILPNRAVASGKDESDWDVMDASYE  
FRFSLCPNDLVEISKKGRIFGYFSGLDRTSNIHIKEHMDKTKGKDGLHRSGLGVKTV  
QSFQKYQIDPLGKEIRLCKAEPRLSVKKS

>II-C\_WP\_246092851.1\_Terrilactibacillus\_laevilacticus

MILLGVKTLKYSIGFDIGISSVGWAVINHDKNRIEDMNVRMFDAENPKNGSSLAEP  
RNARSSRRRLRRRRHRVLRTRQFIINKGLLTNGQMLKLYDWQDEDLDIWLIRVNALE  
RKLTDREFARILIHFAKNRGFKSNRKSEASDKENGVLAAVKENAQLMEEKGYTTVA  
QLLTYESEIFKGRKRNNGEYTHITTRHDLEHEIHIFAKQREFGHIFANKENEETFVNI  
WSSQRPFATKDDISKIGKCTLEPNEKRAPKAAFTQRFRAIDKLNRLRILSESHPSRK  
VSEQERELAIQGLFSRKDVKYTDLRKLFNLNNERFSELFYETDENIHDVEKKVFLSFV  
EHDKVRKIIKVLGKEALTTFRPVDFDCIAQALTVEKDDNDIRDYLLNRYVNQNGKKER  
NLANRAFDKELIEELLTSFSKFGHLSLRALHRILPHLEKGLSYDKACTEAGYSFNNKL  
EKEKSRLPAVNTNEIRNPVIRSLSQTRKVLNAIKRYGSPTSFIELAREMGRPYNE  
RRDLEKLFTKNHATNEKARNARELHPEISAPRGHDILKYKLWQECCCKCAYSFEPISL  
ERLFEPGYVEVDHIIPYSRSFDDSNANKVLVLAKQNNKRNQTPFEWFGHDEIRWGH  
FESFVNAMQIPSKKVAQKKRSLLLKKVFDDEQEEAFKSRHLNDTRYITRYFKNFVEEN  
LLFREELGVKQHVFTVNGAYTSLMRKRWGFNKSREENDLHHAVIDAVIVAVSGPFRH  
QISAYFKQREVSATQLMKRKGDFPQPWPGFGRELRSRTLQDPKLLKMALESLELPS  
YDEKFISGLKPIFISRMPPKRSVKGQIHEETLRRRRGETEKGYVRIITTKKLEDISFNKNG

DFNMYGKESDPKTYQAIKQRYLDSGNDKKKAFIKPLYKPSKMDNAPLIRTIKIEDKKN  
QIVDLGNGAVAFNGSIVRTEVFRHRESGKFYLAPVYIADVLRKRIPEKFITANKPYDQW  
KDIITEYEFKFNLYPNVDVIKIKMSKMSIKNNLGEVVSWSQEGYVYFKGIDAGTGALKIID  
HMNSYKDRIGSQNLSVFEKYQVNPLGFLVKVHGEKRHGI

>II-C\_WP\_188415172.1\_Pedobacter\_mendelii

MKKILGLDLGTNSIGWALIEVDELNGPFQILAMGSRIIPLDADARDQFQKGQAISKNKD  
RTTARTQRKGYDRKQLKKSDDFKYSLLKILEKLNIFPTDELLNLPSLDLWKLRSNAVS  
NNEEISPKQLGRILYLLNQKRGYKSSRSEANADKKD TDYVVEVKGRYKLLKEKNQTL  
GQYFYDELSSANQNNTYFRVKEKVYPREAYIEEFDAIINTQKSKHSFLTDDVIHSLRNE  
IYYYQRKLKSQKGLVSICEFEGFETSYFDKKTQQDKTIFTGPKVAPRTSPLFQLCKIWE  
VVNNISIKIKNPEGSKYKWDDRIPTIEEKQTIVNHLLNENLSFAELLKILELKKEQVYAN  
KQILKGIQGNTTYSaihKITGNSEHVKFDTH TIPSKHFAVLIDKKTGEILDERDSVELNSA  
LEQEPFYQLWHTIYSIKELDECKNALMKRFNFEEEEIAEKLSKIDFNKQAFGNKSNKAM  
RKMLPYLMLGYNQADAESFAGYNRRLTKEEKSKNVSD EPLQLLAKNSLRQPVVEKIL  
NQMINVVNAIEKYGKPEEIRVELARELKQSKDEREDTDKQNGFNKKLNELVVTKLSEL  
GLPTTKRYIQKYKFIFPAKDKNWKEAQVANQCIYCGETFNLTEALSGDNFDVDHIVPK  
ALLFDDSQTNKVLVHRNCNSTKTNTAYDYISKKGGNALGEYILRVDEWFKRGIISYG  
KMQRKLVSFEEYQERKKVGKETEADKRIWENFIDRQLRETAYIARKAKEILGKVCHNV  
TSTEGNVTAKLRGLWGWDVLMNLQLPKYKEIGQTQIKEWTS DHGNRKHQKEEIAN  
WTKRDDHRHH AIDALVIACTQQGFIQRINTLSSSETKDEM QREIDFAKKKYGN NYDEK  
ILNENNEIVKGQKEKNNQLNSYLISQRPENFTTKYIEQEADKILVSFKAGKKVATISKFK  
ATGKNKETGVIVPRGALHEQFVYGKIKTI AKDFKTGLLIKYPIKYL FENPDLIAEPIV KTK  
VEERLND FGRDIKKCIESLKKFP IYLDLAKSIVLISSYCYKDEFVIKYKITDLKNDDVKYII  
DEKVMHLVQERLDAHNGKEKEAFKDILWFNENKRIPILTVRCFTGLSAVQA IKKDENG  
KEIGFAKTGNNHHIALYKDNTGKHIQHLCTFWHAVERKKFKIPYIINNTNTLWNDLINKE  
LPQSFLDKLPTDNLDLQLSMQQNEMFVLGLSQEEFDEAIQQNNKALLSKNIYLVWSIG  
DNQYWFRHHLETKNSELKVTIGAKESRRFYLFKSIGAFLSKNPIKVRLNHLGEITKVGE  
GHTKEQRYISSEVNLLLLNEPTAIYQTQNRLKIFNSFEEAKQDEIDFIIKQSPIDRIKHTV  
DLILKAYGFNEEDLRIRKQNNTIRIIKAE

>II-C\_WP\_230343355.1\_Methylobacillus\_arboreus

MNNNKNFGQLTFGFDIGIASVGWAVLNERRIVDLGVRAFDKAETAKEGESLNLARRT  
ARLMRRRLRRRAWRLTKVARLLKRHGLINDVTIFKPEHPFTFSLWELRKQALDRPLNA  
EEWARVIYHLCKHRGFHWVSKAE EKAEADTS GEGGKV KQGLAATARLMKEKGYRT  
AAEMVLA EFPQAQRNKQGEYSKALSRSLLSEEFSTLFAVQRAFNNPHASLELESKLL  
DRKTGLFWAQKPALS GEKLEMLGYCTFERKKMTGEKDEYRAPKASFSAERHVWLT  
RLNNLRVVM DGVTRPLSQEERQFVLPLPYNQASDLSYKQLKAALVKQGFWSEAVRF  
TGLSYPSQNQVKDGKAKDPESATLVKIPAFQELRKALKEAGLETEWQGMVNAAMSG  
DASLLDQIAWVLSIYKDDDEVERELKALS LPSKTNMVDVLQNI RFDKFSNLSLKALRNI  
VPHMERGLRYDEACIKAGYHHSQLVKPDDEKLHYLPPFY SARDKTGKLICNENIDVPR  
NPVVLRALNQARKVMNALIKKYGSPHQVNIEMARDLSRPFDERRKIKDAQEEFRDRN  
KISRDKFNETFNTAGDLKGRDFEKWQLYEEQECKCAYSREPLDIHRLFEMGYVEIDH  
ALPYSRSFDDSKNNRVLVLAREN RNKGNLTPYEYLDGKNNSERWHQFEVSVNTNKA  
YRQAKRSRLLRKEFGEKESAEFRDRNLNDTRYICRFFKNYVEQFLRLAETSESKRCIV  
LSGQMTAFLRARWGLTKVRTDSDRHHALDAAVVAACSHGMVKRLSDYSRRKELEKV

SEGFIDVTTGEVANIAMDRQLKEHFDPWPHFRRELEARRLVDDLPLMKAELANLGT  
YSEEELDSVKPLFVSRAPQRRNSGAHKETIYGQTERLKKTGSVTQKVAVASMSLKD  
LGEVIDGERVNCCLIDPHRNEKLYAAAIWLQGKDDREKRAKEIEAGARGKEKRALT  
VDEKAEIERLRALPRKPMNDGSIGPIVRTVTMVIDKLSGISIRHGVAKNDTMLRVDVFT  
KANKFHLVPVYVHHAVAKHLPSRAIVAFKDEDEWTLIDESFAFCFSLYPNDFVKISQK  
GKSPVSGYYSSCHRG TG NLNIW SHDRNQEIGKGGIIEGIGVKTALSIEKFNV DVLGNIY  
PVLKEERRGLA

>II-C\_WP\_108911279.1\_Helicobacter\_apodemus

MKILSFDIGVTSIGWAYVEGEKLKDCGVRIFTKAENPENGDSLAAAPRREARGVRRRLA  
RRKVRLNVIKRLLCKEFGLNLSDYFANDGELPKAYETTKDTKSPYELRSLALEQKLEP  
KDLARVILHIAKHRGYGNKHAKRDLEAEEKAKKAAEKVPQEKESTTTNGDSKKEQEK  
VLKALYQNETLLKNYQINYQTNDETNNKTIFLVGKYLYKEFQDKGQRIRNTTNNYQHT  
MRQEWLKDLEFIFAKQREFGFPFSEKLYKVATENNKTKELDFAAIMETA FYQRDLK  
SFEDKVGKCVFYENENRAPKDSL SVM EFVAL TRIINTLKNLEEKSKNLGIGETYGDKDI  
QEILKIVLDKGEMSYKELRKILKLNESILFAKDPKLDYTKGVQEA EKAKFIELKNLKAFK  
KAMGGETEQKANKKTKKTIEVSLESFDRKELDLIATHIALIKSKENLAERLRNYPALSKE  
QVEALSNNLFAKHIDLSLKALDKILPLMREGLRYDEAVKKAGLQEHRKHKQKGEFLIPL  
KEYEPYLANPVVARALSEYRKVLNALLKKYGVHVKIHLEFTREAKLSAKERQKYEKEQ  
REHFEANQNAIKQCEKLDLRLNGANILKMKLWIEQQEYCAYSVKKITTEDLKDSNKLQI  
DHIYPYSRSFDDSYMNKVLVFSKENQNKLNKTPFEAFGSDKEKWDRILSLSAKLPKQ  
KQRRISNTNFKDKEGGFLARNIVDTGYIARLASQWTNHCLKFEDLRENEVTITGEKGS  
SVHIEVVSGSLTATLRHYWGLGDKDRNNHLHHAVDAILAYVNDTMIQAFANFRKNQE  
QNKAKFYAKQIAESEYKAKKAFFEP CENFRDKVLKKLDSIFVSKPPRKRARGALHKET  
FYSFEDEKLLKSYGGKQGVQRAIDLKIRRIGTKIVSNGAMVRVDIFKHKEGKFYAVPI  
YTMDFAFGEWPNKAVATGKDEDGVIKDWIEMDSYDFCFSLFKDDLIKVKKKKMKEA  
EFCYFKRFTSGASIEIVKHDNRADNLNGNQEKIFKVLENQKILDKSLISIQNLEIFEK  
WQVSPLGKVQKAVPESRQGIKLSLKRKG

>II-C\_WP\_101954953.1\_Tenacibaculum\_finnmarkense

MSKILGLDLGTNSIGWALVQNNIILGMGSRIFPEGVVNLGEGEGREASKNASRTEARG  
VRRQIFRRRLRKRYLLKELAKHNFCPISADLIKDWNGTEIFKNQELQDWFKLNPYELR  
AKAIHQEISLVEIGRIFYHIIQRRGFLSNSRSAGADV KETSTLFTGDAKIGKIGISSTLENI  
KNIGNSEENKTLGSYLSIQPKENEPYSECLERIRNRYTTRQMYIDEFELIWEFQKIYH  
SELTENLKS LFGGRKKDGYKDDGVLFHQ RPLRSQKHLVGYCSFEPKKTCKKISAI PFE  
LFRIYQWVN TLSYDFNGEKQEITSEEKEKIIQLLLSKEKVAFAIRKVIDKSESNYQFNY  
KDKDNIVGTH TISQLSNKKFFGKQWFHFSEEEQENIWHVLSFFDDSDKLKEYALKNW  
NFD DKKATEISKFNLKEGYSNLSRK AINTILPFLKQGFMFDIAVGLAGVKNVLSKENNW  
AENEQFVLDNVPEIVRSNIKGGYIEPLKNMLKNEFNCTDKQLEKLYHHSSAIETQIILDK  
LPVNINADREIQKIKNPVVITALFEIRKLVNQIITTYGKPTEIKVELARDLKASKTNRQETR  
RRQQILERENDRVKVELEGLDQRTNHANILKYKLWEECNKTC PFTGKVITLSKLFTGE  
VQIEHIFPWSRSLNDSFN NKTLCFADENRAKGNKTPYEFYTSQGLEKWEAIKLQALSC  
FKTKAHYPNCYNKFKHFVKQKHDEDFISRQLNDTRYISKEAKNYLSKICANVMVAPGQ  
MTAVLRHNWGLNTVLNKEQDAKTRDDHRHHAVDALVMACAKRSHLQELSKRNRHYH  
KNHELKDFPMPWQSFRMDAETA VEQILVSHKKQKNLVTIRTHKTKKNGVEYKNLGVS  
ARGQLHKESVYGKRTAPDSSQAFHIRKPLESLTTEKHLDKIVDATVRLLILKRVKELGG

FVKGKIPADTFFKIDENGIKQAQIFLPNRNGSKVPIKKIRIKENIGGAEKLKEDINQYVNP  
RNNHHVLIYKDEDDNLKEDVVTFTWATERKRQKQDVFQLPEDGKKIVTTLQINDMFL  
GLQNEQVDWDNIDANLLKDHLRYVQKFTSKDYFRKDKESNLDGKLGIAFHYIKGFG  
TGKTGWKTFNPIKVGINSIGVLKKL

>II-C\_WP\_013389026.1\_llyobacter\_polytropus

MKYSIGLDIGIASVGWSVINKDKERIEDMGVRIFQKAENPKDGSSSLASSRREKRGSR  
RNRKKHRLDRIKNILCESGLVKKNEIEKIYKNAYLKSPWELRAKSLEAKISNKEIAQILL  
HIAKRRGFKSFRKTDNRNADDTGKLLSGIQENKKIMEEKGYL TIGDMVAKDPKFNTHVR  
NKAGSYLFSFSRKLEDEVRIQAKQKELGNTHFTDDVLEKYIEVFNSQRNFDEGSPK  
PSPYYSEIGQIAKMIGNCTFESSEKRTAKNTWSGERFVFLQKLNNFRIVGLSGKRPLT  
EEERDIVEKEVYLKKEVRYEKLRLKILYLKEEERFGDLNYSKDEKQDKKTEKTKFISLIG  
NYTIKKLNLSEKLKSEIEEDKSKLDKIIELTFNKSDKTIESNLKKLELSREDIEILLSEEF  
S  
GTLNLSLKAIKKILPYLEKGLSYNEACEKADYDYKNNGIKFKRGELLPVVDKDLIANPVV  
LRAISQTRKVVNAIRKYGTPHTIHVEVARDLAKSYDDRQTIKENKKRELENEKTKKFIS  
EEFGIKNVKGKLLLKYRLYQEQEGRCAYSRLKSLSEVILDESMTDIDHIIPYSRSMDD  
SYSNKVLVLSGENRKKSNLLPKEYFDRQGRDWDTFVLNVKAMKIHPRKKSNNLLKEKF  
TREDNKDWKSRALNDTRYISRFBVANYLENALAYRDDSPKKRVFMIPGQLTAQLRARW  
RLNKVRENGDLHHALDAVVAVTDQKAINNISNISRYKELKNCKDVIPSIEYHADEETG  
EVYFEEVKDTRFPMPWSGFDLELQKRLESENPREEFYNLLSDKRYLGWFNYEEGFIE  
KLRPVFSRMPNRRGVKGQAQETIRSSKKISNQIAVSKKPLNSIKLKDLEKMQGRD  
TD  
RKL YEALKNRLEEYDDKPEKAFAEPFYKPTNSGKRGPLVRGIKVEEKQNVGVYVNGG  
QASNGSMVRIDVFRKNGKFYTVPIYVHQTLLKELPNRAINGKPYKDWDLIDGSFEFLY  
SFYPNDLIEIEFGKSKSIKNDNKLTKTEIPEVNLSEVLGYRGMDDTSTGAATIDTQDGKI  
QMRIGIKTVKNIKKYQVDVLGNVYKVKREKRQTF

>II-C\_WP\_106152678.1\_Marinilabilia\_salmonicolor

MKRILGLDLGTNSIGWALVEKDEKKKEIIGMGSRIPLSPDDKDEFSGKNTISKNHRT  
VKRTQRKLYDRYQLRRNALIEELKNKNMMPDEDLIHVSKFRLWELRSKAVETKLTLP  
E  
IGRILLHLNQKRGYKSARSEENLDKKDTEYVEKVKNRHEELRDAGLTIGQYLLKELKK  
SENFRTKDLVYPREAYIEEFDRIISIQKGYYPEELSEDFIQRIRNEIIFYQRALKSQKGLV  
SICEFEGFWMEKEGKNLYVGPKVAPKSSPLFQIEKIWETINNIRLKDQYGGPIVLGIRE  
KEKLFKVLNIPGAGSEDFDNFSKYSLDRNGKLSYSKLLNLLGLSKDDVYANKQLENGI  
QGNILRQKILNCFNNPDEFSELLRFDISFIKTDKKVHLYDRKTGEVIHEEKESIIANKAED  
IEKEPLYRLWHVLYSIKDREECENILKKKFNIPPD SAEKLSALDFTLMAFGNKSVRAIRR  
ILPYLAEGYDYSESCMLAGYNHNSLT KDQNRDRQIAERLSLLEKNSLRQPVVEKILN  
QMINLVNAVIEEYGHPPDEIRIELARELKQSRQERFDADKAIRKIERDNQSIRKELVENYA  
LRASRNNVIKWRLYHEMNNQKLNAICVYCGQPISFTDAIKGNDVDIEHIIPKAKLFDDS  
QSNKTLSHRSCNRQKDKQTAFDYMKTKEDEFHEYIERVNYLYKNRLIGKAKRDKLL  
MSEDKIPQDFIQRQLRESQYIARKSREILMEVCHDVWTTSGTVTAELRHIWGWDDVL  
MNLQMPKYKKYDLTEEIEFENNGQKHKKERITDWNKRNDHRHHAIDALT VACTEQGF  
IQRLNTLNSGKTRNEMLRINEFQKNASEVDNKRSEFEIEKRKLYQDYLFVNKPFSTN  
EVEQAASKIMISFKPGKKVATWGKRKVKRNGRKT VQSDILVPRGALSEEQVYGRIK  
VLKRENPKVFLFHNPHLIVKERIRKKVQKRLSEYDDNPDKALKSLENEPIYIKDDILLEW  
ASCYSEEVYKYKVDVDFKKADKIIDDRVREIVKQRLSEFDNNPKEAFKDL DGNPIWF  
NEEKNIPIKSVRCFTGLSSIEPLKKNANNEVTGFVKPGNNHHIALYLDENGKLQEHICT

FWHAVERKKYDIPVIIKNTDEIWEKVLSEGEDYFPDDFLEKLPDIKWQFVTSLQQNEM  
FLLGLPEENVNRQINVNQKDNYSFLSDYLYRVQKITSGDYVFRHHLETQIVDTTDSKLC  
KRWYRVQSLKAFAGLNPVKIRVDVLGRMVWYQKCF

>II-C\_WP\_003611034.1\_Methylosinus\_

MRVLGLDAGIASLGWALIEIESNRGELSQGTIIGAGTWMFDAPEEKTQAGAKLKSEQ  
RRTFRGQRRVVRRRRQRMNEVRRILHSHGLLPSSDRDALKQPGLDPWRIRAEALDR  
LLGPVELAVALGHIARHRGFKSNSKGAKTNDPADDTSKMKRAVNETREKLARFGSAA  
KMLVEDESFVLRQTPTKNGASEIVRRFRNREGDYSRSLRDDLAAEMRALFTAQARF  
QSAIATADLQTAFTKAAFFQRPLQDSEKLVGPCPFVDEKRAPKRGYSFELFRFLSRL  
NHVTLRDGKQERTLTRDELALAAADFGAAAKVSFTALRKKLKLPETTVFVGVKADEES  
KLDVVARSGKAAEGTARLRSVIVDALGELAWGALLCSPEKLDKIAEVISFRSDIGRISE  
GLAQAGCNAPLVDALTAASDGRFDPFTGAGHISSKAARNILSGLRQGMYDKACCA  
ADYDHTASRERGAFDVGGHGREALKRILQEERISRELVGSPTARKALIESIKQVKAIVE  
RYGVPDRIHVELARDVGKSIEEREEITRGIEKRNQKDKLRGLFEKEVGRPPQDGAR  
GKEELLRFELWSEQMGRCLYTDYISPSQLVATDDAVQVDHILPWSRFADDSYANKT  
LCMAKANQDKKGRTPYEWFKAEKTDTEWDAFIVRVEALADMKGFKKRNKYLNAEE  
AAAKFRNRNLNDTRWACRLAEALKQLYPKGEKDKDGKERRRVFSRPGALTDRLRR  
AWGLQWMKKSTKGDRIPIIDRRHHALDAIVIAATTESLLQRATREVQEIEDKGLHYDLVK  
NVTTPWPGFREQAVEAVEKVFVARAERRRARGKAHDATIRHIAVREGEQRYERRK  
VAELKLADLDRVKDAERNARLIEKLRNWIEAGSPKDDPPLSPKGDPIFKVRLVTKSKV  
NIALDTGNPKRPGTVDRGEMARVDVFRKASKKGKYEYLLVPIYPHDIATMKTPPIRAV  
QAYKPEDEWPEMDSSYEFCSLVPMTYLQVISSKGEIFEGYYRGMNRSVGAIQLSA  
HSNSSDVVQGIGARTLTEFKKFNVDRFGRKHEVERELRTWRGETWRGKAYI

>II-C\_WP\_214161343.1\_Litoribacter\_ruber

MKKILGLDLGTTSIGWAFVHEKEKEEEEEENRIVDVGVRIVPLTSDEENDFAKGKTISINA  
DRTLKRSARKNLQRFKQRRDALLEIFAGLGYIDSDFSCAESGSKSTYSTLTLRALAVT  
EKISREDFVKVLMALNKKRGYKSSRKVKSEDDDEGVAIDSMGIAKELYQSNLTPGQWV  
FEALKKGRKTIPDFYRSDLEEEFRCIYEFQKEFFPELLPTDFLDNVLGKTKDQTRFYFN  
KLHNIPLAENKGKRPRTLQHYKWRNDALKEKLAPSELAFLFTEINGQISASSGYLGA  
SDRSKELYFNQETVGQYLLKQVEKNRHLRLKGQVFYRQDYLDDEFERIWEKQAAYFP  
ELTEKLKKEVRDITIFYQRRLKSQKHLISTCEFELLQKVAPKSSPIFQDFRIWQNLNHVT  
LTNTETFEKLSLSQEHKEQLQAELTFKKKMAAADVLKEIGYDKNWELNFEELQGNRT  
NHEIFTAFAKIIELEDGNEIKLNKLRADEILDVFSEAFRLRFGINPEYLQVNNGLVGDGFE  
KQPYLQLWHLLYASEDEKKLRKTMVKKFNFKENHAKIISGVSFQQEHCSLSSKAICKIL  
PYLILGLDYSEACEKAGYNHSSFLTKEENDNRPLKDELELVKKNALRNPVVEKILNQLV  
NVVNAIVKDPALGRPDEIRVELARELKANSEQRKKMTNDIGRATKYHQEIRELLKKEF  
GISRVTRNDIIRYKLWKECGGISIYTGRPIEATKLFSKEYDIEHIIPKARLFDDSFNKTIC  
ERQLNADKSNITAFITLQETLSAGDFEQYLARVKGLYGKIGRTKIQKLLMATDKIPTDFI  
ERQIRETQYISRKAKEILLISRNVTVTSGSVTDKLRDDWGLVEVMKELNWDKYNALG  
LTKIEKGKNGERLKKIKDWSKRNDHRHHAMDALAVAYTKPAYIQYLLNLSAKGIEGEK  
GIEIRAIERKYLKFNRLKFISPIPNFREEARFHLGNILVSFKSKNKVVTINKNKIKKAG  
GFLIQTALTPRGQLHKETVYGKQHGYKTKIEKVGGNFDSESIKAVANKKYRDALSLRL  
QQFGNDPKKAFTGKNSLTKLPIYLSLEKNITVPEKVKLVWLEPYTYTIRKNITPDLKVDK  
VIDPKIKSILKNRLTSYNEDPKLAFANIDENPIWLNKEKGIAIKVKISGVSNIVSLHYKKD

NRGNFILNAEGQKEEVNFVSTGNNHHVAIYRDEKGDQLQEEVVSFFEAVERKNQGRA  
VVEKQHEKGWEFLFTLKQNEYFVFPSMDFDPGDIDLMDPTNYSLISPNLYRVQKIAAK  
NYMFRHHLETNVEEPAALKGIAFKSIRSTAPLKGIVKVRLNHLGKIVQIGESSPNILLETI  
AKIKKNEALPEK

>II-C\_WP\_011809011.1\_Verminephrobacter\_eiseniae

MNKAVAYRLALDLGSTSLGWAIFRLNEAREPTAIRAGVRIFSDGRNANSEPLAVHRR  
VARAMRRRRDRLLKRKKRMHDQLVQHGFPAEVAERKELERLNPYQLRAKGLHEAL  
TPGEFARALFHINQRRGFKSNRKTDRKDNDSGALKQAISELRKRIQSDSCATAGEWF  
WKERMQQKPEGVRGQWVRARYRKTPSTTDEGKKRIGYDLYVDRAMVEQEFDALW  
AAQAALQPGLFTEAARAELKDTLLYQRDLRPVKPGRCTLLPAEERAPLALPSTQRFRI  
LQEVNNLRLLDEALREVPLNLAQRDAVVSALLEGKKELSFAAIRKLLTSLGKFNLEDEK  
RAELKGNATSVILARKDLFGDAWAGFDAAVQDEIVWRLVSQESEGALIAWLQQHTGV  
DGVCAEAIVNTRLPDGYGRLSRKALERIVPALQREVCTYDKAVQAAGFAHHSDLGFD  
FDYAEDEVQQVGEHTIASTGEVQAQYAFKQLPYYGKALQRHVAFGSGDAKDHEEKC  
YGKIANPTVHIGLNQVRTVNALIRRYGHPTEVVVELARDLKQSREQKQQTQREQAD  
NQKRNEDIRKRIAPILETSPERVRDRDIKKWILWEELNKKDIADRHCPYSGERISATML  
LSEAVEIEHILPFSRTLDDSLNNRTVAMRRANRIKGDRTPWEARADFEAQGWRYEAIL  
QRAEGMPPRKRYRFAEDGYQRWL GKDRNFLARALNDTSYLSRLAANYLRLVCPQG  
VRVIPGQMTDKLRGKFGLNSVLGLDGKKNRNDHRHHAVDACVIGVTDQGLMQRFAN  
ASKQARENGLTRLVQDMLLPWWPSYYDHVERAVRHIRVSHRPDHGFEGAMMKGTA  
HGIREDGIREDGRIKQRPKAKGSADHKTITLIPIDEPRQLARHGVD AEGKPLPYKGYAS  
GSNYCIEITKNGKGKWEQQVISTFDAYRIKAAADAAARAGQVISTVEADSIVRKSGWE  
RLRGAQSQNGQPLVMRLVIGDSVRMEVDGRDEV MRVVKMSGKEMVFAPVREANVD  
KRNNMPDEQDPFTYTYKRADQLRKAKARQVTISPIGELRDPGFKG