Computational analysis of DNA and RNA editing in human cells

Srinivasan, Harini

2019

Srinivasan, H. (2019). Computational analysis of DNA and RNA editing in human cells.
Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/136576>

<https://doi.org/10.32657/10356/136576>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0
International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:54:51 SGT

COMPUTATIONAL ANALYSIS OF DNA AND RNA EDITING IN HUMAN CELLS

HARINI SRINIVASAN

School of Chemical and Biomedical Engineering

2019

COMPUTATIONAL ANALYSIS OF DNA AND RNA EDITING IN HUMAN CELLS

HARINI SRINIVASAN

Supervisor: Dr. Tan Meng How
Co-supervisor: Dr. Chandra Verma

School of Chemical and Biomedical Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2019

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

HARINI
SRINIVASAN

16 JAN 2019

Date

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

Dr. Tan Meng How

16 JAN 2019

Date

Authorship Attribution Statement

This thesis contains materials from two published papers in peer-reviewed journals as detailed below, where I am a co-author.

1. **Chapter 3** describes the work published in "Wang Y, Liu KI, Sutrisnoh NB, Srinivasan H, Zhang J, Li J, Zhang F, Lalith CRJ, Xing H, Shanmugham R, Foo JN, Yeo HT, Ooi KH, Bleckwehl T, Par YYR, Lee SM, Ismail NNB, Sanwari NAB, Lee STV, Lew J, Tan MH. Systematic evaluation of CRISPR-Cas systems reveals design principles for genome editing in human cells. *Genome Biol.*, 19(1):62, May 2018".

The contribution from authors are listed below:

- Experiments were performed by YW, KIL, NABS, JZ, JL, CRJL, HX, RS, HTY, KHO, TB, YYRP, SML, NNBI, NABS, STVL, JL, and MHT.
- HS developed and fine-tuned the pipeline for analyzing the Illumina sequencing data.
- HS, FZ, JNF, and MHT analyzed the Illumina sequencing data.
- HS provided inputs for the manuscript relating to the computational aspects.

2. **Chapter 6** described the work published in "Shanmugam R, Zhang F, Srinivasan H, Charles Richard JL, Liu KI, Zhang X, Woo CWA, Chua ZHM, Buschdorf JP, Meaney MJ, Tan MH. SRSF9 selectively represses ADAR2-mediated editing of brain-specific sites in primates. *Nucleic Acids Res.*, 46(14):7379-7395, Aug 2018".

The contribution from authors are listed below:

- Experiments were performed by RS, CRJL, KIL, CWA, ZHMC and MHT.
- HS worked with FZ for developing the pipeline for analyzing the Illumina sequencing data.

- FZ, HS, XZ and MHT analyzed the Illumina sequencing data.
- HS, FZ and MHT provided inputs for the manuscript relating to the computational aspects.

HARINI
SRINIVASAN

16 JAN 2019

Date

Acknowledgements

First of all, I wish to thank my supervisor Dr. Tan Meng How for giving me an opportunity to pursue my research under his guidance as a graduate student. Being a part of his research group, I had the freedom to explore and learn many things. Also, being in his group gave me the opportunity to be part of projects that involved both experimental and computational aspects and this gave me a chance to interact with experimental lab members that helped increase my understanding of their work and also taught me to present my work to experimentalist. Regular interactions with him throughout this period helped me focus and structure my research. His constant push for excellence helped get the best out of me. I sincerely thank him for giving me this opportunity and guiding me.

I am equally grateful to my co-supervisor Dr. Chandra Verma at the Bioinformatics Institute, A*STAR, for his guidance, constant support and mentorship during my tenure. Being a part of his group, I learnt a lot about different aspects of molecular modeling and simulation. Apart from this, his constant reminder to keep in touch with latest technology and updates has helped me think about the future and push my scientific endeavors further. He has always been available for discussions and I thank him for his time and support. Finally, his valuable suggestions helped me improve my thesis.

I would like to thank the members of the DaRE (DNA and RNA Editing) lab and Dr. Verma's group. The discussions with the groups during the group meetings helped enhance my critical thinking and research output. The opportunity to present my work to people from different background during these meetings also helped me improve my presentation skills. I would particularly like to thank Dr. Adelene Sim from Verma's group with whom I worked closely for the modeling and simulations project on Cas9 (Chapter 2). The scientific discussions I had with her were very productive and insightful.

I would also like to thank the administrative and support staff from both Bioinformatics Institute (BII) and Genome Institute of Singapore (GIS) for providing a comfortable environment and computational support for carrying out my research at both the institutes.

Last but not the least; I would like to acknowledge my family who have been my pillar of support during these 4 years and without whom this work would not have been possible. I would like to thank my parents, Usha Srinivasan and S.R. Srinivasan, for instilling in me the hunger for knowledge right from a young age and always being supportive in all my endeavours. Also, I would like to specially thank my husband, Anand Bardwaj Araamudhan, and my sister, Ranjani Srinivasan, for constantly motivating and encouraging me to strive harder.

Table of Contents

Acknowledgements	i
Abstract	vii
List of Figures	xi
List of Tables	xiii
List of Abbreviations	xv
1 Introduction	1
1.1 Computational techniques and their basics	2
1.1.1 Molecular modeling and simulation	3
1.1.2 Analyzing high-throughput sequencing data	8
1.2 Editing in human cell	10
1.2.1 DNA editing: CRISPR	10
1.2.2 RNA editing	15
1.3 Scope and objective of this work	17
1.4 Conclusion	19
2 Uncovering the role of highly conserved residues in the HNH nuclease domain of Cas9 using molecular dynamics simulations	21
2.1 Background	22
2.2 Methodology	24
2.2.1 Sequence analysis of type II Cas9	24
2.2.2 Initial structure preparation for MD simulations	25
2.2.3 Structural Analysis	27
2.3 Results	28

2.3.1	Sequence analysis of Cas9 reveals minimal sequence conservation	28
2.3.2	P843 maintains the conformation of the active site cleft . . .	36
2.3.3	K866 forms interactions with the catalytic residues and Mg^{2+}	38
2.3.4	Zn^{2+} cannot substitute for Mg^{2+} in coordinating the highly conserved residues within the active site of the HNH domain	43
2.4	Discussion	46
2.5	Conclusion	48
Preface to Chapter 3		49
3	Developing computational pipeline for a systematic evaluation of CRISPR-Cas systems in human cells	51
3.1	Background	52
3.2	Design of the experimental study	54
3.3	Pipeline developed	55
3.4	Results and biological inferences	60
3.4.1	"Performance of the CRISPR-Cas systems in NHEJ-mediated genome editing"	60
3.4.2	"Performance of the CRISPR-Cas systems in HDR-mediated genome editing with single-stranded oligodeoxynucleotide donor"	70
3.4.3	Optimization of ssODN donor templates	73
3.4.4	"Enhancement of error-prone repair with long single-stranded DNA"	77
3.4.5	"Performance of CRISPR-Cas systems in HDR-mediated genome editing with plasmid donor"	79
3.5	Discussion	80
3.6	Conclusion	84
Preface to Chapter 4		85
4	Characterizing non-random mutations introduced by CRISPR-Cas systems in NHEJ-mediated genome editing	87
4.1	Background	88
4.2	Pipeline developed	90
4.3	Results and discussion	92

4.3.1	Insertions are prominent signature repair outcomes while using SpCas9	94
4.3.2	Prominent indel classes identified from the mutations introduced by the CRISPR-Cas enzymes vary with the enzymes .	95
4.3.3	Correlation between indel classes introduced and spacer length of the locus	97
4.4	Future work	98
4.5	Conclusion	99
Preface to Chapter 5		100
5	Characterizing glutamine synthase knock-out based on HEK293 for use as a human expression system for stable production of recombinant erythropoietin	103
5.1	Background	104
5.2	Design of the experimental study	107
5.3	Pipeline developed	107
5.4	Results and biological inferences	108
5.4.1	Gene expression based classification of cell lines	109
5.4.2	Differentially expressed genes between wildtype and GLUL knockout cell lines	110
5.4.3	Differentially expressed genes between adherent and suspension cell cultures	112
5.4.4	Glycoprotein biosynthesis in GLUL knockout cell lines . . .	112
5.5	Conclusion	113
Preface to Chapter 6		115
6	Genome-wide editing analysis identified SRSF9 as selective repressor of ADAR2-mediated editing of brain-specific sites in primates	117
6.1	Background	118
6.2	Experiments and biological inferences	120
6.3	Pipeline developed	122
6.3.1	Pipeline for analyzing bulk RNA sequencing	122
6.3.2	eCLIP data analysis pipeline	123
6.4	Results and biological inferences	125

6.4.1	SRSF9 down-regulating ADAR2-mediated RNA editing at RADAR sites	125
6.4.2	Editing sites directly regulated by SRSF9 identified using eCLIP-Seq	126
6.4.3	Down-regulation of ADAR2-mediated editing by SRSF9 in multiple human tissues	130
6.4.4	Physiological consequence of SRSF9 mediated editing re- pression	131
6.5	Discussion	137
6.6	Conclusion	139
7	Conclusion: summary, and future directions	141
7.1	Main contributions and significance of work	141
7.1.1	Computational contribution	142
7.1.2	Biological insights	145
7.2	Summary, impact and future directions	146
7.2.1	Uncovering the role of highly conserved residues in the HNH nuclease domain of Cas9 using molecular dynamics simulations	146
7.2.2	Developing computational pipeline for a systematic evalua- tion of CRISPR-Cas systems in human cells	147
7.2.3	Characterizing non-random mutations introduced by CRISPR- Cas systems in NHEJ-mediated genome editing	148
7.2.4	Characterizing glutamine synthase knock-out based on HEK293 for use as a human expression system for stable production of recombinant erythropoietin	149
7.2.5	Genome-wide editing analysis identified SRSF9 as selective repressor of "ADAR2-mediated editing of brain-specific sites in primates"	150
7.3	Concluding Remarks	151
	Author's Publications	152

Abstract

Use of computational techniques is inevitable in present day biological research and this work describes my efforts in utilizing these computational techniques on biological systems in combination with experiments to increase our understanding of them. DNA editing (using CRISPR-Cas system) and RNA editing (ADAR-mediated) in human cells are studied in this work with an aim to increase our understanding of functioning of these editing mechanisms.

CRISPR-Cas systems are RNA-guided DNA editing systems that have their origin in bacteria and archaea. They hold huge potential for genome engineering due to their ability to introduce a double-stranded break at any desired genomic location. In an attempt to increase the understanding of the functioning of these CRISPR-Cas editing systems, a structural study was designed to probe the role of conserved residues in the active site of the HNH nuclease domain of Cas9 (refer chapter 2). This study highlighted the likely role of highly conserved residues in maintaining the structure of the active site cleft in the HNH nuclease domain and thereby, in the functioning of the enzyme. Succeeding these efforts, another study based on activity of enzymes was designed to gain insights into the functioning of the CRISPR-Cas enzymes. Here a large scale characterization of the functioning of five CRISPR-Cas enzymes was done by comparing their performance in genome editing pathways using a combination of experiments and computational pipelines developed to analyze and quantify the NGS data from the experiments (refer chapter 3). This study highlighted the factors affecting the functioning of the five enzymes and their performances. Additionally, the mutations introduced by the five enzymes in NHEJ-mediated genome editing was identified to reveal non-random patterns in the next study (refer chapter 4). This study enabled characterizing the most probable insertion and deletion outcomes from the NHEJ-mediated genome editing using the five enzymes. The results from both these studies give us insights that can serve as a guidance for designing future genome

engineering studies and also enable using the CRISPR-Cas technology safely to its fullest potential. Further, an application of CRISPR-Cas editing system was showcased by using it to produce a glutamine synthase knockout cells based on HEK293 cells in the next study (refer chapter 5). These knockout cells were characterized for their utility as a bioproduction platform for stable production of glycoproteins with resemblance to human proteins. These cells are then developed into a suitable expression system for stable production of recombinant erythropoietin (EPO) with fully human glycosylation. Finally, another aspect of gaining insights into the functioning of a system involves understanding its regulators and RNA editing system was employed for this study. RNA editing refers to site-specific nucleotide change in RNA after transcription, and adenosine to inosine (A-to-I) editing is the most prevalent modification in higher eukaryotes mediated by ADAR family of enzymes. In this study, regulators of ADAR-mediated RNA editing were probed for using a combination of experiments and genome-wide computational study (refer chapter 6) with an objective of gaining insights into its mechanism. This study revealed splicing factor, SRSF9, as a key repressor of ADAR2-mediated editing of brain-specific sites in non-brain tissues, thereby, enabling the shaping of the editing landscape.

List of Figures

1.1	Algorithm of molecular dynamics simulations	8
1.2	Schematic of analyzing NGS data	9
1.3	Schematic of the domain architecture of ADAR family of proteins .	16
1.4	Computational Analysis of Editing	19
2.1	Domain organization in SpCas9	23
2.2	Local rearrangement of atoms around the active site of SpCas9 . .	24
2.3	Active state of HNH domain of SpCas9 with Mg^{2+} and short single-stranded DNA in the active site	26
2.4	Multiple Sequence Alignment of Cas9 sequences	32
2.5	Multiple sequence alignment of the of HNH domain of Cas9s	33
2.6	Multiple sequence alignment of the Pfam HNH_4 family	35
2.7	Orientation of conserved residues in the HNH domain in the presence and absence of ion in the active site	36
2.8	Root Mean Square Deviation for the simulation	37
2.9	Interplay between two loops surrounding the active site of the HNH domain	39
2.10	Fluctuations in DNA in the wildtype and mutant simulations	40
2.11	An interaction network between three highly conserved residues in the $\beta\beta\alpha$ -motif and the Mg^{2+} cation	41
2.12	Secondary structure of the HNH domain in the wildtype and mutant simulations	42
2.13	Interactions between the $\beta\beta\alpha$ -motif and the DNA	44
2.14	Effect of Zn^{2+} cation in the active site of the HNH domain	45
3.1	Expression of selected target genes in HEK293T cell line	55
3.2	Two endogenous repair pathways for comparing the Cas endonucleases	56

3.3	Schematic of the computational pipeline developed to analyze the NGS data	59
3.4	Evaluating Cas enzymes in NHEJ-mediated genome editing using matched spacer	61
3.5	Effect of genomic location of the target site in NHEJ-mediated genome editing	63
3.6	Evaluating Cas enzymes in NHEJ-mediated genome editing with overlapping seeds	64
3.7	Evaluating NmCas9 and Cpf1 enzymes in NHEJ-mediated genome editing with same spacer at higher spacer lengths	65
3.8	Relationship between DNA cleavage efficiency and gene expression .	67
3.9	Comparing cleavage efficiencies of Cas enzymes at different spacer lengths	68
3.10	Relationship of DNA cleavage efficiency with target specificity . . .	69
3.11	Evaluation of CRISPR-Cas systems in HDR-mediated genome editing at loci with overlapping seeds and using ssODN as donor templates	71
3.12	Extent of HDR-mediated genome editing with different amounts of ssODN donors	72
3.13	Correct versus incorrect integration of XbaI restriction site using HDR-mediated genome editing	72
3.14	Measurement of cleavage efficiency at loci where HDR-mediated genome editing is performed using ssODN	73
3.15	Evaluating HDR efficiency using symmetric ssODN donors with different orientations	75
3.16	Measurement of cleavage efficiency in HDR-mediated genome editing with NT or T ssODNs	76
3.17	Evaluating HDR efficiency using symmetric and asymmetric ssODN donor templates with different orientations	78
3.18	Cleavage efficiencies in the presence and absence of ssODNs	79
3.19	Gene disruption efficiency in the presence of different single-stranded DNA donors	80
3.20	Guideline for selecting appropriate CRISPR-Cas system depending on the application	81
3.21	Pairwise comparison of SpCas9 and LbCpf1 performance	82
4.1	Nomenclature for naming indel start position	90

4.2	Schematic of pipeline developed to identify indels and prominent patterns	93
4.3	Characterizing CRISPR-Cas enzymes ability to introduce insertions and deletions	94
4.4	Prominent indel classes introduced by CRISPR-Cas enzymes	96
4.5	Prominent indel class for the CRISPR-Cas enzymes grouped by spacer length of the locus	98
5.1	Computational pipeline developed for characterizing the GLUL knock-out cell lines	108
5.2	Principal component analysis of wildtype HEK293 and GLUL knock-out cell lines	109
5.3	Heatmap based on gene expression of wildtype HEK293 and GLUL knockout cell lines	110
5.4	GO analysis of differentially expressed genes between wildtype and GLUL knockout cells	111
5.5	GO analysis of differentially expressed genes between adherent and suspension cells	113
5.6	Expression profile of genes involved in glycoprotein biosynthesis . .	114
6.1	Schematic of the computational pipeline developed for transcriptome and editing analysis	124
6.2	ADAR2-mediated RNA editing down-regulated by SRSF9 at known RNA-editing sites	126
6.3	eCLIP-seq guided identification of editing regulon of SRSF9	128
6.4	Distribution of editing level of genes bound but not regulated by SRSF9	129
6.5	Distance distribution between editing site and closest SRSF9 binding site	130
6.6	Editing profile of GTEx tissues showing SRSF9-mediated regulation of brain-specific sites	132
6.7	Transcriptome analysis of genetically perturbed SH-SY5Y cells . . .	134
6.8	Editing analysis of genetically perturbed SH-SY5Y cells	136
7.1	Integrated schematic of the computational pipelines developed in the thesis to analyze NGS data	143

List of Tables

1.1	List of Cas effector enzymes acting on DNA and RNA	14
2.1	List of simulations performed	27
3.1	Operators in CIGAR string	58

List of Abbreviations

- ADAR: Adenosine Deaminases Acting on RNA
- ALS: Amyotrophic Lateral Sclerosis
- AMBER: Assisted Model Building and Energy Refinement
- AnCas9: *Actinomyces naeslundii* Cas9
- AsCpf1: *Acidaminococcus* sp. BV3L6 Cpf1
- A-to-I: Adenosine-to-Inosine
- BHK: Baby Hamster Kidney
- BWA : Burrows-Wheeler Aligner
- Cas proteins: CRISPR Associated proteins
- CHARMM: Chemistry at HARvard Molecular Mechanics
- ChIP: Chromatin ImmunoPrecipitation
- CHO: Chinese hamster ovary
- CIGAR: Concise Idiosyncratic Gapped Alignment Report
- CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats
- crRNA: CRISPR RNA
- DBS: Double Stranded Break
- DNA: Deoxyribonucleic Acid
- dsRBDs: Double-stranded RNA-Binding Domains
- EBV: Epstein–Barr virus

- eCLIP-seq: enhanced crosslinking followed by immunoprecipitation and deep sequencing
- EMA: European Medicines Agency
- EPO: Erythropoietin
- FDA: Food and Drug Administration
- FnCas9: *Francisella novicida* Cas9
- FPKM: Fragments Per Kilobase Million
- GATK: Genome Analysis Toolkit
- GLUL: Glutamine Synthase
- GO: Gene Ontology
- GROMOS: GRONingen MOlecular Simulation
- GTEx: Genotype-Tissue Expression
- HDR: Homology-Directed Repair
- Indels: Insertions and Deletions
- KEGG: Kyoto Encyclopedia of Genes and Genomes
- LbCpf1: *Lachnospiraceae* bacterium ND2006 Cpf1
- MD simulations: Molecular dynamics simulations
- MM: Molecular Mechanics
- MMEJ: Microhomology Mediated End Joining
- MSA: Multiple Sequence Alignment
- MSX: Methionine SulfoXimine
- NCBI: National Centre for Biotechnology Information
- NGS: Next Generation Sequencing
- NHEJ: Non-Homologous End Joining
- NmCas9: *Neisseria meningitidis* Cas9
- PAM: Protospacer Adjacent Motif

- PCA: Principal component analysis
- PCR: Polymerase chain reaction
- PDB: Protein Data Bank
- PME: Particle Mesh Ewald
- PTM: Post-Translational Modifications
- qPCR: quantitative real-time PCR
- RMSD: Root Mean Square Deviation
- RMSF: Root Mean Square Fluctuation
- RNA: Ribonucleic Acid
- RNase H: Ribonuclease H
- RNA-seq: RNA sequencing
- RRM: RNA recognition motif
- SaCas9: *Staphylococcus aureus* Cas9
- sgRNA: single guide RNA
- SpCas9: *Streptococcus pyogenes* Cas9
- ssODN: single-stranded oligodeoxynucleotide
- tracrRNA: trans activating crRNA
- tRNA: transfer RNA
- WGS: Whole-genome Sequencing

Chapter 1

Introduction

Using of computers has become an integral part of biological research since the late 20th century. With the sequencing of human genome a significant effort in the omics era is towards sequencing the genomes or proteomes of different organisms and thereby, scientists are compiling huge amounts of data at both these levels. Though, the sequencing data is available, a lot more effort needs to be put in annotating the data, understanding the interactions between them and relating them at the level of an entire organism. Besides, piecing this information together by understanding all possible interactions at the organism level, could only help us in gaining useful insights from this huge amount of data. However, the handling of data of this scale and extracting useful information by omitting the background noise is humanly impossible without the use of computers. Thus, with the advent of omics era scientists embarked on a new phase in biological research where handling big data was essential, thereby laying more emphasis on the need for computers to work hand-in-hand with experiments.

Computers can handle enormous data and help structuring them thereby, reducing the processing time significantly. Their importance is not only due to their ability to handle vast amount of data or keep track of them but also because they

can efficiently mine them for necessary conditions. Though, these might be affected by the scope of human errors in programming, but they will be unbiased towards the familiar which could frequently result from human inspection. Also, they help dealing with the complexities relating to compiling the huge scale of information, their interconnections and can easily handle different levels of data in isolation and combinations. Thus, making it very convenient to get a big picture. Also, presently a lot of efforts is invested in testing and training their ability based on available data and making predictions based on new data that has never been encountered before. These features can give us enormous power to make fast progress in biological research, and as a result making it inevitable to use computers in almost every field of biological research today.

1.1 Computational techniques and their basics

Computers help use a combination of mathematics, chemical and physical concepts along with informational techniques to solve biological problems. At the sequence level, analyzing the DNA and protein sequences, while more recently with the advent of next generation sequencing (NGS) technologies assembling and comparing entire genomes give insights relating to function, homology, and evolutionary relationships. Whereas, at the structure level, modeling and comparing the secondary plus tertiary structures reveal functional and catalytic information. Gene expression profiling resulting from microarray data or NGS helps gain a global picture of cellular functioning. Finally, at the level of integrating all the information comes the networks that include gene regulatory, metabolic, signaling and molecular interaction networks that are usually modeled as graphs to understand the biological entities at a systemic level and analyze them as not just individual systems but as a part of huge interconnected systems.

Two of the above mentioned computational techniques have been utilized by me in

this thesis for studying biological systems. Firstly, molecular modeling and simulation on macromolecules and secondly, developing computational pipelines to analyze huge RNA-Seq data generated from sequencing efforts. The following sections, would explain the background and basics of these computational techniques employed.

1.1.1 Molecular modeling and simulation

Molecular modeling and simulations is the art of studying structure and function of biomolecules by building models of the molecule's structure to represent them and performing computations on them. Building models of macromolecules is one of the preliminary steps towards understanding a system at a molecular level and is of great importance as it is known to give insights about its functioning. Although, structure can be determined using experimental techniques but this can be challenging in some cases. Thus, computational techniques come handy in such situations by building models. These can be broadly classified into three categories based on the approach used to build the model. Homology modeling and threading use structural information from similar molecules, while *ab initio* is template free approach and is based on first principles. Specifically, homology modeling is technique used for modeling structure of biomolecules, predominantly proteins, on the basis of an assumption that similar sequences give rise to similar folds and thus, similar structure.

As mentioned, homology modeling is used to model biomolecules that have homologous biomolecules with experimentally determined structures. The methodology involves aligning the target with an identified template (homolog), followed by constructing the initial model from the template structure and finally refining the model using appropriate techniques to avoid any steric clashes. α -lactalbumin is the very first protein structure predicted using homology modeling [1]. X-ray

structure of lysozyme served as the template for this prediction. Modeller, developed by Sali and Blundell in 1993, remains one of the most widely used packages for modeling [2].

With the molecular model ready, the next step is to gain insights about the functioning and that requires systematically exploring the structure and dynamics using simulation techniques. Simulation techniques can be broadly classified as based on *ab initio* quantum mechanics, and molecular mechanics (MM). *Ab initio* quantum mechanics based methods attempt to solve the electronic Schrodinger equation. These methods are probably the most accurate but are very computationally expensive plus time consuming and thus, can be employed only on small systems and for problems requiring precision to the maximum. The molecular mechanics based methods utilize classical mechanics, where the Born-Oppenheimer approximation is valid and the potential energy associated with a system is calculated as a function of its nuclear coordinates using force fields.

A simple force field definition involves two components bonded and non-bonded terms as shown in equation 1.1.

$$U_{\text{total}} = U_{\text{bonded}} + U_{\text{non-bonded}} \quad (1.1)$$

The bonded term involves contributions from bond stretching, angle bending and bond rotation due to torsions; while the non-bonded term usually consists of electrostatic and van der Waals interactions as shown in equations 1.2 and 1.3.

$$U_{\text{bonded}} = U_{\text{bond}} + U_{\text{angle}} + U_{\text{torsion}} \quad (1.2)$$

$$U_{\text{non-bonded}} = U_{\text{electrostatic}} + U_{\text{van der Waals}} \quad (1.3)$$

The bond stretching and angle bending terms are modeled using a harmonic potential, while the torsional potentials are usually cosine series expansion as defined

in the equation 1.4.

$$U_{\text{bonded}} = \sum_{\text{bonds}, i} \frac{k_{b,i}}{2} (b_i - b_{i,0})^2 + \sum_{\text{angles}, i} \frac{k_{\theta,i}}{2} (\theta_i - \theta_{i,0})^2 + \sum_{\text{dihedrals}, i} \frac{k_{\phi,i}}{2} (1 + \cos(n\phi_i - \delta_{i,0})) \quad (1.4)$$

where

$k_{b,i}$, $k_{\theta,i}$, and $k_{\phi,i}$ are the force constants associated with a particular bond length, angle and torsion respectively.

b_i , and $b_{i,0}$ correspond to the bond length and equilibrium bond length respectively.

θ_i , and $\theta_{i,0}$ correspond to the bond angle and equilibrium bond angle respectively. ϕ_i , and $\delta_{i,0}$ corresponds to the torsion/ dihedral angle and phase angle respectively.

The electrostatic interactions are modelled using a Coulomb potential, while the van der Waals is done using Lennard-Jones potential as defined in the equation 1.5.

$$U_{\text{non-bonded}} = \sum_{\text{atoms}, i,j} \frac{1}{4\pi\epsilon_0} \frac{q_i q_j}{r_{ij}} + \sum_{\text{atoms}, i,j} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^6 \right] \quad (1.5)$$

where

q_i , and q_j are partial charges of atoms i and j respectively.

ϵ_0 represents dielectric constant.

r_{ij} corresponds to interatomic distance.

ϵ_{ij} corresponds to the optimum energy between the two atoms that also represents the depth of the potential well.

σ_{ij} is the finite distance at which the van der Waals potential is zero.

The functional form of the force field alone is not sufficient but parameterization is also essential before the force field can be used. A lot of effort goes into deriving the parameters based on experimental observations on small molecules and this has given rise to different force fields available for use today based on the purpose. Some of the popular forced fields include Assisted Model Building and Energy Refinement (AMBER) [3], Chemistry at HARvard Molecular Mechanics (CHARMM) [4] [5], GROningen MOlecular Simulation (GROMOS) [6], etc.

These force fields developed can be used in conjugation with computer simulation methods to enable studying macroscopic systems and their properties by modeling small replications of them. Some of the most widely used simulation techniques used in molecular modeling include molecular dynamics and Monte Carlo methods. While Monte Carlo methods rely on repeated random sampling where each configuration depends only on its predecessor and is generated according to Boltzmann probabilities without any time evolution, the molecular dynamics methods are used to study the time evolution of motion of atoms and molecules in a system. In molecular dynamics, the successive configurations of a systems are obtained by integrating the Newton's laws of motion and thus the resulting trajectory indicates how the position and velocity of each particle in the system varies with time. This trajectory is determined by numerically solving the differential equations of Newton's second law of motion ($F = ma$):

$$\frac{d^2x_i}{dt^2} = \frac{F_{x_i}}{m_i} \quad (1.6)$$

where

m_i represents the mass of the particle

x_i represents the coordinates of the particle

F_{x_i} represents the force acting on the particle

In order to realistically capture the properties of a system, one needs to use a continuous potential where motion of all particles are coupled and force on each particle changes with change in its position or change in position of any other particle it interacts with. Under these circumstances, the equations of motion cannot be solved analytically and so they are integrated using a finite difference method. The integration is broken down into small time steps, δt and can be done using different algorithms. These algorithms assume that the positions, velocities and accelerations can be approximated as Taylor series expansions. The velocity Verlet method [7] is one of the common methods used where the position and velocity at time $t + \delta t$ is defined as in equations 1.7 and 1.8.

$$r(t + \delta t) = r(t) + v(t)\delta t + \frac{1}{2}a(t)\delta t^2 \quad (1.7)$$

$$v(t + \delta t) = v(t) + \frac{1}{2}\delta t[a(t) + a(t + \delta t)] \quad (1.8)$$

Thus, using this approach based on the initial position and velocity, the force acting on each particle in a system at time t is computed based on the force field as the vector sum of its interactions and this is used to compute the acceleration on the particle. This acceleration in combination with the position and velocity at time t can be used to compute the new position and velocity of the particle at time $t + \delta t$ by using one of the different integration algorithms [for eg: like the one shown in equations 1.7 and 1.8]. Then, the cycle is repeated by calculating the new force acting on the particle at time $t + \delta t$ that leads to computing new positions and velocities at time $t + 2\delta t$ and so on. The figure 1.1 illustrates the algorithm of MD simulations.

Figure 1.1: This figure summarizes the general protocol/ algorithm of molecular dynamics simulations.

1.1.2 Analyzing high-throughput sequencing data

Sequencing technologies have improved from the simple Sanger or Maxam-Gilbert sequencing to a wide-range of high-throughput methods available today. This improvement in technology has expanded the scope of sequencing from just a short stretch of DNA to whole-genome sequencing (WGS), RNA sequencing (RNA-Seq), targeted sequencing, protein-binding regions (ChIP seq), etc. In order to be able to get meaningful insights from the sequencing efforts one needs to analyze the huge data generated by these techniques and this can be done by developing computational pipelines. But, no one pipeline exists for all analysis as the pipeline varies depending on what is sequenced and what information one expects to gather from the data. Though custom pipelines need to be developed for extracting information from the sequencing data, certain preliminary steps are common as

shown in figure 1.2 and described below.

Figure 1.2: A schematic of the common steps involved in analyzing NGS data.

The very first step in analyzing the NGS generated data is checking if the raw reads contain primers and processing to remove the primers. Also, considering the sequencing accuracy is generally poor at the ends of the read sequences, one might want to trim the edges to improve the quality and confidence of the sequenced read. Further, once the reads are trimmed, they need to be mapped to a reference library that could be a genome or a transcriptome. There are a lot of available tools for mapping that include BWA [8], STAR [9] [10], Top Hat [11], etc. Uniquely mapped reads are then identified using MAPQ score and filtered to be processed further. These reads are then subjected to local realignment to reduce the number of mismatches across all reads. This is essential because reads are considered separately by genome aligners and their scoring techniques have limitations while trying to align reads with insertions and deletions (indels). On

the other hand, the realigner considers reads spanning a given region together and thus, can detect the presence of an indel with greater confidence. Another common step that the reads are subjected to is recalibration which is done for the base quality score. Base quality scores are assigned to each base in the sequenced read that indicates the per-base estimate of error emitted by the sequencing machines. Variant calling methods and further post-processing might rely on these quality scores that may be over- or under-estimates of the actual score due to systematic technical error. Thus, these scores are adjusted during the recalibration process. Both the local realignment and recalibration steps can be performed using Genome Analysis Toolkit (GATK) [12] [13] [14].

The downstream processing after this depends on the application that one is studying and may vary from differential gene expression calculations, variant analysis (RNA editing calculations), identification of insertions and deletions, etc that are explored in this thesis. Thus, the computational pipeline can be developed depending on the particular application that one intends to explore.

1.2 Editing in human cell

1.2.1 DNA editing: CRISPR

CRISPR (clustered regularly interspaced short palindromic repeats) has its origins as an adaptive immune system in prokaryotic organisms: bacteria and archaea [15]. These are short DNA fragments that are incorporated into the prokaryotic genome from the foreign genetic elements that have previously infected the prokaryote and serve as a memory to destroy similar genetic elements during subsequent infections, thus forming an integral part of the prokaryote's acquired immune system by functioning on the self-nonself discrimination principle [16]. Cas (CRISPR associated) proteins are involved in processing the CRISPR loci and guided by the transcribed

CRISPR sequences to recognize and cleave specific region in the genome that are complementary to the CRISPR sequence [17] [18] [19] [20] [21] [22].

Discovery of CRISPR-Cas

The story of CRISPR discovery started with the identification of new type of repeat sequences in prokaryotes in 1987 [23]. In 2000, Mojica et al. described these repeats as clusters of short repeated palindromic sequences (20-40bp) separated by intervening sequences (20-60bp) found in most archaea and bacteria [24]. In 2002, Jansen et al. coined the term CRISPR for these repeats [25]. These repeats were found to lie in the vicinity of a family of genes known as a 'DNA repair system' that are now called Cas proteins [26]. The origin of the spacer remained unknown till 2005 when three groups independently revealed their bacteriophage origin [27] [28] [29]. In 2006, Makarova et al. suggested CRISPR-Cas system to be a prokaryotic RNA interference-based immune system and later in 2007, Barrangou et al. demonstrated that CRISPR-Cas system functions as the adaptive immune system in archaea and bacteria against bacteriophages [22] [30]. Following this, attempts at understanding the molecular mechanism of action, structural and functional characterization of the system and its components gave rise to the CRISPR-Cas9 editing system.

Mechanism of CRISPR-Cas

The functioning of the CRISPR-Cas system is broadly described in three stages:

- **Spacer acquisition:** During this phase, the Cas proteins recognize and integrate a short piece of foreign DNA into the CRISPR region [31] [32] [33] [34]. Two proteins, Cas1 and Cas2, that are conserved through most known CRISPR types are involved in this phase [33].
- **Expression and processing (crRNA generation):** CRISPR repeat-spacer ar-

ray is transcribed into a long pre-crRNA and then processed into small individual mature crRNAs composed of a repeat portion and an invader targeting spacer portion [35] [36].

- Interference or silencing: crRNA combines with the Cas proteins to form an effector complex that is then directed by the crRNA to recognize the target sequence in the foreign invasive nucleic acid and thereby, induce sequence-specific cleavage [37].

The proteins involved in the expression and processing stage and the interference or silencing stage vary from one CRISPR-Cas system to another and the same enzyme may also participate in both these stages.

Classification of CRISPR-Cas systems

CRISPR-Cas systems known so far are classified into two classes based on the effector complex design principles [38] [39] [40]. Class 1 consists of the common and diversified type I and type III along with the rare type IV [41]. The effector module in this case comprises of more than one protein and varies with the CRISPR-Cas system. Class 2 consists of type II, type V and type VI that comprises a single large multidomain protein being the effector complex [41]. The phylogeny of the Cas1 protein that is the most conserved protein across the different types also agrees with this classification system [38].

Application of CRISPR-Cas systems

The ability of these Cas proteins to cleave a genomic region guided by a RNA (ribonucleic acid) molecule has been leveraged to develop the CRISPR-Cas genome editing toolkit with a wide range of applications. Class 2 CRISPR-Cas systems have been the choice for developing the new generation genome editing technologies as a result of their simple effector complex that consists of just one protein. Thus,

as a powerful RNA-guided DNA targeting platform they are used for genome editing, genome imaging, transcription perturbation, and epigenetic modulation [42] [43] [44]. The active Cas enzyme has been used for genome editing with the application of appropriate DNA repair mechanisms (NHEJ or HDR), while the mutant Cas9 devoid of cleavage activity (dCas9) has been used for perturbation of gene expression [45] [46] [47]. Further, CRISPR technology has been used for editing micro-RNA and long-noncoding RNA in plants after the identification of a CRISPR enzyme that can be used to edit RNA [48]. A list of Cas enzymes that are being utilized for DNA and RNA editing is listed in Table 1.1

Challenges of using CRISPR-Cas systems

Although, CRISPR technology has seen enormous progress in terms of application, yet some prominent challenges including off-target effects, restriction for targeting range due to PAM requirement, efficacy, large size and delivery issues limit safe usage and extensive use of this technology for therapeutic purposes. Even as roughly a 20 base long spacer region on the sgRNA guides the CRISPR complex to the target site on the genome, many identical or highly homologous regions to the spacer can be found on a large genome that can be mistakenly recognized by the complex and thus, introducing unwarranted changes in the genome. These modifications, termed as off-targets, can be lethal. A host of studies have been designed to understand, estimate and alleviate the off-target effects as detailed in [63] [64] [65] [66]. The targetable space on the genome increases with the discovery of more CRISPR enzymes by adding new PAM sequences and also with efforts to decipher altered PAM specificities by mutating the CRISPR enzymes [67] [68] [69] [70] [71] [72]. The large size and off-target effects compound the delivery issues and technologies for effective delivery will enable safe usage of the technology as detailed in reviews [73] [74] [75].

Table 1.1: List of Cas effector enzymes acting on DNA and RNA

SNo.	Enzyme	Species	Size (in amino acids)	Purpose	References
1	Cas9	<i>Streptococcus pyogenes</i>	1368	DNA editing	[49]
2	Cas9	<i>Staphylococcus aureus</i>	1053	DNA editing	[50]
3	Cas9	<i>Neisseria meningitidis</i>	1082	DNA editing	[51]
4	Cas9	<i>Francisella novicida</i>	1629	DNA editing	[52]
5	Cas9	<i>Actinomyces naeslundii</i>	1101	DNA editing	[53]
6	Cas9	<i>Campylobacter jejuni</i>	984	DNA editing	[52]
7	Cas9	<i>Streptococcus thermophilus</i>	1409	DNA editing	[54]
8	Cpf1 or Cas12a	<i>Lachnospiraceae bacterium</i>	1231	DNA editing	[55]
9	Cpf1 or Cas12a	<i>Acidaminococcus sp.</i>	1307	DNA editing	[55]
10	Cpf1 or Cas12a	<i>Francisella novicida</i>	1300	DNA editing	[56]
11	Cpf1 or Cas12a	<i>Moraxella bovoculi 237</i>	1419	DNA editing	[57]
12	Cas13a	<i>Leptotrichia wadei</i>	1182	RNA editing	[58]
13	Cas13a	<i>Leptotrichia shahii</i>	1389	RNA editing	[59]
14	Cas13a	<i>Lachnospiraceae bacterium</i>	1437	RNA editing	[58]
15	Cas13a	<i>Leptotrichia buccalis</i>	1159	RNA editing	[60]
16	Cas13b	<i>Prevotella sp.</i>	1127	RNA editing	[61]
17	Cas13b	<i>Bergeyella zoohelcum</i>	1224	RNA editing	[61]
18	Cas13d	<i>Ruminococcus sp.</i>	930	RNA editing	[62]
19	Cas13d	<i>[Eubacterium] sinuatum</i>	954	RNA editing	[62]

Studying this system and the related technology offers great scope for research that can add value to the growing technology. In this thesis, the functioning and characterization of this system is done by designing different studies to contribute to our understanding about the system and the technology.

1.2.2 RNA editing

RNA editing is site-specific nucleotide change within the RNA after transcription for fine-tuning gene products [76]. The first reports of identification of RNA editing was in trypanosome mitochondrial mRNA where insertions or deletions of uridine were identified that led to generation of functional proteins for this protozoa [77]. Since then, a number of RNA editing phenomena have been identified that including modifications such as cytidine to uridine and adenosine to inosine by deamination, as well as insertion or deletion of nucleotide like the one found in [77]. Of the different RNA editing processes the most prevalent one in higher eukaryotes is adenosine to inosine (A-to-I) conversion [78] [79] [80] [81]. A-to-I editing involves the deamination of an adenosine and this deamination reaction is catalyzed by a family of enzymes called adenosine deaminases acting on RNA (ADAR) [79] [82] [83].

ADAR family

ADAR family of enzymes recognize double-stranded RNA structures and catalyze the deamination of select adenosines to inosines. Three members of the ADAR family have been identified in mammals namely ADAR1, ADAR2 and ADAR3 [84] [85] [86] [87] [88] [89] [90]. They contain a C-terminal deaminase domain and two or three double-stranded RNA-binding domains (dsRBDs) in the N-terminal that are conserved in the three enzymes as shown in figure 1.3 [91]. ADAR1 also contains Z-DNA binding domain, while ADAR3 contains R-domain shown to be

important for single-stranded RNA binding. Mammalian ADAR1 and ADAR2 are ubiquitously expressed in many tissues, while ADAR3 is restricted to brain [82] [85] [92] [86] [87] [88] [89] [93]. ADAR1 and ADAR2 have distinct substrates and have demonstrable editing activity, while ADAR3 has no known editing activity.

Figure 1.3: A schematic of the domain architecture of ADAR family of proteins.

Consequence of RNA editing and disease implications

A-to-I modification in the RNA leads to physiological change as a result of nucleotide substitution as the cellular machinery interprets I as G [94]. Therefore, RNA editing can give rise to protein isoforms, influence RNA stability and affect RNA splicing. Editing in the coding region could lead to codon changes that results in new protein isoforms and diversification of protein functions as is typically found in neurotransmitter receptors and ion channels expressed in the central nervous system [95]. Editing at the splice sites/ protein coding regions can impact the RNA splicing: editing 3' splice site can modify AG -> IG read as GG by the cellular machinery leading to loss of splice site, editing at a nonsplice site AA -> AI can create a splice site as it is read as AG by the cellular machinery, and finally 5' splice donor site can also be created by editing of a site with AU -> IU that is read as GU by the cellular machinery. ADAR2 pre-mRNA transcript is an example of splice site editing where the -1 position is edited (AA -> AI) pro-

ducing alternative 3' splice site by ADAR2 protein [96]. Alteration in microRNA targets, impact on microRNA maturation are other consequences of RNA editing observed.

Dysfunction of the RNA editing mechanism is implicated in many diseases including neurological disorders and multiple cancer types [97] [98] [99] [100] [101] [102] [103]. For example, underediting of Q/R site of GluR-B pre-mRNA causes death of motor neurons in amyotrophic lateral sclerosis (ALS) patients and is also shown to be responsible for neuronal death during ischemia resulting from cardiac arrest [99] [104] [93]. Several studies have shown that transcriptomic and proteomic diversity in cancer can be attributed to RNA editing where the overall editing levels correlate with the expression of the ADAR enzymes [105] [106] [107] [108]. Although, correlation between editing levels and tumorigenic potential of cancer is unclear as there are reports for both increased and decreased editing linked to enhanced tumorigenesis [105] [106] [107] [109] [110] [111]. There are also reports linking RNA editing to diabetes mellitus and some viral infections including those caused by influenza virus, measles, human immunodeficiency virus-1 [112] [113] [114].

Thus, the wide-range of consequences and disease implications due to RNA editing motivate the need to better understand the mechanism of editing and its regulatory control points.

1.3 Scope and objective of this work

In this thesis, I present my work which focuses on my efforts in working closely with experimentalists in order to enhance our understanding of editing (DNA and RNA) in human cells using the computational techniques and contributing to the field. Three main aspects of the editing system are studied that are critical for understanding and utilizing a system for different applications. Firstly,

understanding the functioning of the enzymes involved is an important aspect for appropriate usage of the system and thus, I studied the functioning of the enzymes involved in DNA editing system (CRISPR-Cas) from both structural perspective and through the enzymes activity. Secondly, showing applicability of the system allows to exemplify its usage and thus, as an application of the CRISPR system, knock-cell lines developed using this technology was utilized for glycoprotein production. Computational pipeline is designed to characterize the knock-cells for this application. Finally, the functioning of enzymes is tightly regulated within the biological context by a host of other factors and identifying those regulators are key for enhancing our understanding of these system and proper applicability. Therefore, a computational study is designed in conjugation with experiments to identify one of the regulators of RNA editing system. The figure 1.4 summarizes the problems addressed in this work.

Further, two main computational methods are employed. First being computational modeling and simulation techniques that I have used to gain insights into the protein structure and functioning. Second includes developing computational pipelines to analyze the huge data resulting from next generation sequencing techniques to gain meaningful insights. At the bottom of figure 1.4 these methods are colored coded and the biological problems addressed using these techniques colored coded accordingly. In addition, in one of the studies, computational pipeline has been used for identification of patterns. This could potentially be used for developing and training networks using machine learning techniques that can be utilized for predictive purposes, in line with the present trend. But, in this current work I have not explored this aspect.

Figure 1.4: This figure summarizes the biological problems addressed in this thesis in order to enhance our understanding of DNA and RNA editing in human cells. The biological problems are color coded depending on the computational techniques used in the study.

1.4 Conclusion

Biological editing systems (working both at the level of DNA and RNA) are studied in this work by employing a host of computational techniques that include structural modeling for inspecting the protein structure and by developing customized pipelines for extracting useful information from huge data resulting from NGS. The outcome from employing these computational techniques are put in the biological context to add to our wealth of knowledge on these two systems, enable better understanding of them and powerful usage of the technologies resulting from them.

Chapter 2

Uncovering the role of highly conserved residues in the HNH nuclease domain of Cas9 using molecular dynamics simulations

Cas9, the effector protein of type II CRISPR-Cas systems, is a genome editing tool with wide-ranging applications. It consists of two nuclease domains, each of which cleaves a different strand of the DNA, thereby resulting in a double-stranded break. Gaining insights into the functioning of the nuclease domains may guide us in tuning the active sites so as to improve the efficacy and specificity of the enzyme. Here, the HNH nuclease domain of Cas9 is investigated using molecular dynamics (MD) simulations and thereby, identify likely roles for the highly conserved residues in the neighbourhood of the active site. Results suggest that these highly conserved residues are important for maintaining the structure of the $\beta\beta\alpha$ -motif and enable proper positioning of the DNA, thereby assisting catalysis.

2.1 Background

The type II CRISPR-Cas system is referred to as the minimal CRISPR-Cas system as its surveillance complex comprises of just one protein, Cas9, and a mature crRNA [115] [49] [116]. Cas9 is an RNA-guided DNA endonuclease and the RNA that guides it is formed by a hybrid structure of the mature crRNA and tracrRNA (trans activating crRNA) or a chimeric single guide RNA (sgRNA) [49] [117]. In order for Cas9 to cleave the DNA, a sequence motif called the protospacer adjacent motif (PAM) must be present at the target locus [49]. As a result of a relatively simple surveillance complex that can be readily modified, Cas9 has been exploited to engineer genomes for various applications, including genome editing, transcriptional control, and genetic screening [118].

Structural studies on the Cas9 protein from *Streptococcus pyogenes* (SpCas9) [119] [53], *Actinomyces naeslundii* (AnCas9) [53], *Staphylococcus aureus* (SaCas9) [120], and *Francisella novicida* (FnCas9) [121] have shown that the overall domain organization and assembly is similar for all the Cas9 proteins, although they vary in their sequence lengths considerably. It consists of a bi-lobed architecture (recognition and nuclease lobes) with a central channel between them accommodating the DNA [119] [122] [123]. The recognition lobe consists of several recognition domains and a long α -helix abundant in arginine (referred to as bridge helix), while the nuclease lobe consists of two nuclease domains (RuvC and HNH) and the PAM-interacting domain.

The RuvC and HNH nuclease domains cut opposite strands of the DNA to generate a double-stranded break. While, the RuvC domain shares a similarity with the ribonuclease H (RNase H) fold and employs a two-metal-ion mechanism to cleave the DNA, the HNH domain is characterized by the $\beta\beta\alpha$ -motif that utilizes a one-metal-ion mechanism for catalysis [119] [124]. The reason behind the utilization of different mechanisms for cutting DNA between the two nuclease domains is

currently unclear. Additionally, while the RuvC domain is assembled from three separate motifs, the HNH domain is composed of a single continuous globular fold, as indicated in figure 2.1. Three amino acids within the $\beta\beta\alpha$ -motif, namely D839, H840, and N863, have been identified as the catalytic residues of the HNH domain [119]. The metal ion, Mg^{2+} , has been found to be coordinated to D839 and N863, while it has been proposed that H840 is responsible for initiating the nucleophilic attack that results in DNA cleavage [53]. Experimental studies have also shown that mutating H840 or N863 results in loss of function of the domain [119].

Figure 2.1: The domain organization of SpCas9 showing the RuvC domain split into three motifs by long insertions between them and the HNH domain present as one continuous motif in the sequence space.

Although, the key catalytic residues for the RuvC and HNH domains have been identified by comparison to other nucleases containing similar folds, most of the mechanistic details underpinning the functioning of the active sites remain largely unknown [119]. Recently, MD simulations have been employed to probe the conformational changes of Cas9 during the gene editing process [125] [126]. In one study, Palermo et al. demonstrated a role for the non-target DNA strand in priming the HNH domain for catalysis [125]. In another study, Zuo et al. found that Mg^{2+} ions in the RuvC domain helped to induce conformational changes that resulted in the active site of the domain being brought sufficiently close to the non-target DNA strand for cleavage to occur [126]. However, the authors, who performed all-atom MD simulations of SpCas9 complexed with DNA and its sgRNA scaffold, were unable to sample the active conformational state of the HNH domain, despite the relatively long time scale ($\sim 1.5 \mu s$ in [125] or 200-300 ns in [126]) of their simulations .

In this study, the active state of the HNH domain is investigated, particularly by concentrating on the highly conserved residues located within the $\beta\beta\alpha$ -motif. A comparison of existing Cas9 structures with and without a divalent cation in the HNH domain reveals a local rearrangement of atoms around the active site near the catalytic N863 residue as indicated in figure 2.2 [53] [120] [123]. Hence, MD simulations are carried out on the HNH domain and found to reveal novel roles for the highly conserved catalytic and non-catalytic residues in positioning the DNA and in maintaining the $\beta\beta\alpha$ -motif. The findings from this study may guide us in tuning the active site of the HNH domain to improve the enzyme's efficacy and specificity in the future.

Figure 2.2: The $\beta\beta\alpha$ -motif of AnCas9 (PDB ID: 4OGE in green) and SpCas9 (PDB ID: 5F9R in blue) superposed to show the variation in position of the catalytic asparagines (shown in sticks). The Mg^{2+} ion present in the AnCas9 structure, but absent in the SpCas9 structure, is represented as a bead in magenta in the active site.

2.2 Methodology

2.2.1 Sequence analysis of type II Cas9

Sequences of type II Cas9 from different organisms were extracted from Swiss-Prot [127] and their distinct full-length sequences were aligned using Clustal

Omega [128] [129] [130]. Jalview [131] was used to visualize the Multiple Sequence Alignment (MSA).

2.2.2 Initial structure preparation for MD simulations

The HNH domain of SpCas9 was simulated with the ion (Mg^{2+}) in the active site and a short single-stranded DNA (5' AGCTCC 3') in the catalytic site as shown in figure 2.3. The HNH domain was extracted from a structure of SpCas9 taken from the Protein Data Bank (PDB) [132] (PDB ID: 5F9R chain B, resolution 3.4 Å [123]) comprising residues 781 to 906. The N- and C-terminal ends of the protein were capped using C-acetyl and N-methyl caps as only the HNH domain of SpCas9 was extracted from the full-length SpCas9 structure. The ion is absent in the active site of the HNH domain in 5F9R, while the DNA is ~ 30 Å away from the active site [123]. Hence, from this structure with no ion and no DNA, a structure with an ion complexed to the two catalytic residues (D839 and N863) is modeled. Upon inspecting the Cas9 structures available in the PDB, it was observed that in the presence of an ion in the active site, residue D839, which is in a β -strand, does not change conformation, while the helix containing the catalytic residue N863 (numbered as in SpCas9) adopts a different conformation as shown previously in figure 2.2. Therefore, the helix in 5F9R was modeled using the structure of AnCas9 (PDB ID: 4OGE chain A [53] containing a Mg^{2+} ion in the active site) as template to generate the active state of 5F9R. Next, the Mg^{2+} ion in the active site of the active state was modeled using the same AnCas9 structure. Furthermore, a short segment of DNA (6 nt long) was modeled using the crystal structure of T4 Endonuclease VII (PDB ID: 2QNC chain A [133]), as it shares the same nuclease fold as the HNH domain, is crystallized with DNA in the active site, and has a Mg^{2+} ion coordinated in the active site. The system with the protein, ion, and DNA (or just protein) was then solvated in a water

box having box edges at least 12 Å away from the protein/ DNA, followed by neutralization of the system using counter-ions (Na^+). In addition, two control simulations lacking the ion and DNA were also performed. One of them was the HNH conformation taken as is from the PDB (PDB ID: 5F9R chain B) and was referred to as the inactive state, while the other was the starting active state HNH conformation without the ion and DNA.

Figure 2.3: The HNH domain of SpCas9 with the Mg^{2+} (in magenta) and a short single-stranded DNA (represented as sticks with element wise color scheme – C in green, N in blue, O in red, and P in orange) in the catalytic site. The catalytic residues D839, H840 and N863 are shown in sticks with element wise color scheme (C in white, N in blue and O in red). The coordination of the catalytic residues, D839 and N863, with ion is denoted by black lines.

Simulations were performed using the AMBER 14 package with the AMBER ff14SB force field [3] [134] [135]. The divalent cation was treated using the 12-6-4 LJ-type non-bonded model as it best reproduces the experimental hydration free energies, ion-oxygen distances, and coordination number values at the same time [136]. Explicit waters were modeled using the TIP3P water model [137]. The SHAKE algorithm was used to constrain the covalent bonds involving hydrogen atoms, enabling a 2 fs timestep during the simulations. A cut-off of 8 Å was used for the non-bonded interactions and long-range electrostatics were treated with Particle Mesh Ewald (PME) beyond the cut-off [138]. Minimization was performed using 2500 cycles of steepest descent followed by another 2500 cycles

of conjugate gradient. The system was gradually heated to a temperature of 300 K prior to equilibration. Three replicate simulations of each setup, seeded by different random numbers, with a 5 ns equilibration followed by a production run of 100 ns were carried out. The highly conserved residues (P843 and K866) in the $\beta\beta\alpha$ -motif of the HNH domain were mutated to alanine and subjected to MD simulations in order to gather more insights into their possible roles. The Pymol mutagenesis plugin was used to model these mutants [139]. Further, Mg^{2+} , the ion in the active site is replaced by Zn^{2+} in the active state HNH conformation in the presence of DNA in the active site and the system is subjected to simulation as well. A list of simulations performed is summarized in Table 2.1.

Table 2.1: List of simulation performed in the study

System	Number of replicates	Simulation time per replicate
HNH domain + Ion (Mg^{2+}) + DNA – Active state	3	100 ns
HNH domain (PDB ID: 5F9R chain B) – In-active state	3	100 ns
Starting active state HNH domain with the ion and DNA removed	3	100 ns
P843A mutant + Ion (Mg^{2+}) + DNA	3	100 ns
K866A mutant + Ion (Mg^{2+}) + DNA	3	100 ns
HNH domain + Ion (Zn^{2+}) + DNA – Active state with Zn^{2+}	3	100 ns

2.2.3 Structural Analysis

All analysis was performed using the cpptraj module from AMBERTools 14 [134]. Only the last 80 ns of each simulation was considered for analysis. The trajectories were visualized using VMD [140], while Pymol [139] and Chimera [141] were used for crystal structure visualization and for preparing the figures.

2.3 Results

2.3.1 Sequence analysis of Cas9 reveals minimal sequence conservation

Manually curated sequences of type II Cas9 enzymes are retrieved from Swiss-Prot [127] and aligned using Clustal Omega [128] [129] [130]. Despite the large size of Cas9, only 19 residues show 100% identity across all the Cas9 proteins examined as depicted in figure 2.4. In the bridge helix, R66 and R69 (numbered according to SpCas9) lie in close proximity to, and form polar contacts with the sgRNA, based on the published crystal structures [123] [124]. Three residues in the REC lobe (E418, Q426, and K510) are perfectly conserved, with K510 lying in close proximity to the bound DNA. In the RuvC nuclease domain, seven amino acids (D10, G12, G17, P731, H983, D986, and A987) are 100% identical in all the Cas9 sequences, including three of the catalytic residues (D10, H983, and D986). The fourth catalytic residue of the RuvC domain is not identical in all sequences, but is nevertheless constrained to be either a D or an E, which have similar side chains.

In the HNH nuclease domain, there are seven residues that show 100% identity across all Cas9 proteins, namely Y823, D839, H840, P843, N854, N863, and K866, with five of them lying within the core of the $\beta\beta\alpha$ -motif as shown in figures 2.5 and 2.2.

Furthermore, four of these seven residues are also found to be identical across the entire Pfam family HNH_4, which comprises HNH nuclease domains from a larger set of CRISPR-related proteins as shown in figure 2.6. The catalytic residue D839 is one of the three residues that shows variation across the Pfam family and is replaced by a conserved mutation to E in some of the proteins, while the aromatic amino acid Y823 is the other and is substituted by another aromatic amino acid

[illegible]

816	SAADNAHMOQ	K I K I T E P V E D N K D K I I L S A K A Q R L P A I T R I V D G A V K K M A T I L A K N I V D O N W Q I K Q V L S A K H Q L H I P I I T E S N A F E F E P A L A D V K G S L K D R R K K A L E R I	927
485		PAEA I N A P V G N P S V D R T L K I V G R Y L S A V E S M W G T P E V I H V E H V R D G F T E R M A D E R O K A N R R Y N D N Q E A M K I Q R D Y G - K E	545
456		P T P R I G E P V G N A P V A D R L K T V R A M L E S A T K T W G A P E R V I I E H V R G F V T E R M A D E R O K A N R R A A R N A K L F Q E W K E L N - V Q	538
742		I K E V V K S L P G S P A I K G I L Q S I K I V D E L V K Y M G G R K P E S I V V	824
721		I K E V V K S L P G S P A I K G I L Q S I K I V D E L V K Y M G G R K P E S I V V E H V R G F V T E R M A D E R O K A N R R A A R N A K L F Q E W K E L N - V Q	803
724		I Q S I V A D L A G S P A I K G I L Q S I K I V D E L V K Y M G G R K P E S I V V	802
720		L H E H I A N L A G S P A I K G I L Q S I K I V D E L V K Y M G G R K P E S I V V	802
720		L N Q V S D I A G S P A I K G I L Q S I K I V D E L V K Y M G G R K P E S I V V	799
439		N E T Y V D E T V P N V L R A I K E Y K V L N A L K Y G K V H K I	515
460		PAQ E I R N P V L R T L S O A R K V I N A I I R O Y G S P A R V H I	534
468		PAQ E I R N P V L R T L S O A R K V I N A I I R O Y G S P A R V H I	543
468		PAQ E I R N P V L R T L S O A R K V I N A I I R O Y G S P A R V H I	543
437		P T T L V D E F I L S P V K R S F I Q S I K V I N A I I K Y G L P N D I	511
469		DEK L L T E E I Y N P V A K S V R O A I K I V N A A I K E Y G D F D N I V	553
928		N R I K E F A K G I S A V G A N T I D G D F D G A K E E L D H I	1043
546		G Y I S R G D I V R L D A L Q G C A L I G G T I I G Y H T C O L D H I V D A G P G S N	644
537		G P S R A D L W R Y Q S V G N C O C A Y C G S P I T F S N S E M O H I V A G O G S T	635
804		K L S K I D N N A L Q N D R L Y L Y L O N G D M T G D D L D I R L S N Y D I D H I I D A F L K D	520
825		K L S K I D N N A L Q N D R L Y L Y L O N G D M T G D D L D I R L S N Y D I D H I I D A F L K D	899
903		P T D N Q E L R N N R L Y L Y L O N G D M T G D D L D I R L S N Y D I D H I I D A F L K D	895
900		P V E N T O L N E K L Y L Y L O N G D M T G D D L D I R L S N Y D I D H I I D A F L K D	892
800		P V E N S O L Q N D R L Y L Y L O N G D M T G E E L D I D Y L S Q Y D I D H I I D A F L K D	892
515		L G L K I N S K N I L K L R F Q E O K E F C A Y S G E K I T S O L Q D E K M L E I D H I I D A F L K D	907
544		F S E P K S D I L K F R L Y E Q O H G K L Y S G K E I N L R L N E K G Y V E I D H A L F S R T W D	632
544		F V G E P K S D I L K L R L Y E Q O H G K L Y S G K E I N L R L N E K G Y V E I D H A L F S R T W D	640
512		T G K E N A K Y L E K I L H D M Q E G K L Y S L E A I P L E D L N N P F A Y E V D H I P R S V S F D	609
554		S V F H G H K O L A T K I R L W H Q Q G E R C L T G K T I S I H D L I N N S O F E V D H I L E L S I T F D	649
1044		G N Y R S F I N L T P Q E A K F R H L F L A D E N P I K Q A V I A N N R N R I F V N G T O R Y F A E V L A N N I Y I R A K K E N L N T D K I	1156
645		T S S E D L R L K K E V I A R L R R T Q E D P E I D E R S M E S	730
636		R S T D F K F T K A V E R F O R A T W D E E I D A R S M E S	721
921		S - Q R K F D N L T K A E R G G L L P E D K A G F I Q P O L V	1005
900		S - Q R K F D N L T K A E R G G L L P E D K A G F I Q P O L V	984
996		S - K R K F D N L T K A E R G G L T E A D K A R F I H Q P O L V	980
993		S - K R K F D N L T K A E R G G L T E A D K A R F I H Q P O L V	977
993		S - K R K F D N L T K A E R G G L T E A D K A R F I H Q P O L V	977
606		A K N L P T K Q K R I L D K N Y K D K E Q K N F K D R L N	701
633		G S O C S A A K Q R L L T O V I D N K F I D E R L N	709
641		T S R F P R S K Q R I L L O K F D E D G F K E R N L N	717
641		T S R F P R S K Q R I L L O K F D E D G F K E R N L N	717
610		N L A K G R I S K T K E Y L L E E D I N R F S V Q K D F I N R L V	717
650		V R E S K T L S N K K K E Y L L T E E D I S K F D V R K K F I E N L V	733
1157		Q A S Y S L L D A M L A F C I A A D E H R N D G S I G L E I D K N Y S L Y P L D K N T G E V F T K D I F S O I K I T D	1273
731		D R R H A V C A S V A L M E A S V A K T I A E R S L R Q E O R L T G K E D T W K Y T G S T	827
732		D R R H A V C A S V A L M E A S V A K T I A E R S L R Q E O R L T G K E D T W K Y T G S T	827
1006		I N D F H A I D A Y L N A V S A L L K K Y P K L E P E F	1101
946		I N D F H A I D A Y L N A V S A L L K K Y P K L E P E F	1061
941		V N D Y H A I D A Y L N A V S A L L K K Y P K L E P E F	1061
978		I N D Y H A I D A Y L N A V S A L L K K Y P K L E P E F	1069
978		I N D Y H A I D A Y L N A V S A L L K K Y P K L E P E F	1069
710		N N H L H A I D A Y L N A V S A L L K K Y P K L E P E F	704
710		N N H L H A I D A Y L N A V S A L L K K Y P K L E P E F	704
718		N D R H A I D A Y L N A V S A L L K K Y P K L E P E F	780
718		N D R H A I D A Y L N A V S A L L K K Y P K L E P E F	780
696		K Y H A I D A Y L N A V S A L L K K Y P K L E P E F	768
734		T Y H H A V D A L I I A A S - Q L N L - W K K	807
516		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	516
517		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	517
518		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	518
519		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	519
520		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	520
521		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	521
522		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	522
523		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	523
524		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	524
525		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	525
526		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	526
527		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	527
528		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	528
529		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	529
530		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	530
531		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	531
532		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	532
533		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	533
534		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	534
535		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	535
536		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	536
537		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	537
538		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	538
539		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	539
540		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	540
541		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	541
542		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	542
543		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	543
544		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	544
545		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	545
546		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	546
547		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	547
548		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	548
549		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	549
550		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	550
551		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	551
552		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	552
553		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	553
554		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	554
555		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	555
556		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	556
557		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	557
558		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	558
559		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	559
560		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	560
561		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	561
562		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	562
563		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	563
564		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	564
565		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	565
566		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	566
567		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	567
568		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	568
569		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	569
570		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	570
571		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	571
572		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	572
573		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	573
574		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	574
575		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	575
576		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	576
577		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	577
578		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	578
579		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	579
580		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	580
581		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	581
582		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	582
583		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	583
584		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	584
585		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	585
586		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	586
587		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	587
588		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	588
589		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	589
590		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	590
591		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	591
592		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	592
593		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	593
594		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	594
595		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	595
596		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	596
597		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	597
598		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	598
599		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	599
600		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	600
601		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	601
602		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	602
603		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	603
604		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	604
605		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	605
606		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	606
607		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	607
608		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	608
609		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	609
610		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	610
611		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	611
612		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	612
613		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	613
614		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	614
615		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	615
616		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	616
617		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	617
618		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	618
619		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	619
620		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	620
621		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	621
622		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	622
623		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	623
624		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	624
625		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	625
626		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	626
627		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	627
628		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	628
629		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	629
630		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	630
631		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	631
632		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	632
633		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	633
634		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	634
635		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	635
636		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	636
637		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	637
638		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	638
639		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	639
640		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	640
641		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	641
642		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1 6 2 9	642
643		Q O Q 3 Y 7 C A S 9 _ F R A T N V I - 1	

[illegible]

Figure 2.4: Multiple sequence alignment of all the manually curated Cas9 proteins retrieved from Swiss-Prot with highly conserved residues highlighted in blue.

```

sp|A0Q5Y3|CAS9_FRATW/1-1629      928  -- -- SPENIFKDKN -- -- NRIKEFAKGI S A Y S G A N L T D G D F D - - G A K E E L D H I I P R S H K K Y G T L N D E A L I C V T R G D N K N G N R I F C L R D L A D N Y K L K Q F E T T D D L E I E K K I A D T I W D A N K K D F K F 1043
sp|J3F2B0|CAS9_ACTNW/1-1101     546  -- -- -- G Y I S R G D I V R L D A L E L Q G C A C L Y C G T T I G Y H T - - - - - C Q L D H I V P Q A G P G S N - - N R R G L V A V C E R C N S K S N T - P F A Y W - A Q - - - - - K C G I P H V G V K E A I G R V R G W R K Q T P N - - 644
sp|C6AKI3|CAS9_COROI/1-1064     537  -- -- -- G K P S R A D L W R Y Q S V Q R Q C Q C A Y C G S P I T F S N - - - - - S E M D H I V P R A G Q G S T - - N T R E L V A V C H R C N Q S K G N T - P F A I W - A K - - - - - N T S I E G V S V K E A V E R T R H W T D T G M - - 635
sp|G3EQI1|CAS9_STRTR/1-1409     825  K L S K I D N N A L Q N D R L Y L Y L Q N G K D M Y T G D D L D I D R L S - - - - - N Y D I D H I I P Q A F L K - D - - N S I D N K V L V S S A S N R G S D D F P S L E V - V K - - - - - K R K T - - - - - F W Y Q L L K S K L I 920
sp|Q3J161|CAS9B_STRTD/1-1388    804  K L S K I D N N A L Q N D R L Y L Y L Q N G K D M Y T G D D L D I D R L S - - - - - N Y D I D H I I P Q A F L K - D - - N S I D N K V L V S S A S N R G S D D F P S L E V - V K - - - - - K R K T - - - - - F W Y Q L L K S K L I 899
sp|Q927P4|CAS9_LISIN/1-1334     803  -- -- -- P T D N Q E L R N N R L Y L Y L Q N G K D M Y T G D D L D I H N L S - - - - - N Y D I D H I V P Q S F I T - D - - N S I D N K V L V T S S A G N R G S D D V P P L E I - V R - - - - - K R K V - - - - - F W E K L Y Q G N L M 895
sp|Q92W02|CAS9_STRP/1-1368     800  -- -- -- P V E N T Q L Q N E K L Y L Y L Q N G R D M Y V D Q E L D I N R L S - - - - - D Y D V D H I V P Q S F L K - D - - D S I D N K V L T R S D K N R G S D N V P S E E V - V K - - - - - K M K N - - - - - Y W R Q L L N A K L I 892
sp|Q8DTE3|CAS9_STRM/1-1345     800  -- -- -- P V E N S Q L Q N D R L F L Y L Q N G R D M Y T G E L D I D Y L S - - - - - Q Y D I D H I I P Q A F I K - D - - N S I D N K V L T S S K E N R G S D D V P S K D V - V R - - - - - K M K S - - - - - Y W S K L L S A K L I 892
sp|QDP897|CAS9_CAMEI/1-984      515  -- -- L G L K I N S K I L K L R L F K E Q K E F C A Y S G E K I K I S D L Q - D E K M L E I D H I V P Y S R S F - D - - D S Y M N K V L V T K Q N Q E L N Q T P F E A F - G N - - - - - D S - - - - - A K W Q K I E V - - L 607
sp|Q9CLT2|CAS9_PASW/1-1056     536  -- -- F S S E P K S K D I L K F R L Y E Q Q H G K C L Y S G K E I N I H R L N - E K G Y V E I D H A L P F S R T W - D - - D S F N K V L V L A S E N Q N G N Q T P Y E W L - Q G - - - - - K I N S - - - - - E R W K N F V A L V L 632
sp|C9XIG5|CAS9_NEIN/1-1062     544  -- -- F V G E P K S K D I L K L R L Y E Q Q H G K C L Y S G K E I N L G R L N - E K G Y V E I D H A L P F S R T W - D - - D S F N K V L V L G S E N Q N G N Q T P Y E Y F - N G - - - - - K D N S - - - - - R E W Q E F K A R V E 640
sp|A1I1Q68|CAS9_NEIM/1-1082     544  -- -- F V G E P K S K D I L K L R L Y E Q Q H G K C L Y S G K E I N L G R L N - E K G Y V E I D H A L P F S R T W - D - - D S F N K V L V L G S E N Q N G N Q T P Y E Y F - N G - - - - - K D N S - - - - - R E W Q E F K A R V E 640
sp|J7RU45|CAS9_STAAU/1-1053     512  -- -- T G K E N A K Y L I E K I K L H D M Q E G K C L Y S L E A I P L D L L N N F N Y E V D H I I P R S V S F - D - - N S F N K V L V K Q E E N S K K G N R T P F Q Y L - S S - - - - - S D S K - - - - - I S Y E T F K K H I L 609
sp|Q33LF7|CAS9A_STRTD/1-1121   554  -- -- S V F H G H K Q L A T K I R L W H Q Q G E R C L Y T G K T I S I H D L I N N S N Q F E V D H I I L P L S I T F - D - - D S L A N K V L V Y A T A N Q E I K G Q R T P Y Q A L - D S - - - - - M D D A - - - - - W S F R E L K A F - - 649

```

Figure 2.5: Multiple sequence alignment of the HNH domain of Cas9s with the seven highly conserved residues highlighted in blue.

F in some of the enzymes. The third is amino acid N854 that is replaced by an S in some homologs.

From existing structures of Cas9, it is observed that the five highly conserved residues lying in the core of the $\beta\beta\alpha$ -motif differed in their spatial orientations between the structures crystallized with and without divalent cations in the active site as depicted in figure 2.7 [53] [120] [123]. This suggests that coordination of the cation and interactions in the active site may be essential for stabilizing the structure necessary for cleavage to occur. Hence, MD simulations are performed on the HNH domain (see section 2.2) to monitor the dynamics of these conserved residues in the presence and absence of an ion in the active site. The root mean square deviation (RMSD) changes as a function of time is shown in figure 2.8 indicating that the simulations were stable.

Figure 2.6: Multiple sequence alignment of the seed sequences of Pfam HNH_4 family with the four identical residues highlighted in blue color and the three varying residues indicated by a red arrow.

Figure 2.7: (a) AnCas9 HNH domain structure (PDB ID: 4OGE) showing the five highly conserved residues in the core of the $\beta\beta\alpha$ -motif and their orientation in the presence of an ion (Mg^{2+} in magenta) in the active site. Catalytic residues are depicted in white sticks and the other conserved residues are depicted in orange with an element-wise colour scheme (C in orange/ white, O in red and N in blue); and (b) SpCas9 HNH domain structure (PDB ID: 5F9R) showing the five highly conserved residues and their orientation in the absence of an ion in the active site. The same colour scheme as (a) is used. Also, the side chain atoms of K are missing as they are not resolved in the crystal structure.

2.3.2 P843 maintains the conformation of the active site cleft

P843 is the first residue examined, which sits in the loop connecting the two β strands of the $\beta\beta\alpha$ -motif as shown in figure 2.9(a). Simulations of the wild type protein in both active and inactive states reveal a concerted motion between the loop containing P843 (loop 1) and the loop containing residues S867, D868, and N869 (loop 2). Expectedly, changing P843 to alanine in the simulations resulted in the loss of a kink in loop 1. Importantly, it is observed that this loop containing residue P843 also appeared to move away from the active site. This movement is quantified by calculating the distance between the backbone atoms of D839 and residues 842-844 (refer figure 2.9(b)); it is clear that the P843A mutant sampled

Figure 2.8: Root mean square deviation (RMSD) of the individual replicates of the (a) wild type active state simulations with Mg^{2+} ion; (b) wild type inactive state simulations; (c) wild type active state simulations without any ion; (d) wild type active state simulations with Zn^{2+} ion; (e) P843A mutant simulations; and (f) K866A mutant simulations.

larger distances than the wild type protein.

Coupled to the movement of loop 1 away from the active site, loop2 appeared to collapse into the active site cleft (refer figure 2.9(c)). To quantify the movement of loop 2 into the active site cleft, the distance between the backbone atoms of D839

and residues 867-869 is computed. The wild type simulations mainly sampled distances greater than 10 Å and exhibited two peaks, one around 13 Å (closed state) and the other around 18 Å (open state) as shown in figure 2.9(d). The open state corresponded to loop 2 reorienting itself due to a transient loss of contact between D839 and K866. In contrast, the P843A mutant simulations exhibited not only the peak around 18 Å (open state), but also exhibited a new larger peak around 7.5 Å (collapsed state).

The collapse of loop 2 may affect the positioning of DNA in the active site cleft. Indeed, in the P843A mutant simulations, the DNA including its backbone underwent larger fluctuations in comparison to the wild type (refer figure 2.10(a),(b) and (d)). This highly dynamic behaviour of the DNA, including large shifts in the backbone atoms during the P843A mutant simulations, indicates that proper placement of the DNA in the active site is strongly affected. Collectively, the simulation results show that P843 plays a crucial role in maintaining the active site cleft conformation by conferring a kink in loop 1, which then prevents the collapse of loop 2 into the active site cleft, thereby allowing DNA to be placed appropriately for catalysis.

2.3.3 K866 forms interactions with the catalytic residues and Mg^{2+}

From the simulations, an interaction network formed between K866-D839- Mg^{2+} -N863 is identified, with the side chain of K866 engaged in hydrogen bonds with the catalytic residue D839 40% of the time (refer figure 2.11(a)). However, the network is disrupted in the absence of the divalent cation (refer figure 2.11(b)). A distribution of the distance between K866 NZ and D839 OD1 (or OD2) peaked at ~ 3 Å in simulations with an ion (the active state), suggesting the presence of a hydrogen bond or salt bridge, while the distance distribution peaked at ~ 13 -14 Å

Figure 2.9: (a) A snapshot from the wild type (active state) simulation with ion (Mg^{2+}) showing the positioning of P843 (in loop 1 colored green) with respect to the catalytic residues (D839 and N863 shown in sticks) and also indicating another loop (loop 2 colored orange) consisting of residues S867, D868, and N869; (b) Distance distribution showing the slight drifting apart of loop 1 (containing residues V842, P843, and Q844) from active site cleft in the P843A mutant simulations compared to wild type simulations; (c) A snapshot from the mutant P843A simulation with ion (Mg^{2+}) showing the collapse of loop 2 into the active site cleft and the loop 1 drifting away from the active site cleft; and (d) Distance distribution showing the collapse of loop 2 (containing residues S867, D868, and N869) into the active site cleft in the mutant P843A simulations in comparison to wild type active state simulations.

in simulations without any ion (the inactive state) (refer figure 2.11(c)).

Further, the K866A mutant is simulated, and it is observed that the percentage of time that the two catalytic residues, D839 and N863, were coordinated with the Mg^{2+} ion decreased in comparison to the wild type (refer figure 2.11(d) and (e)). Hence, it is likely that K866 is important for maintaining the functional spatial arrangement of the catalytic residues.

Figure 2.10: Time step evolutions (from red to blue) of the DNA's position in (a) wild type (active state) simulation; (b) P843A mutant simulation; (c) K866A mutant simulation. Residues D839, P843/ A843, N863 and K866/ A866 are represented by sticks with an element-wise color scheme (C in cyan, N in blue, and O in red). The protein is depicted in grey and the Mg²⁺ ion is depicted in magenta. In both the P843A and K866A mutant simulations, the DNA showed greater fluctuations in space compared to the wild type simulation; and (d) Root Mean Square Fluctuation (RMSF) values for the DNA backbone atoms in the wild type active state (in black), the P843A mutant (in red) and the K866A mutant (in green) simulations. The DNA was less stably bound and underwent larger spatial movements during the mutant simulations than the wild type.

Next, the structure of the $\beta\beta\alpha$ -motif is examined in greater detail. Both the α -helix containing the catalytic asparagine and the β -sheet containing the catalytic aspartate are largely stable during our simulation of the wild type protein, as shown by the secondary structure plot in figure 2.12(a). However, in the K866A mutant simulation, the α -helix with N863 is destabilized as depicted in figure

Figure 2.11: (a) A snapshot of the active state simulation in the presence of a divalent cation (Mg^{2+}) showing the formation of interactions between D839, N863, K866, and the ion; (b) A snapshot of the inactive state simulation in the absence of an ion showing they were oriented differently in space and were too far apart to interact; (c) Distribution plot of the distance between the K866 NZ atom and the D839 OD1 (or OD2) atom in the active state and inactive state simulations; (d-e) 3D histograms of the distances between the catalytic residues, D839 and N863, and the Mg^{2+} ion for (d) the wild type (active state) and (e) the K866A mutant simulations. The amount of simulation time that both the catalytic residues were concurrently coordinated with the ion was reduced from around 78% to 50% when K866 was mutated to alanine.

2.12(b). In contrast, the β -sheet with D839 remained stable, with the residue undergoing only some minor reorientation in its side chain. Subsequently, two additional simulations were carried out with wild type Cas9, one in which the starting structure was the inactive state (figure 2.12(c)) and the other in which the Mg^{2+} ion was removed from the active state (figure 2.12(d)). In both scenarios, the helical region containing N863, but not the β -sheet containing D839, is destabilized in a similar manner to that seen in the K866A mutant simulations.

Figure 2.12: Secondary structure assignment plots for (a) wild type (active state) simulations with ion; (b) K866A mutant simulations with ion; (c) wild type (inactive state) simulations; and (d) wild type (active state) simulations without ion. In (a) the α -helix containing the catalytic N863 residue is preserved throughout the simulations. While in (b), (c) and (d) the helicity is frequently lost. The purple arrow indicates the helical region to focus on in all of the four plots.

Destabilization of the α -helix might affect its interaction with DNA. Hence, the interactions of N863 and R864 with the DNA nucleotides C3 and T4 (between

which lies the cleavage site) is examined in the wild type active state and the K866A mutant. In the wild type simulations, the N863 ND2 atom is involved in a hydrogen bond interaction with O5' atom of the DNA nucleotide T4, while R864 NE and NH2 are involved in hydrogen bonds with O2 and N3 atoms of nucleotide C3 respectively (refer figure 2.13(a)). However, the distance distribution plots from the simulations showed that the interactions of N863 and R864 with DNA are dramatically reduced in the K866A mutant (refer figure 2.13(b) and (c)). On further inspecting the positioning of the DNA in the active site cleft over the course of the simulations it is observed that the DNA was stably bound in the wild type simulations but underwent large fluctuations and greater displacements relative to the original position in the K866A mutant simulations (refer figure 2.10(a), (c) and (d)). Collectively, the simulation results indicate that K866 plays an important role in positioning the DNA in the active site by stabilizing the $\beta\beta\alpha$ -motif structure and is consistent with a recent experimental study that reported the loss of function of Cas9_{K866A} [142].

2.3.4 Zn^{2+} cannot substitute for Mg^{2+} in coordinating the highly conserved residues within the active site of the HNH domain

Previous experiments have shown that the SpCas9 enzyme exhibits preferences for particular divalent cations, in particular magnesium [49]. To probe the reason underlying this ion selectivity, MD simulations of the HNH domain with the Zn^{2+} cation replacing Mg^{2+} were performed. Strikingly, it is found that the K866 NZ atom is no longer involved in a hydrogen bond interaction with the D839 OD1 (or OD2) atom, as depicted by the distance distribution plot between the two atoms in figure 2.14(a). It is also observed that the percentage of time that the catalytic residues D839 and N863 are both simultaneously coordinated with the

Figure 2.13: (a) A snapshot showing the interactions of N863 or R864 with DNA. 1 represents the interaction between N863 ND2 atom and T4 O5' atom, 2 represents the interaction between R864 NE atom and C3 O2 atom, and 3 represents the interaction between R864 NH2 atom and C3 N3 atom; (b) Distribution profiles of the distance between amino acid N863 and nucleotide T4. The wild type active state simulation is shown in black, while the K866A mutant simulation is shown in green; and (c) Distribution profiles of the distance between amino acid R864 and nucleotide C3 (wild type in black and K866A in green). The DNA moves further away from the α -helix in the K866A mutant simulation when compared to wild type.

divalent cation is considerably reduced to only 4% (refer figure 2.14(b)). Furthermore, the α -helix containing the catalytic N863 residue is destabilized (refer figure 2.14(c)), with the DNA in the active site cleft undergoing considerable amount of spatial shifts as shown in figure 2.14(d). Collectively, the analysis revealed that replacement of Mg^{2+} with Zn^{2+} in the HNH domain led to a disruption of the K866-D839-ion-N863 interaction network. The difference in properties between Zn^{2+} and Mg^{2+} including size of the divalent ions could be the reason for this

but more detailed study with simulations using divalent ions of different sizes are required to confirm. The loss of the K866-D839-ion-N863 interaction network indicates change in the active site and this possibly explains the reason behind loss of catalysis with Zn^{2+} .

Figure 2.14: (a) Distribution plot of the distance between K866 NZ atom and D839 OD1 (or OD2) atom in the HNH domain simulations with Zn^{2+} ion in the active site showing that almost no hydrogen bond interaction existed; (b) 3D histogram of the distances of two catalytic residues D839 and N863 from the ion (Zn^{2+}) in the HNH domain simulations with Zn^{2+} ion in the active site showing a drastic reduction in simulation time that both the residues were coordinated with the ion; (c) Secondary structure assignment plot for the simulations of the HNH domain with Zn^{2+} ion in the active site showing that the α -helix containing the catalytic residue N863 was lost predominantly; and (d) A time step evolution (from red to blue representing 0 to 100 ns) of the DNA's position during simulations of the HNH domain with Zn^{2+} in the active site showing greater displacement in space of the DNA.

2.4 Discussion

Cas9 is a versatile tool for genome engineering and can be deployed in a wide range of applications. Yet, much about the protein and its catalytic mechanism are not fully understood. In this study, the catalytically active state of the HNH domain is modelled and novel functions of the highly conserved residues in the $\beta\beta\alpha$ -motif of the domain are uncovered. Specifically, P843 is essential for maintaining the cavity in the active site cleft for DNA to dock and the results highlight the importance of proper van der Waals interactions between two loops (loop 1 and loop 2 in figure 2.9) surrounding the active site. Also, a K866-D839-Mg²⁺-N863 interaction network is identified that appears to be important for positioning the catalytic residues and for maintaining the stability of the $\beta\beta\alpha$ -motif structure.

One of the drawbacks of the current work is that the HNH domain is modelled in isolation in order to gain insights into its active state. Conceivably, the HNH domain may behave differently when it is joined to all the other domains of the Cas9 protein. However, the size of the enzyme is large, being typically more than 1000 amino acids long. To completely sample all possible conformational states of such a large protein complexed with other biomacromolecules requires significant computational power. Zuo et al. attempted to simulate the entire SpCas9-sgRNA-DNA complex for 200-300 ns, but they were unable to sample the active conformational state of the HNH domain, which remained at a considerable distance from the target DNA throughout their simulations [126]. In another recent study, the structure of Cas9 was resolved using cryo-EM and the authors claimed that the HNH domain has moved significantly closer (~ 13 Å) to the cleavage site in comparison to previous structures [143]. But, even in this structure the helix containing N863 adopts a conformation similar to that observed in the absence of an ion and this could possibly be because the atomic model was built by homology modeling using PDB ID: 4OO8 [119] as the template, where the helix adopts a

conformation similar to that in the absence of ion. Hence, there is an unfortunate trade-off between tractability and realism.

Further, our model involves only a single-stranded DNA in the active site cleft in coordination with the divalent cation, while in reality the assembly contains a DNA-RNA heteroduplex. Modeling a DNA-RNA heteroduplex in the active site of the HNH domain would require a proper heteroduplex template to begin with. The resolved structures of $\beta\beta\alpha$ -motifs in active state deposited in PDB all involved a DNA duplex and none with a DNA-RNA heteroduplex. While a DNA-DNA duplex assumes a B-form, a DNA-RNA hybrid is likely to assume an intermediate conformation between the A-form and B-form [144]. And, thus without a template it is challenging to build a stable model of the structure in the active site.

Several avenues of research may be pursued in the future. First, although I have simulated the HNH domain in the presence of Mg^{2+} or Zn^{2+} , more can be done to better understand the ion selectivity of Cas9. Previous experiments have demonstrated that there are other divalent cations, such as Co^{2+} and Ni^{2+} , that can also catalyze DNA cleavage but at a lower efficiency than magnesium [49]. Subsequent simulations may be targeted at these ions to determine how different ions modulate the dynamic behaviour of the Cas9 protein. Second, other CRISPR-Cas systems have recently emerged as powerful new tools for genome engineering. One example is the Cpf1 enzyme, which exhibits properties that are distinct from those of Cas9 [55]. Although, the crystal structures of Cpf1 from different bacteria have been solved [145] [146], it will also be useful to examine the conformational dynamics of the enzyme in the absence and presence of nucleic acids.

2.5 Conclusion

In conclusion, MD simulations can help to provide novel insights into the functioning of CRISPR-Cas9 systems, which may not be easily obtained from experimental methods like X-ray crystallography. The analysis in this study focusing on the HNH domain have demonstrated that the highly conserved residues play important roles in maintaining the proper structure of the active site cleft, which allows for correct docking of DNA for cleavage to occur. Results from this and other recent computational studies [125] [126] can serve as a springboard for downstream experimental work aimed at either modifying the Cas9 enzyme for higher efficiency and specificity or fully understanding the dynamical properties of CRISPR-Cas9 systems.

Preface to Chapter 3

In an attempt to understand the functioning of the editing system (DNA editing system - CRISPR-Cas), MD simulations on the HNH nuclease domain of SpCas9 helped uncover the role of conserved residues from the structural and functional perspective as described in chapter 2. But, as mentioned in the previous chapter, this study had its limitations in terms of building models of the domain in isolation and modeling a single-stranded DNA in the active site instead of the DNA-RNA heteroduplex. The unavailability of structural templates and calibration data for building and refining the models, shifted the focus on to characterizing the enzymes using activity based studies.

In the following chapter (Chapter 3) as a continuation to understanding the functioning of the editing system, a large scale characterization is attempted by using 5 CRISPR-Cas systems to evaluate their functioning under different circumstances using experiments to probe their activity. Although, a number of CRISPR-Cas systems have been identified over a period of time, a single platform comparing their performances could enable identifying the pros and cons of using a particular CRISPR-Cas system and thereby, facilitate choosing an appropriate system for future studies depending on the circumstances.

Chapter 3

Developing computational pipeline for a systematic evaluation of CRISPR-Cas systems in human cells

Although, CRISPR-Cas systems have wide range of applications in genome editing and regulation, but how the systems perform in comparison to one another is yet unclear. In an attempt to probe the functioning and characterizing the CRISPR-Cas systems, this study was designed "to compare 5 CRISPR-Cas systems based on their performance in human cells by targeting over 90 sites in genes with varying levels of expression. The performance of the 5 systems in non-homologous end joining (NHEJ)-mediated and homology-directed repair(HDR)-mediated genome editing is evaluated" by designing a rigorous computational pipeline in order to process the Illumina deep sequencing data from the experiments that assisted in quantification. Apart from a trade-off observed between the cleavage efficiency and target specificity in the enzymes, the editing by smaller enzymes were less

affected by the change in gene expression than the larger ones. Further, it is observed that Cpf1s outperform the Cas9s in homology-directed repair while using single-stranded oligodeoxynucleotide (ssODN) donor templates and exhibit preference for ssODNs of the non-target strand sequence. "In contrast, *Streptococcus pyogenes* (SpCas9) exhibits a preference for ssODNs of target strand sequence". Also, asymmetric donors with longer arms 5' improved the HDR efficiencies of Cpf1s and SpCas9. The results from the study collectively helped in building a guideline that could potentially be useful in designing future genome engineering studies.

3.1 Background

CRISPR-Cas systems comprise of an endonuclease which is guided to a desired location on the genome using a chimeric single guide RNA (sgRNA). The presence of a short sequence motif, protospacer adjacent motif (PAM), at the target locus is essential for the endonuclease to cleave the DNA. Thus, the PAM constraints the parts of genome that can be targeted by a particular Cas endonuclease. And, this is tackled by incorporating Cas enzymes from different bacterial species as they recognize different PAMs, or by modification of existing Cas enzymes to recognize varying PAM sequences. Therefore, the engineering toolbox is enhanced by incorporating various CRISPR-Cas systems to expand the range of sites targeted. In order to obtain desired genome engineering outcome, distinct endogenous DNA repair pathways are utilized [147] [148]. A Cas endonuclease causes a double-stranded break at the target site that is typically repaired by either the non-homologous end joining (NHEJ) repair or the homology-directed repair (HDR). In the absence of a repair template, the former is activated and often results in error-prone repair causing insertions and deletions (indels). While, the presence of a homologous repair template activates the latter resulting in precise DNA changes

as specified in the template at the target site. The dominant repair pathway in higher eukaryotes is NHEJ and thus, HDR usually occurs at a very low frequency. Wide range of genome engineering studies in human cells have been performed by using CRISPR-Cas systems from different bacterial species [149] [150] [151] [56] [50] [152] [153]. *Streptococcus pyogenes* Cas9 (SpCas9), being the most characterized Cas enzyme and the first to be successfully used in human cells, is largely the first choice for both NHEJ-mediated and HDR-mediated genome editing [149] [150] [151]. Although, simple PAM requirement of SpCas9 (NGG) is an advantage, its large size (1368 amino acids) in comparison to other Cas9 homologs makes it difficult to employ it for certain in vivo therapeutic applications. While, on the other hand smaller Cas9 homologs like *Staphylococcus aureus* Cas9 (SaCas9) with 1053 amino acids and *Neisseria meningitidis* Cas9 (NmCas9) with 1082 amino acids have an advantage of size but are restricted due to complex PAM requirements (NNGRRT for SaCas9 and NNNNGATT for NmCas9) [154].

As more and more Cas systems are discovered, it has become essential to know the relative performance of the systems in engineering mammalian genomes to be able to utilize the optimal system depending on the application. Although, abundant literature is available showcasing the activity of the CRISPR-Cas systems, it is not straight forward to compare the relative performance of the different systems as the cellular context, protein expression levels, target sites and experimental conditions vary in the different studies. Two studies attempted to compare SpCas9 with Cpf1 nucleases but focused mainly on the specificities of Cpf1s and thus were limited in scope [155] [156]. Thus, this study was designed with the objective of "comparing the performances of three Cas9 nucleases (SpCas9, SaCas9 and NmCas9) and two Cpf1 nucleases (from *Acidaminococcus* sp. BV3L6, AsCpf1, and from *Lachnospiraceae* bacterium ND2006, LbCpf1) at different target sites in human cells". The data from the high-throughput sequencing after performing

the experiments helped in quantification of the performances of the 5 enzymes and thus, can serve as a guidance for those who intend to use the CRISPR-Cas technology for performing genome editing.

3.2 Design of the experimental study

A framework needed to be established for evaluating the 5 Cas endonucleases (SpCas9, SaCas9, NmCas9, LbCpf1 and AsCpf1) in an unbiased way. Target genomic loci should ideally be the same to be able to compare the performance of the 5 enzymes but that is not straight forward as the PAM requirement is not the same for all the nucleases examined. Thus, it was decided that locus will be selected such that they are flanked by the PAM for Cpf1 on one side and Cas9 on the other side. But, the PAM requirement for the 3 Cas9s examined were different and were incompatible for SaCas9 (NGGRRT) and NmCas9 (NNNGATT). Therefore, 61 locus flanked by TTTN (PAM for Cpf1s) on one side and by "either NGGRRT (combined PAM for SpCas9 and SaCas9) or NGGNGATT (combined PAM for SpCas9 and NmCas9)" on the other side were selected as targets. These targets varied in length from 17 to 23 nucleotides. Also, as the chromatin accessibility may have an influence on the activity, "genes with varying expression levels in HEK293T cells" were chosen because the chromatin architecture is known to control the gene transcription [157]. The expression level of the chosen genes calculated from the RNA-Seq data, as shown in figure 3.1, indicates over 4000-fold difference in expression.

For the NHEJ-mediated genome editing, the locus selected were targeted by the Cas enzymes without providing any repair templates in the experimental set up. While, for the HDR-mediated genome editing, the repair templates provided in the experimental set up were designed in such a way that a XbaI restriction site is incorporated at the cut site if the experiment is successful. Figure 3.2 depicts

Figure 3.1: Expression levels of five selected genes (ALK, EGFR, NF1, KDM6A, and STAG2) based on RNA-Seq experiments. Data represent mean $\pm$ s.e.m (n = 6).

the two endogenous repair pathways tested in this study. The experimental set up, material used, loci targeted and procedures are explained in greater details in [158].

3.3 Pipeline developed

Computational pipeline needed to be developed to analyze the illumina deep sequenced reads with the objective of comparing the performance of the Cas endonucleases in both NHEJ-mediated and HDR-mediated genome editing. The performance in NHEJ-mediated genome editing is quantified based on the frequency at which indels are introduced in the vicinity of the cut site, while for the HDR-mediated genome editing the rate of incorporation of the restriction site at the precise location is used as a measure.

The very first step in processing the sequenced reads is mapping them to a reference. "A local reference library is built using the amplicon sequences of the target genes to be used for mapping the reads against". The building of local reference library enabled reducing the complexity relating to computational time, space required and also facilitated fine tuning of parameters used for the mapping. The

Figure 3.2: The two endogenous repair pathways that the 5 Cas enzymes were compared for performance include the NHEJ and HDR. As depicted in the figure in NHEJ no repair template is provided, thus during the process of repair random insertions and deletions occur close to the cut site. While for HDR a repair template is provided and in this study the repair template provided are designed in such a way that a restriction site is incorporated as a precise insertion at the cut site. Figure obtained from [159].

parameters utilized for mapping needed to be fine-tuned to enable mapping to correct locus along with sensitivity to very small (1-2 bp) and large (50-70 bp) indels both of these kinds are characteristic indel patterns associated with CRISPR induced indel patterns. Thus, a range of values were employed for the following parameters that include match score, mismatch penalty, gap open and extension penalty, and clipping penalty to verify the capture of the above described features. Randomly reads were scrutinized to see if they map to the correct locus and capture the insertions and deletions accurately with the different values for the parameters tested and the set of values that best capture the actual scenario is further chosen. Finally, the sequenced reads were mapped to the local reference library using BWA-MEM" [8] using this set of parameters that best capture the actual scenario and they are as defined below:

- Match score, a score given when there is a match of base between the refer-

ence and aligned read, set to 1

- Mismatch penalty, a penalty given when there is a mismatch of base between the reference and aligned read, set to 2
- Gap open penalty, a penalty given for opening a gap in the alignment, set to 6
- Gap extension penalty, a penalty for extending a gap by one base, set to 1
- Clipping penalty, a penalty for 5'- and 3'-end clipping, set to 8. In local alignment (Smith-Waterman), a read may be not be aligned from the starting base to the last base to a position in the reference and as a result the ends may be clipped off and the penalty for performing such a clipping is given by the clipping penalty.

"The uniquely mapped reads having a mapping quality ≥ 20 were sorted. Then Picard, a set of command line tools developed to manipulate high throughput sequencing data, was used to assign group information using '*AddOrReplaceRead-Groups*' " [160]. Local realignment and recalibration of base quality scores was performed using GATK toolkit [12] [13] [14].

The next step involved identification of indels. CIGAR string which is an output parameter from the mapping tool is used for this. CIGAR string describes how the read aligns to the reference using a combination of operators and numbers. The operators are described in table 3.1 and denote the state of alignment, while the number represent the number of bases that the operator is used to describe. For example: a CIGAR string *8M2I5M1D10M5S*, indicates that 8 bases are aligned with the reference as match or mismatch followed by and insertion of 2 bases in the read and then 5 bases aligned followed by a deletion of 1 base in the read and then 10 bases aligned and finally 5 bases that are clipped. So, now by processing the CIGAR string, indels are identified and each read is tagged as either an "I" or "D" or "M". I indicates an insertion is found in the read, D indicates a deletion,

while M indicates that the read align with the reference without any insertions or deletions. Only a region of 40 bp on either side of the spacer was included in the indel calculations as only these are considered to be introduced due to the CRISPR-Cas related editing at the locus of interest. Finally, the total number of reads tagged as "I" or "D" is used to calculate the percentage of indels for each sample. Also, some reads were randomly selected and aligned with the relevant reference sequence using Needleman-Wunsch algorithm. A manual comparison of the output from these alignments with the script output revealed that the analysis and the pipeline developed is accurate and captures the actual scenario perfectly.

Table 3.1: List of operators used to describe the read alignment to reference in the CIGAR string

Operator	Descriptions
M	Match that denotes an alignment match or mismatch
I	Insertion denoting an insertion in the read in comparison to the reference
D	Deletion denoting a deletion in the read in comparison to the reference
S	Soft clipping denoting a region at the start or end of read that is not a part of alignment with the reference but is still represented in the SEQ string
H	Hard clipping denoting a region at the start or end of read that is not a part of alignment with the reference and is not present in the SEQ string. Also, this region usually aligns elsewhere in the reference

"Another local reference library is built consisting of the modified genes for characterizing the HDR-mediated genome editing". The steps described above were all repeated again from mapping to tagging the reads as "I", "D", or "M", with the only difference being in the reference library used for mapping, which in this case was the newly built reference library consisting of the modified gene sequences [that encompasses the HDR introduced modifications at the locus particularly in

this case consists of a restriction site incorporated at the cut site]. The reads are further processed and binned into three categories: "No incorporation", when the restriction site is not present in the read; "Correct incorporation", when the restriction site is present in the read and the read is tagged as "M"; and "Incorrect incorporation", when the restriction site is present in the read and the read is tagged as either "I" or "D". Finally, the total number of reads binned in each of these three classes is used to compute the rate of correct and incorrect incorporations.

The pipeline developed is illustrated in figure 3.3.

Figure 3.3: A schematic of the computational pipeline developed to analyze the NGS data for both quantifying the performance of the enzymes in NHEJ-mediated and HDR-mediated genome editing.

3.4 Results and biological inferences

The pipelines described in the above section are used to calculate the indel frequency and restriction site incorporation rates for each experimental sample. These numbers from the individual samples are then combined based on the conditions compared to evaluate the performance of the enzymes. The following sections detail the different findings from the study based on calculated numbers and their biological significance.

3.4.1 "Performance of the CRISPR-Cas systems in NHEJ-mediated genome editing"

The editing efficiencies of the 5 Cas enzymes in the NHEJ-mediated genome editing, at the chosen 61 genomic loci in HEK293T cells, is examined. Grouping the loci based on spacer lengths indicated that at lower spacer lengths between 17-20 nt, SpCas9 performed the best with highest cleavage efficiencies (refer figure 3.4(a),(b)). While, at higher spacer lengths between 21-23 nt SaCas9 and LbCpf1 exhibited highest editing (refer figure 3.4(a),(b)). The location of the target sites whether in intronic region or protein-coding region had no effect on the editing activity observed and a similar trend was found irrespective of the location (refer figure 3.5).

"The PAM-proximal seed region of the DNA target is critical for proper recruitment of the CRISPR-Cas systems". But, with the present design of locus selection, the PAMs are on opposite sides for Cas9 and Cpf1, thereby causing the seed region to vary for the enzymes. Thus, to investigate and compensate for any effect due to seed, new target sites were selected where both Cas9 and Cpf1 has overlapping seed regions as shown in figure 3.6. The editing activity indicated a similar trend as observed with same spacer.

Figure 3.4: (a) NHEJ-mediated editing activity measured as percentage indels introduced for SpCas9, SaCas9, LbCpf1 and AsCpf1 based on the illumina deep sequencing data; and (b) NHEJ-mediated editing activity measured as percentage indels introduced for SpCas9, NmCas9, LbCpf1 and AsCpf1 based on the illumina deep sequencing data. The mean of the editing activities is represented by a horizontal bar and the results are grouped based on spacer length of the target genes. The results indicate SpCas9 is effective at lower spacer lengths (17-20 nt), while SaCas9 and LbCpf1 are effective at higher spacer lengths (21-23 nt).

Figure 3.5: Extent of genome modification represented by percentage indels determined from the deep sequencing data for loci with varying spacer length between 17-23 nt long. Data is represented as mean $\pm$ s.e.m. Intronic sites in genes with varying expression levels including (a) ALK; (b) EGFR; (c) NF1; (d)KDM6A; and (e) STAG2. Two panels represent locus selected for evaluating either SaCas9 (left panel) or NmCas9 (right panel) with SpCas9, LbCpf1 and AsCpf1. Protein coding regions from genes including (f) APC; (g) ATM; and (h) KDM5C (left panel) and ALK (right panel). The percentage indels indicate that irrespective of the location of the target site in intronic region or protein coding region, SpCas9 is most efficient at lower spacer lengths and SaCas9, LbCpf1 at higher spacer lengths.

Figure 3.6: Extent of genome modification denoted by percentage indels at target locus in (a) CACNA1D and (b) PP1R12C gene where "Cas9 and Cpf1 had overlapping seed region and three different spacer lengths of 17, 20 and 23 nt examined". Data is represented as mean $\pm$ s.e.m. A similar trend is observed for the editing activity with SpCas9 being most active at lower spacer lengths and SaCas9, LbCpf1 active at higher spacer lengths.

Predominantly, at most of the target sites examined NmCas9 performed poorly in comparison to the other nucleases (refer figure 3.4(b)). As the length of the naturally occurring crRNA spacers in *N. meningitides* is 24 nt, further nine target sites were identified with spacer length of 24-25 nt long, flanked by the PAM sequence for Cpf1 on one side and NmCas9 on the other side. The chosen sites had high expression levels but the editing activity for NmCas9 was again low in comparison to LbCpf1 and AsCpf1 at these sites too (refer figure 3.7). Experimentally, NmCas9 was verified to have poorer performance in other cells lines as

well [158]. Thereby, collectively suggesting that NmCas9 is probably not an ideal Cas nuclease for many genome editing application.

Figure 3.7: (a) Expression levels of six genes based on RNA-Seq experiments where the nine additional selected sites were present. Data is represented as mean $\pm$ s.e.m; (b) Editing efficiency represented by percentage indels for NmCas9 and two Cpf1 nuclease at the nine additionally chosen sites with longer spacer length calculated from deep sequencing data. Data is represented as mean $\pm$ s.e.m; and (c) Strip chart indicating the individual editing activities at the nine chosen sites having matching spacers of length 24-25 nt summarized by the nuclease for the three nucleases examined. The results show that NmCas9 has poor editing activity even at the selected target sites with higher spacer lengths of 24-25 nt.

Next, the editing activities of the enzymes were compared with the expression level of the target sites in order to identify any effect due to local chromatin context. Grouping the target sites based on their expression levels for each nuclease and comparing the editing efficiency depicted by the percentage indels indicated that for SpCas9, AsCpf1 and LbCpf1 editing is significantly affected (p-value <0.05 , Wilcoxon rank sum test) by expression as shown in figure 3.8. This is also consistent with prior studies that have reported that chromatin structure may affect the efficiency of the CRISPR-mediated genome editing [161] [162] [163] [164]. How-

ever, for SaCas9 and NmCas9 the target site expression levels did not seem to impact the editing efficacies (refer figure 3.8). This is possibly because these two enzymes are smaller in size in comparison to others and so may be that eases the access to heterochromatin. Restricting the target sites to only those having sgRNA with optimal length for each enzyme also displayed the similar results (refer figure 3.8 (b)).

At most of the target sites examined, although AsCpf1 showed a relatively good performance, it was the "best performing only in a minority of the sites" and usually outperformed by some other Cas enzyme as is seen from figure 3.9. "A pairwise comparison of AsCpf1 with either SpCas9 or LbCpf1 at only the sites with sgRNA of optimal length" also indicated that AsCpf1 has significantly lower cleavage efficiencies than SpCas9 or LbCpf1 as shown in figure 3.10 (p-value <0.05 , Wilcoxon rank sum test)". All three nucleases showed higher tolerance to mismatches around the middle of the spacer (refer figure 3.10). But, AsCpf1 exhibited lower tolerance to single base mismatches between sgRNA and target DNA, thereby, indicating higher specificity than the other two nucleases. Overall, the results suggest a trade-off between editing activity and specificity in the naturally occurring Cas endonucleases.

Figure 3.8: (a) Cleavage efficiency denoted by percentage indel is grouped based on expression level of the target site as lowly expressed (FPKM < 25, blue boxplots) and highly expressed (FPKM ≥ 25, red boxplots) for all the sites examined for the five Cas enzymes; and (b) Plot similar to (a) except that it is restricted to sites having sgRNA of optimal length for the nucleases that include 17-22 nt for SpCas9, ≥ 21 nt for SaCas9, ≥ 19 nt for NmCas9, LbCpf1 and AsCpf1. A comparison of the boxplots shows that SpCas9, AsCpf1 and LbCpf1 do have significant change in editing activity depending on the expression level of the target site, while for smaller enzymes SaCas9 and NmCas9 the difference is not profound.

Figure 3.9: At each target site examined the nuclease that exhibited highest cleavage efficiency is identified and then summation of the number of sites where a particular nuclease exhibits highest editing activity is determined after grouping by spacer length. (a) Percentage of Group A target sites where each enzyme performed best; and (b) Percentage of Group B target sites where each enzyme performed best.

Figure 3.10: (a) Pairwise comparison of AsCpf1 with either SpCas9 or LbCpf1 restricting to sites with sgRNAs of optimal length (19-22 nt inclusive) only; (b) A list of single base mismatches introduced at different positions of the spacer in sgRNA designed to monitor the tolerance of enzymes (SpCas9, LbCpf1, AsCpf1) to these mismatches that can serve as an indication of their specificity; and (c) Editing activity of SpCas9, AsCpf1 measured using the sgRNAs with single base mismatches as indicated in (b) using analysis of deep sequencing data. The results indicate that although AsCpf1 has reduced activity in comparison to SpCas9 and LbCpf1, it shows greater specificity.

3.4.2 "Performance of the CRISPR-Cas systems in HDR-mediated genome editing with single-stranded oligodeoxynucleotide donor"

In order to compare the performances of the 5 CRISPR-Cas system in HDR-mediated genome editing two genomic loci having matched seeds for Cas9 and Cpf1 are targeted. A single-stranded oligodeoxynucleotide (ssODN) is co-transfected as donor that would "introduce a XbaI restriction site between the cleavage sites of Cas9 and Cpf1". The donor template consisted of the restriction site flanked by 47 nt of homology on both sides and complementary to the target strand. The pipeline developed was used to analyze the illumina deep sequencing data from these experiments to identify the percentage of corrected targeted reads in order to quantify the performance of the enzymes in HDR-mediated genome editing. The results presented in figure 3.11 show that SpCas9 successfully inserted the "restriction site when the spacer length is 17 nt, while for spacer lengths of 20 and 23 nt AsCpf1 and LbCpf1" performed better. Consistently, for all 3 spacer lengths examined (17, 20 and 23 nt) both SaCas9 and NmCas9 performed poorly and were not able to introduce the XbaI restriction site. Further, varying the length of the "homology arm of the donor template from 47 to 27 nt" for the target site with a spacer length of 20 nt had no significant difference in activity and the "Cpf1 nucleases continued to exhibit higher HDR efficiency than SpCas9" (p-value <0.05, Student's t-test; refer figure 3.11 (c), (d)).

Varying the amount of donor template supplied between 100 and 300 ng during the experiment also produced comparable results (refer figure 3.12). Along with the rate of correct incorporation, the pipeline was computed the rate of incorrect incorporation and these numbers are 6.4-fold and 12.4-fold lower in comparison to the rate of correct incorporation for the two loci targeted (CACNA1D and

Figure 3.11: The target site original sequence and edited sequence of (a) CACNA1D and (b) PP1R12C gene along with the ssODN introduced represented by the brown bar that has a 47 nt homology arm on both sides and is of non-target strand sequence denoted by NT. Extent of XbaI restriction site incorporated into the target site is represented by the percentage correctly targeted using three spacer lengths that include 17, 20 and 23 nt at both the sites; (c) and (d) Extent of XbaI restriction site incorporated into the target site using a spacer length of 20 nt by varying the homology arm length of the ssODN between 47-27 nt. Data represents mean $\pm$ s.e.m with *p-value <0.05 , **p-value <0.01 , ***p-value <0.001 using Student's t-test. The results indicate SpCas9 outperforms Cpf1 nucleases at lower spacer length (17 nt), while at higher spacer length (20, 23 nt) Cpf1 nucleases perform better and the variation in homology arm length of the ssODN donor template did not affect the editing efficiency much.

PP1R12C respectively, refer figure 3.13).

Figure 3.12: Extent of incorporation of XbaI restriction site denoted by percentage correctly targeted at the two loci (a) CACNA1D and (b) PPP1R12C targeted using varying amounts of ssODN donors (100, 200, 300 ng). The results indicate that similar HDR frequency is observed irrespective of the amount of ssODN used in the experiment.

Figure 3.13: Rate of correct and incorrect incorporations for the XbaI restriction site is computed from the deep sequencing data for (a) CACNA1D and (b) PPP1R12C genomic locus. Extent of XbaI incorporation is computed from the NHEJ-mediated genome editing experiments as well where no donor template is provided and used as the baseline. Blue diamond corresponds to SpCas9, grey triangle to AsCpf1 and orange triangle to LbCpf1. The results indicate that the rate of incorrect incorporation is much lower in comparison to the rate of correct incorporations for both the locus examined for all 3 nucleases.

Further, HDR-mediated genome editing was also performed at 8 other selected loci

and the results from them are similar to those presented here. The details about the results from those loci can be found in [158]. Based on the trends observed here for HDR-mediated genome editing, the next question of interest was to investigate if the trend was a result of variation in cleavage efficiencies itself at the particular target site by the nuclease. But, the sequencing results confirmed that SpCas9 was capable of producing random indels at almost the same rate as that of AsCpf1 and LbCpf1 (refer figure 3.14). Therefore, confirming that the lower HDR efficiency "exhibited by SpCas9 in comparison to Cpf1 nucleases using ssODNs of non-target strand sequence was not due to" differences in cleavage efficiency at the site by the nucleases.

Figure 3.14: Extent of indel formation which is an indication of cleavage efficiency is computed from the deep sequencing data for (a) CACNA1D and (b) PPP1R12C target loci. The results indicate that SpCas9 was equally efficient as Cpf1 nucleases in introducing random indels at the two sites. Data is represented as mean $\pm$ s.e.m with * p-value < 0.05 , ** p-value < 0.01 , *** p-values < 0.001 using Student's t-test.

3.4.3 Optimization of ssODN donor templates

Previous studies have reported that the design of ssODN template can influence HDR efficiency [165] [166] [167] [168]. The experiments reported in the previous section were all based on ssODN that had the non-target strand sequence. Therefore, in order to investigate the effect on editing activity due to the orientation

of the donor template, ssODN complementary to both the target and non-target strand were examined for six loci with a homology arm length of 27 nt. The results fail to detect any consistent strand bias for each Cas nuclease (refer figure 3.15). Although, in five out of the six loci editing activity changes in the same direction when altering the ssODN orientation, therefore implicating "that each locus may have an inherent ssODN strand preference". For example, PPP1R12C shows a preference for ssODN with target strand sequence, while A11 locus shows a preference for ssODN with non-target strand sequence. Further, quantification of cleavage rates confirmed that "the change in HDR frequencies were not simple due to differences in cleavage rates" (refer figure 3.16).

Previous study reported that "the strand bias of SpCas9 at an AAVS1 genomic locus became prominent with longer donor templates" [169]. In order detect for any bias as reported previously, ssODNs with longer arm lengths (37 nt) were used to edit CACNA1D and PPP1R12C loci. The quantification from the deep sequencing data indicated the presence of bias, whereby SpCas9 exhibited higher HDR efficiency when using T ssODN at both loci, while Cpf1 nucleases preferred NT ssODN at both the loci (refer figure 3.17). Thereby, indicating that Cas9 and Cpf1 prefer donor templates of opposite orientations.

An earlier study reported enhancement in homology-directed editing by SpCas9 using asymmetric donors [170]. Further, to investigate if the structure of ssODN could impact the editing efficiency, asymmetric donors were designed by either extending the "PAM-proximal or PAM-distal side of the ssODN from 37 to 77 nt" as shown in figure 3.17 (a). These asymmetric donors were tested at CACNA1D and PPP1R12C loci. SpCas9 preferred a ssODN with longer PAM-distal side and target strand sequence, while ssODN longer PAM-proximal side either remained neutral or decreased the efficiency of HDR-mediated genome editing. These results were consistent with those reported earlier [170]. As opposed to SpCas9,

Figure 3.15: Extent of HDR-mediated genome editing using ssODN donor either complementary to target strand (denoted as NT as it has the same sequence as non-target strand) or non-target strand (denoted as T as it has the same sequence as target strand) having a homology arm length of 27 or 17 nt, targeting genomic loci (a) CACNA1D (27 nt), (b) PPP1R12C (27 nt), (c) A3 (27 nt), (d) A11 (27 nt), (e) B18 (27 nt), (f) B8 (27 nt), and (g) B8 (17 nt) quantified from deep sequencing data. The results indicate the absence of any consistent strand bias, although each locus may possibly have an inherent strand preference. Data is represented as mean $\pm$ s.e.m with * p-value <0.05 , ** p-value <0.01 , *** p-values <0.001 , N.S.: not significant using Student's t-test.

Figure 3.16: Extent of indel formation which is an indication of cleavage efficiency is computed from the deep sequencing data for (a) CACNA1D (27 nt), (b) PPP1R12C (27 nt), (c) A3 (27 nt), (d) A11 (27 nt), (e) B18 (27 nt), (f) B8 (27 nt), and (g) B8 (17 nt) target loci. The results indicate that 3 nucleases were equally efficient in introducing random indels at all the sites irrespective of whether the donor supplied was NT or T ssODN. Data is represented as mean $\pm$ s.e.m with * p-value <0.05 , ** p-value <0.01 , *** p-values <0.001 using Student's t-test.

Cpf1 nucleases preferred a ssODN with longer PAM-proximal side and non-target strand sequence, while longer PAM-distal side decreased the HDR efficiency. Also, comparing the HDR efficiency by using different ssODN designs it is observed that LbCpf1 performs the best at CACNA1D (refer figure 3.17 (b)) and SpCas9 performs the best at PPP1R12C (refer figure 3.17 (c)). Together, "the results indicate that SpCas9 or LbCpf1 may be used for HDR-mediated genome editing using ssODN", although the strand preference and the structure of donor template need to be carefully considered.

3.4.4 "Enhancement of error-prone repair with long single-stranded DNA"

Deep sequencing of HDR-mediated genome editing experiments provided an opportunity to monitor the effect of the presence of different donors on the cleavage efficiency of the Cas nucleases. Presence of single-stranded DNA of length between 40 to 100 nt (corresponding to ssODN with homology arm length of 17 - 40 nt) did not impact the cleavage efficiency much as is seen from figures 3.14, 3.16 and 3.18. But, the presence of a longer asymmetric ssODN (120 nt long) increased the cleavage efficiency for all the nucleases at two loci examined (CACNA1D, and PPP1R12C), with more prominent effect for ssODN having a longer PAM-proximal arm (refer figure 3.19).

The increased cleavage efficiency observed here could not be the reason for increased HDR frequency with the optimized ssODN donor templates as suboptimal donors that caused a decrease in HDR frequency also resulted in increased cleavage efficiency (refer figure 3.17). Further, a non-homologous 127-mer single-stranded DNA was reported to stimulate gene disruption by SpCas9 in an earlier study [171]. Therefore, these results suggest a possible way to enhance gene knockout efficiency "by introducing long ssODN donor containing a frameshift or a nonsense mutation"

Figure 3.17: (a) Different ssODN donor template designs tested that include template complementary to target strand denoted as NT, or complementary to non-target strand denoted as T, with varying arm lengths (symmetric - 37/37 nt, asymmetric - 37/77 nt or 77/37 nt). The restriction site designed to be incorporated at cut site is shown in green; Extent of HDR-mediated gene editing by SpCas9, AsCpf1 and LbCpf1 using the designed ssODN donors at (b) CACNA1D; and (c) PPP1R12C loci. The results indicate opposite strand preference for the two nuclease at both the loci. Data is represented as mean $\pm$ s.e.m with * p-value < 0.05, ** p-value < 0.01, *** p-values < 0.001 using Student's t-test.

Figure 3.18: Cleavage efficiency of the CRISPR-Cas systems measured as percentage indels introduced by them at (a) A12 locus in the presence/ absence of symmetric ssODN with homology arm length of 47 nt; and (b) B8 locus in the presence/ absence of symmetric ssODN with homology arm length of 27 nt. The results show that the presence of short single-stranded DNA (40-100 nt) has no effect on the cleavage efficiency. Data is represented as mean $\pm$ s.e.m with * p-value <0.05 , N.S.: not significant using Student's t-test.

that could inactivate the target gene not only through "NHEJ repair pathway but also by the HDR pathway using an optimized single-stranded DNA donor".

3.4.5 "Performance of CRISPR-Cas systems in HDR-mediated genome editing with plasmid donor"

Also, experiments were performed to evaluate the performance of SpCas9, AsCpf1 and LbCpf1 in HDR-mediated genome editing using linearized plasmids as donors at two loci, namely CLTA and GLUL. The results from these experiments indicate that SpCas9 out-performs Cpf1 nucleases in precise genome editing when linearized plasmids were used as donors at these two loci. The details of the experiments and the results pertaining to them can be found in [158].

Figure 3.19: Extent of indel formation at CACNA1D genomic loci using ssODN of different orientations, structure (symmetric/ asymmetric) and arm lengths denoted by (a) bar plots; (b) box plots by pooling the rates depending on ssODN structure and length; Rate of indel formation at PPP1R12C genomic loci using ssODN of different orientations, structure (symmetric/ asymmetric) and arm lengths denoted by (c) bar plots; and (d) box plots by pooling the rates depending on ssODN structure and length. Data in barplots represent as mean $\pm$ s.e.m with * p-value <0.05 , N.S.: not significant using Student's t-test. The rates were pooled for T and NT ssODNs as the difference between the respective indel frequencies were insignificant. The results indicate long single-stranded DNA with longer PAM-proximal arm had maximum positive impact on the frequency of indel formation.

3.5 Discussion

Despite CRISPR-Cas systems having the potential for engineering complex genomes, the critical choice to make frequently is which particular system to use. Comparing performance of different systems based on previous published studies is unreliable due to many confounding factors that include variation in target sites, cellular context, experimental procedures adopted, etc. Thus, SpCas9 is chosen as default in most cases. Here, in this study five distinct CRISPR-Cas enzymes are compared over a range of target sites to gain insights into their functioning and performance efficiencies in both NHEJ-mediated and HDR-mediated genome editing. Based on the results from the study, a set of guidelines are derived that can guide in

selecting the appropriate system depending on the application as shown in figure 3.20.

Figure 3.20: Flowchart to guide the choice of CRISPR-Cas system for a particular experiment depends on different considerations that include type of genome modification desired, characteristic of the target site, donor template and its properties, etc.

For shorter spacer length constrained between 17-20 nt, SpCas9 exhibited the highest cleavage efficiency at most sites examined in comparison to other nucleases as shown in figure 3.9. While, for longer spacer length of 21-23 nt, SaCas9 and LbCpf1 out-performed the other nucleases. A pairwise comparison of the performance of SpCas9 and LbCpf1 with "only the sgRNAs of optimal lengths for the two enzymes" indicates that they generate indels at comparable frequencies in genes with varying expression levels (refer figure 3.21). Thus, either of these enzymes can potentially be used to generate routine gene knockouts via NHEJ pathway. But, in situations where target specificity is an issue while targeting repetitive genomic region or genes with close paralogs, AsCpf1 can be a better choice "due to lower tolerance to mismatches between sgRNA and the target DNA" (reference to figure in results). Also, SaCas9 may be used due to its requirement for longer spacers (atleast 21 nt).

Figure 3.21: The frequency of indels generated by SpCas9 and LbCpf1 considering only the sgRNAs of optimal length for both (19-22 nt inclusive) using deep sequencing analysis indicate comparable performances for both the nucleases in both lowly and highly expressed genes.

High cleavage efficiency along with high target specificity are the desired qualities for the model Cas endonuclease that can be used in future applications. Although, "two recent studies reported that AsCpf1 and LbCpf1 satisfy both" the criteria, the findings from this study indicate that "a trade-off between editing activity and target specificity exists in naturally occurring Cas enzymes" (refer figure 3.10) [172] [173]. Despite showing higher editing efficiency both SpCas9 and LbCpf1 also have higher tolerance for mismatches between the sgRNA and target site which means they are less specific, while AsCpf1 shows lower editing efficiency and higher specificity in comparison to SpCas9 and LbCpf1. The genome-wide methods used in the two studies may possibly have some limitations that lead to the non detection of certain off-target sites cleaved by Cpf1 [172] [173]. The detection of cut sites by digenome-seq requires very high sequencing depth [174] [175]. Incorporation of blunt double-stranded oligodeoxynucleotides (dsODN) at the cut site is required for detection of cleavage site by GUIDE-seq but this is biased against the staggered cut generated by the Cpf1 [176]. And, the tag integration efficiency was reported to be higher for SpCas9 than LbCpf1 and AsCpf1 [177]. Therefore, more thorough investigation with more sensitive off-target detection

methods might give us insights into the relationship between editing activity and target specificity of these Cas enzymes.

Precise genome editing through HDR occurs at low frequencies and thus, an improvement in these frequencies are essential before CRISPR-Cas systems can be utilized for gene therapies to provide permanent cure. Over the years, a lot of efforts have been done for improving the efficiency [165] [167] [178] [179] [180] [181] [182] [183] [184]. The results from the current study indicate that Cpf1 nucleases are more efficient than SpCas9 in introducing precise genome edits in human cells while using ssODN as donor templates (refer figure 3.11 and figure 3.17). Cpf1 and Cas9 show preference for ssODN of opposite orientation and the further improvement is observed for Cpf1 when ssODN with longer PAM-proximal arm is used. These results lead us to hypothesize that "Cpf1 may asymmetrically release the 3' end of the cleaved target strand, in turn allowing the shorter arm of the donor template to anneal and thereby, resulting in the longer arm of the template to invade and displace the base-paired non-target DNA strand at the other side of the break". The results from this study in human cells are comparable to those from another recent study in zebrafish embryos [185].

The nature of cut, staggered, induced by Cpf1 nucleases and the position of the cut-site that lies outside the critical seed region resulting in possible repeated targeting tend to bias the contemplation in favor of Cpf1 nucleases performing better than SpCas9 in precise genome editing. Infact, the results of this study does indicate that Cpf1 nucleases (both AsCpf1 and LbCpf1) out-perform SpCas9 while ssODN of optimal design is used at most sites examined (refer figure 3.15). However, other confounding factors might exist like localized bias due to chromatin structure leading to inherent strand preference at a locus (as observed for PPP1R12C loci for SpCas9 with ssODN complementary to non-target strand). Therefore, further work is needed to carefully uncover the underlying mechanisms for the preferences

identified from this study.

3.6 Conclusion

A systematic analysis of "the ability of different Cas endonucleases to perform NHEJ and HDR-mediated genome editing in human cells" is carried out. Over 90 sites were targeted to obtain a clear picture of the performance and in order to have a fair comparison. A computational pipeline is developed along with parameter fine tuning to analyze the huge data generated from Illumina deep sequencing in order to quantify the performance of the various Cas endonucleases. The extensive study and quantification enabled development of set of rules and design parameters that may be used in future for other genome editing experiments with CRISPR-Cas (refer figure 3.20).

Preface to Chapter 4

A combination of experiments with computational techniques developed to analyze the Illumina deep sequencing data from the experiments helped compare the performance of 5 CRISPR-Cas systems and characterize their functioning. Each of the systems worked best in certain conditions and were ideal for certain applications as characterized by the study described in chapter 3. This work, thus, helped create a guideline for using these 5 systems to their optimal best as is characterized in figure 3.20.

Further, to enhance our understanding of the functioning of these enzymes and to guide proper utilization of them, the next study probed the outcomes from the NHEJ-mediated genome editing. By definition, the NHEJ-mediated genome editing is known to introduce random insertions and deletions while repairing the break in DNA, as no homology template is provided. Analyzing the outcomes from the NHEJ-mediated genome editing and identification of any preference for particular patterns can prove to be very useful in designing future genome engineering experiments.

Chapter 4

Characterizing non-random mutations introduced by CRISPR-Cas systems in NHEJ-mediated genome editing

The mutations introduced by the CRISPR-Cas enzymes during NHEJ-mediated genome editing were revealed to be non-random. Increasingly, efforts are to identify the mutations introduced and identify patterns in them. In this present study, the experiments performed in the previous study (described in chapter 3) are utilized and subjected to a computational pipeline developed to identify indel patterns introduced by the 5 enzymes. These patterns can then be used in designing future genome engineering studies and also for developing a predictive tool that can be used to predict the mutational outcome of targeting a new locus using these enzymes.

4.1 Background

CRISPR-Cas system's ability to perform genome editing arises from its capability to be guided by a RNA to a desired genomic location and cleave it causing a double-stranded break (DSB) [49]. This DSB then leads to the recruitment of the cellular DNA repair machinery to fix the DNA damage. Two major DNA repair pathways are involved in the repair of these CRISPR-induced double-stranded breaks that include non-homologous end joining (NHEJ), and microhomology mediated end joining (MMEJ) [186] [187]. In the former, the broken ends are re-ligated but often errors (insertions/ deletions) are introduced in the process, while in the latter a short (2-25 bp) microhomologous sequence is used for aligning the broken ends before ligation and results in the deletion of flanking the cut-site [186] [187]. A number of factors influence the choice of repair pathway employed including cell cycle stage and availability of repair enzymes [188] [189]. Further, the type of mutations at the cut-site following the repair is also modulated by numerous factors including the repair pathway employed, endonuclease used, chemical or genetic modulation of DNA repair enzymes, cell cycle, level and duration of endonuclease treatment, genomic context of the target site, and nucleotides surrounding the DSB [165] [190] [191] [192] [193] [194] [195] [196] [197] [198].

Although, the key components and mechanisms of these DNA repair pathways have been characterized, their interplay with the CRISPR endonucleases and the mutations introduced are yet to be characterized. Despite, these error prone repair pathways being used only in a minority of CRISPR-mediated genome engineering applications that include disrupting the function of genes or regulatory elements by introducing short indels in a targeted fashion or recording cell lineage or molecular events by introducing irreversible changes, it is inevitable to characterize the mutations introduced [149] [151] [199] [200] [201] [202] [203] [204] [205]. Thus, understanding the behaviour of the endonucleases and their interplay with the DNA

repair machinery is critical as this would help maximize the potential and safety of CRISPR technology in achieving precise and predictable genome editing.

Previously, studies have been performed to characterize the outcomes of end-joining pathways for specialized nucleases including RAG-induced V(D)J recombination, I-SceI meganuclease, and HO endonuclease [206] [207] [208] [209] [210] [211] [212] [213]. The difference between these studies and designing a similar study for CRISPR-endonucleases is that with the former nucleases the range of sequences for which DSB repair outcomes were measured was limited but for the latter due to the ability for the CRISPR-endonuclease to target to any site in a genome containing a protospacer-adjacent motif (PAM) large scale analysis throughout the genome is required.

Previous studies have employed Bioinformatics analysis to uncover the mutational patterns introduced by Cas9 activity. Many of these highlight the importance of the sequence context [191] [194] [197]. Bae et al. demonstrated that 52.7% of all deletions introduced by Cas9 in human cells are associated with microhomologies flanking the site of DSB [191]. Van Overbeek et al. studies on multiple cell lines including HEK293, HCT116, and K562 cells revealed that genome-editing outcome are non-random and change over time [194]. Additionally, these studies have lead to development of predictive model using machine learning and development of a suite of computational tools that can predict outcomes from Cas9's activity [198] [214].

Most of the above mentioned efforts in understanding the mutational outcomes of CRISPR-endonuclease are restricted to SpCas9. In the present study, I present a computational pipeline developed to accurately identify the indel patterns and preliminary results of observed patterns of mutations from outcome of 5 CRISPR-Cas enzymes in error-prone genome editing.

4.2 Pipeline developed

The NHEJ-mediated genome editing experiments performed using the 5 CRISPR-Cas systems in our previous study described in chapter 3 are utilized in this study to probe for mutations in the error-prone genome editing outcomes. The deep sequencing data from the experiments is subjected to the developed computational pipeline to first identify the indels in each read and then calculate their frequency of occurrence and thereby, identify the prominent repetitive patterns for each of the 5 endonucleases.

In order to identify indel events, a method was required to distinguish one indel from another occurring in a read or different reads. So, it was decided to use three properties to label/ characterize each indel:

- Indel type: Insertion or Deletion
- Indel position: The start position of the particular insertion or deletion. The position is labeled using the cut-site as the reference point as illustrated in figure 4.1.
- Indel length: The length of the particular insertion or deletion

Therefore, by using this nomenclature each read is labeled using a string "Indel-Type_IndelPosition_IndelLength" and in case more than one indel occurs in the same read then we use a concatenation of the strings describing each indel.

Figure 4.1: Schematic describes nomenclature used for naming the indel start position for (a) Cas9 enzymes and (b) Cpf1 enzymes. The cut-site is denoted by the black line. The positions upstream the cut-site are labeled as U1, U2,...,Un, while those downstream the cut-site are labeled as D1, D2,...,Dn. For Cpf1 as the enzyme generates a staggered cut, the positions in the staggered cut region are labeled as C1, C2,..., C5.

With the labeling nomenclature in place, the next step was the development of

the pipeline for identification of the indel and using the decided nomenclature to name them. The first part of the pipeline is similar to the one developed in chapter 3 for quantification of performance in NHEJ-mediated genome editing. The first step being mapping the sequenced reads to a local reference library built using the amplicon sequences of the target genes. The sequenced reads in this case too were mapped using BWA-MEM to the local reference library with parameters as identified in the previous study described in chapter 3 (match score = 1, mismatch penalty = 2, gap open penalty = 6, gap extension penalty = 1, clipping penalty = 8) [8]. The uniquely mapped reads having a mapping quality ≥ 20 were sorted and Picard was used to assign group information [160]. Local realignment and recalibration of base quality scores was performed using GATK toolkit [12] [13] [14]. The next step involved identification of indels in the reads using CIGAR string, an output parameter from the mapping tool, as described in chapter 3. Again, in this case too only a 40bp length span is considered on either side of the spacer for identification of indel patterns introduced as a result of CRISPR-Cas induced editing. Using a self developed script, the reads are then labeled based on the indel occurring in them using the nomenclature described above by identifying the indel type, their start position and length from processing their CIGAR string.

Further, to process for recurrent/ prominent indel patterns grouping was done as combinations of individual patterns possible were very high. Therefore, the indels were grouped into 5 classes based on indel type and length of indel, namely:

- Insertion Small - Insertion of length 2 or less
- Insertion Long - Insertion of length 3 or more
- Deletion Small - Deletion of length 2 or less
- Deletion Long - Deletion of length 3 or more
- Complex - Read contains more than one indel event occurring simultane-

ously.

The frequency of occurrence for each indel class is calculated for each sample. But, in order to identify recurrent or prominent patterns some form statistical test needed to be employed as the rate of indel introduction varied for different enzymes and also locus. Therefore, in this case hypergeometric test is chosen to help compute over-representation of a particular indel class for an enzyme at a locus by pooling together information from replicates. In this test, a hypergeometric distribution is employed to calculate the statistical significance of observing a specific number of occurrences of an indel class (desired outcomes) for an enzyme out of total indels observed (total outcomes) for that enzyme, given the total number of occurrences of that indel class in all enzymes. The "*hypergeom*" function from the Scipy Python library is employed for this purpose after pooling the information from all replicates relevant each case [215].

A schematic of the pipeline developed is illustrated in figure 4.2.

4.3 Results and discussion

The pipeline described in the above section is used to label the sequenced reads depending on the indel in them and group them into classes. The frequency of occurrence for each class is calculated sample-wise and combined over replicates at a locus for an enzyme. A hypergeometric test is performed on this consolidated frequency of occurrence of each class for an enzyme at a locus to identify the over represented classes at the particular target locus for an enzyme. The following sections details the preliminary results and biological significance based on the results from the hypergeometric test.

Figure 4.2: A schematic of pipeline developed to identify indels and prominent patterns in NHEJ-mediated genome editing using CRISPR-Cas enzymes.

4.3.1 Insertions are prominent signature repair outcomes while using SpCas9

Of the 5 CRISPR-Cas enzymes compared (AsCpf1, LbCpf1, NmCas9, SaCas9, and SpCas9) at over 90 locus, it is observed that SpCas9 is the lone enzyme where insertions are prominent at a majority of locus (in ~50% of locus examined) as shown in figure 4.3. The other 4 enzymes introduce deletions more frequently with AsCpf1 being the top (in ~80% of locus examined). Additionally, complex indel patterns (representing more than one indel event occurring in the same read) are the least frequent for all enzymes as expected, with LbCpf1 and SpCas9 showing the maximum probability in about 20% of the locus examined.

Figure 4.3: The percentage of locus where each of the CRISPR-Cas enzymes examined show a preference for introducing a insertion or deletion or a complex indel (more than one indel in the same read). Results indicate SpCas9 is most likely to introduce insertions, while AsCpf1 shows highest likelihood for introducing deletions.

4.3.2 Prominent indel classes identified from the mutations introduced by the CRISPR-Cas enzymes vary with the enzymes

Based on the indels identified in the reads, they were grouped into different classes depending on the indel type and their length as described in section 4.2. The results indicate that the prominent indel classes found in the mutations introduced by the enzymes are variable (refer figure 4.4). As mentioned in the previous section, AsCpf1 tends to introduce deletions more frequently and these were more often found to be of length longer than 2 bases. Also, insertions amounted to less than 10% of the introduced mutations. While, with LbCpf1, though longer deletions are the most prominent class (about 29%), small deletions (deletions of length less than or equal to 2 bases) and long insertions also seemed to be found with high prominence (25% and 21% respectively). With NmCas9, mutations of shorter length (2 bases or less) were more common about 74% of the times and nearly two-thirds of this was deletions (~51%) and rest was insertions (~23%). SaCas9 also introduced smaller deletions with highest prominence (about 45%), while small insertions and longer deletions were also observed at notable rates (about 17% and 19% respectively). SpCas9 as mentioned in the previous section introduced insertions about 50% of the time and smaller insertions were found to be of greater prominence here (about 36%). Also, small deletions were introduced with notable rates. LbCpf1 and SpCas9 were also found to introduce complex indel class (with simultaneous indels at more than one location in a read) with notable rates (18% and 19% respectively). In totality, the patterns observed varied highly depending on the enzyme, although longer mutations seemed to be more prominent with Cpf1 enzymes, while smaller ones were prominent with Cas9 enzymes.

The results from our study are consistent with previous report on SpCas9. Taheri-

Figure 4.4: The percentage of locus where each of the CRISPR-Cas enzymes examined show a preference for introducing a insertion or deletion or a complex indel (more than one indel in the same read). The results indicate SpCas9 is most likely to introduce insertions while AsCpf1 shows highest likelihood for introducing deletions.

Ghahfarokhi et al. also report that small mutations comprise about 60% of the mutations introduced by SpCas9 and our results also indicate that with SpCas9 small insertions and deletions are the most prominent at about 55% of the sites examined (refer figure 4.4). High frequency for single base insertion by SpCas9 has previously been reported in [198] [216] and in our current study we observed small insertions as the most prominent indel class introduced at majority of the locus examined. Although, some previous studies report deletions are slightly more prominent than insertions for SpCas9, but our current results indicate otherwise. Some reasons for this could be the variation in locus examined in the studies as

it is known that the patterns of mutations are sequence context dependent or variation in duration of treatment that is known to have an effect on the indels introduced [194].

4.3.3 Correlation between indel classes introduced and spacer length of the locus

In our present study we targeted locus with varying spacer lengths and hence, we sought to investigate any correlation between spacer length and the indel classes introduced by the enzymes. And grouping the indel classes introduced based on the spacer length for each enzyme does reveal variation in patterns created (refer figure 4.5). At higher spacer lengths where Cpf1 enzymes are known to be active (based on our results mentioned in chapter 3), long deletions seem to be the most prominent indel class introduced; while at shorter spacer lengths small deletions are found more often. With NmCas9, the activity was predominantly poor at most of the targeted sites (as mentioned in the results in chapter 3), but at spacer lengths closer to native spacer length small deletions seem to be more prominent. For SaCas9, small deletions are the over-represented indel class at most locus examined. Short insertions are the most prominent indel class irrespective of the spacer length for SpCas9, while mixed signals are observed for the other indel classes at lower spacer lengths where the enzyme is active.

Although, these results indicate some variability, but the more comprehensive analysis is required to ensure that a correlation exists directly between the spacer length and indel patterns created, rather than this being a confounding effect due relationship between activity of the enzyme and spacer length.

Figure 4.5: The number of locus where each of the CRISPR-Cas enzymes examined shows a preference for the indel classes grouped based on spacer length of the locus targeted. The results indicate some variability when the locus are grouped based on spacer length.

4.4 Future work

The preliminary results from the present study has been successful in identifying indels in reads and preferred patterns based on categorizing the indels into different groups. But, probing for specific indel patterns within each prominently expressed groups could highlight preferred indel introduced as well for each enzyme. In the present study, a comparison based on indel type and length is performed but one can also explore the possibility of identifying preferred position for introducing indel, though intuitively it is expected to be closer to the cut-

site. Further, utilizing NGS data previously reported from other studies deposited in the National Centre for Biotechnology Information (NCBI) - Sequence Read Archive (SRA; <https://www.ncbi.nlm.nih.gov/sra/>), and machine learning techniques one can create a tool for predicting the indel outcomes for the 5 different CRISPR-Cas enzymes.

4.5 Conclusion

With reports that the NHEJ-mediated genome editing outcome using the CRISPR-Cas enzymes are non-random, a computational pipeline is developed to analyze the NGS data generated from NHEJ-mediated genome editing performed by 5 CRISPR-Cas enzymes (as described in chapter 3). The indel patterns in the sequenced reads were identified and grouped in to classes to identify prominent indel classes generated using a particular enzyme. The preliminary results reported in this chapter indicate a preference for each of the enzyme, while small indel patterns (2 bases or less in length) are more commonly introduced by Cas9 enzymes, longer indel patterns (3 or more bases) are more common with Cpf1 enzymes. Further, analysis on specific indel patterns and utilizing machine learning techniques with larger dataset of sequencing data available from other studies can help design future genome engineering studies by utilizing the CRISPR technology safely and to its fullest potential.

Preface to Chapter 5

In pursuit of improving the understanding of the DNA editing system (CRISPR-Cas), the indel patterns introduced by the enzymes in the NHEJ-mediated genome editing were probed for as described in chapter 4. Recognition of these favored patterns not only enhances our understanding about these systems and their functioning but also serves as a critical aid in designing future experiments by exploiting these characteristics showcased by the specific CRISPR-Cas enzyme to our advantage. Further, though not performed in the current study, incorporating more sequencing data available online into the present pipeline to recognize the indel patterns in combination with machine learning techniques can guide development of a platform for recognition of the probable indel patterns that may be introduced by a particular enzyme at a given locus.

Succeeding the efforts to probing the functioning of the CRISPR-Cas system at different levels and using different techniques in the studies described in chapters 2, 3 and 4, the next attempt is made at using this CRISPR-Cas technology for a real-time application. A GLUL (glutamine synthase) knockout in HEK293 cells developed using CRISPR-Cas9 system is characterized for its utility to be used as a platform for stable production of a recombinant protein, erythropoietin.

Chapter 5

Characterizing glutamine synthase knock-out based on HEK293 for use as a human expression system for stable production of recombinant erythropoietin

A GLUL (glutamine synthetase) knockout is introduced in HEK293 cells using the CRISPR-Cas9 technology. The knockout is characterized using a computational pipeline designed to analyze the Illumina deep sequencing data for the production of recombinant protein. Further, the fully human cell line with a selectable and amplifiable marker is used for stable production of recombinant erythropoietin (EPO) with fully human glycosylation. This is significant as the existing and most utilized mammalian host systems that include Chinese hamster ovary (CHO) cells

as well as Sp2/0 and NS0 mouse myeloma cells, produce recombinant proteins with non-human glycans that can potentially illicit immunogenic response, while a fully human cell line like the one developed here is the most preferred expression systems for manufacture of complex therapeutics and recombinant proteins with glycosylation patterns resembling those endogenous to human proteins.

5.1 Background

For over three decades, mammalian cells have been used for the production of recombinant protein therapeutics due to the requirement of post-translational modifications (PTMs) for activity [217]. Chinese hamster ovary (CHO) cells, baby hamster kidney (BHK) cells, Sp2/0 murine hybridoma cells, NS0 murine myeloma cells, C127 murine mammary gland cells, HT-1080 human fibrosarcoma cells and human embryonic kidney cells 293 (HEK293) are the mammalian cell lines approved so far for the manufacture of recombinant protein therapeutics [218]. Predominantly, CHO cells are used as mammalian host cell lines for the manufacture of recombinant protein therapeutics because of their advantages that include their ability to confer human-like glycosylation features to glycoprotein products, capability of high protein productivity with titers up to grams per liter due to established stable cell line generation strategies, ability to propagate well in serum-free single-cell suspension cultures enabling ease of scale-up, refractory to a number of human viruses providing lower biosafety risks, and finally their established regulatory track records for the manufacture of recombinant protein therapeutics [219] [220] [221] [222]. Nevertheless, CHO cells have been reported to generate non-human glycan structures including N-glycolylneuraminic acid (Neu5Gc) and galactose- α -1,3-galactose group (α -Gal) on glycoprotein products which confers increased risk of immunogenicity [223] [224]. Also, CHO-derived glycoproteins lack certain human glycans, including bisecting GlcNAc, Lewisx, sialyl Lewisx, and

$\alpha(2,6)$ -linked sialic acid [222] [224]. Moreover, reports suggest CHO cells may not adequately produce other PTMs (such as glutamic acid γ -carboxylation) required for activity of some recombinant protein therapeutics [225].

Pharmaceutical companies have utilized various human cell lines, such as HEK293, HT-1080, AGE1.HN, CAP, HKB11 and PER.C6 for glycoprotein production in efforts to obtain more human-like protein therapeutics [218] [224]. While products from AGE1.HN, CAP, HKB11 and PER.C6 are still at different phases of preclinical and clinical development, several protein therapeutics from HEK293 and HT-1080, including recombinant factor VIII-Fc, Dulaglutide, Idursulfase and Velaglucerase alfa have already been approved by the US Food and Drug Administration (FDA) and European Medicines Agency (EMA) [218]. These regulatory approvals for HEK293 and HT-1080 demonstrate successful circumvention of potential biosafety risks from human virus contamination and provide a useful basis for future assessment of other therapeutics produced using human cell lines.

HEK293 cells serve as an attractive expression system for recombinant protein production as they offer multiple advantages besides being able to generate human glycosylation profiles and having an established regulatory track record. First, they are exceptionally efficient in glutamic acid γ -carboxylation and tyrosine sulfation that is required for some therapeutic products like Drotrecogin alfa and recombinant factor IX-Fc [218] [225] [226] [227]. Second, their ease of manipulation and ability for rapid production of recombinant proteins via transient gene expression [228] [229]. Third, they can be used for stable recombinant protein production, which has been reported for IFN α 2b using HEK293-6E cells with titers of up to 333 mg/L and antibodies using HEK293F cells with titers of up to 600 mg/L [230] [231].

Despite the potential of HEK293 as a human bioproduction cell line, stable recombinant protein production in HEK293 has been achieved mainly using xenogeneic

genes as selection markers. For example, HEK293-6E cells constitutively expressing the non-native Epstein–Barr virus (EBV) EBNA1 protein have been utilized for the amplification of expression vectors with a suitable viral origin of replication [230]. Antibiotic selection strategies involving resistance cassettes from other organisms have also been deployed. The usage of xenogeneic genes will result in higher expression of non-human proteins in the HEK293 producer cell lines on top of the E1 adenoviral genes that the HEK293 expresses, and potentially trace amounts of these proteins can contaminate the final product. Such contamination may trigger allergic and adverse reactions, and/or immunogenicity to the product. To develop a bioproduction system with low expression of heterologous proteins, the use of human glutamine synthetase (GLUL) gene as a selection marker for the development of producer cell lines is explored. Glutamine is an amino acid required for the biosynthesis of several amino acids, pyrimidines and purines, and it can be synthesized from glutamate and ammonia by GLUL or can be exogenously supplemented to culture medium to support cell growth. As a selection marker, GLUL expression becomes essential in the absence of glutamine in the culture media. Coupled with the expression of the trans gene-of-interest and upon addition of sub-lethal doses of methionine sulfoximine (MSX), a GLUL inhibitor, cells expressing high levels of GLUL with simultaneously increased expression of the target transgene were selected for. This GLUL-mediated gene selection strategy was first demonstrated in CHO cells, and subsequently implemented in HEK293E cells but with low titers [232] [233] [234].

Here, CRISPR-Cas9 system is employed to knockout the endogenous GLUL gene in HEK293 cells. In this study, a computational pipeline is developed to characterize the knockout cells developed for utility as a bioproduction platform for the production of human erythropoietin (EPO) as a model product. The characterization further enabled development of these knockout cells as a bioproduction

platform.

5.2 Design of the experimental study

In order to prevent the endogenous GLUL protein from interfering with the gene selection strategy, a knockout of the native GLUL gene in HEK293 was performed using CRISPR-Cas9 system. The knockout was verified using qPCR (quantitative real-time PCR) and monitoring growth of cells in media with and without glutamine-supplemented media. The cells were subjected to illumina deep sequencing. A computational pipeline is developed in order to analyze the data from sequencing and thereby characterize the cells for utility as a platform for production of glycoproteins.

5.3 Pipeline developed

The GLUL knockout cell lines were subjected to illumina deep sequencing and the objective is to look at the gene expression variation between different conditions in order to develop the knockout cell lines to be used effectively as a bioproduction platform. The wildtype and GLUL knockout cell lines were also grown in two different culture conditions (adherent and suspension). Cell lines grown in each of the four conditions were sequenced. The reads were mapped to human genome assembly hg19 using STAR with maximum number of mismatches allowed per pair (*outFilterMismatchNmax*) assigned to 6 and maximum number of mismatches per pair relative to read length (*outFilterMismatchNoverLmx*) assigned to 0.05 [9] [10]. The next step being to compute the gene expression observed in the different samples and the *summarizeOverlaps* function is used to obtain the raw counts from the aligned reads. Further, DESeq2 package is used to calculate the normalized gene expression counts and perform the differential gene expression

analysis [235]. With list of differentially expressed genes in hand it was essential to know what phenotypical/ functional effect they would have. In order to understand that enrichment analysis is performed on the differentially expressed genes to get insights about their functional profile using DAVID [236]. The pipeline developed is summarized in figure 5.1.

Figure 5.1: The pipeline developed to analyze the deep sequencing data for characterizing the GLUL knockout cell lines.

5.4 Results and biological inferences

To gain insights into the molecular changes in the GLUL knockout clones and during adherent and suspension cultures, the transcriptomes of GLUL knockout cell lines and original wildtype HEK293 cell lines during both adherent and suspension cultures are analyzed using their RNA-Seq on the Illumina platform.

5.4.1 Gene expression based classification of cell lines

Principal component analysis (PCA) based on the gene expression reveals the samples can be clearly separated based on experimental conditions (refer figure 5.2). Further, gene expression clustering based on Euclidean distances, as shown in figure 5.3, also clusters the samples based on experimental conditions. Both PCA and clustering results also indicate that the differences between the wildtype and knockout cell lines are larger than the differences between adherent and suspension cell cultures.

Figure 5.2: The principal component analysis based on gene expression for the wildtype HEK293 and GLUL knockout cell lines in adherent and suspension cell cultures showing samples segregated based on experimental condition.

Figure 5.3: Heatmap based on euclidean distances calculated from gene expression values of the wildtype HEK293 and GLUL knockout cell lines in adherent and suspension cell cultures indicating their clustering. The clustering shows samples with same experimental conditions are more similar to one another than those from other conditions and that the differences between the wildtype and GLUL knockout is more prominent than the differences between adherent and suspension cell cultures.

5.4.2 Differentially expressed genes between wildtype and GLUL knockout cell lines

The differentially expressed genes between wildtype and GLUL knockout cell lines were compared by pairing the corresponding adherent and suspension samples with one another using DESeq2. A total of 915 genes were significantly up-regulated, while 1,142 genes were down-regulated in the GLUL knockout cells using a cutoff of 0.0001 for adjusted p-value. The p-value is fairly stringent as it was a 8 vs. 8 comparison and a random checking was done to verify the significance of the genes

picked using the cutoff. Gene Ontology (GO) analysis performed using DAVID, as shown in figure 5.4, revealed that the up-regulated genes were significantly enriched for functions related to protein translation, possibly in response to the loss of GLUL, a key enzyme involved in amino acid metabolism. Additionally, the down-regulated genes were significantly enriched for functions related to the extracellular matrix, cell adhesion, cytoskeleton, and interferon signaling.

Figure 5.4: GO analysis of the genes that were significantly (a) up-regulated and (b) down-regulated in the GLUL knockout cells when compared to the wildtype HEK293 cells.

5.4.3 Differentially expressed genes between adherent and suspension cell cultures

In order to understand the adaptation when moving from adherent to suspension cultures, a comparison was done between them pairing the corresponding wildtype and GLUL knockout samples. 549 genes were significantly up-regulated, while 596 genes were down-regulated in the cells grown in suspension culture in comparison to adherent (adjusted p-value < 0.0001). GO analysis on these differentially expressed genes, as shown in figure 5.5, reveal that functions related to cell membrane and RNA metabolism are up-regulated in suspension cells, perhaps reflecting the fact that these cells were no longer attached to a surface and therefore required remodeling of the cell envelope. While, functions related to cell-cell adhesion and cell division are enriched in the down-regulated suspension cells, which may account for the slower growth rates observed in these cultures.

5.4.4 Glycoprotein biosynthesis in GLUL knockout cell lines

Glycoprotein synthesis related pathways are not picked in the GO analysis comparison between wildtype and knockout cells which indicates that the genes involved in this pathway were not differentially expressed between these two conditions. Further, in order confirm that the knockout cells did not have any impact on the glycoprotein synthesis, the genes involved in glycan biosynthesis were retrieved from KEGG pathways and their expression is compared in both the wildtype and knockout cells. A heatmap of the expression, as shown in figure 5.6, indicates that predominantly the genes have similar expression profile in both the conditions (wildtype and knockout). Thus, confirming that the glycoprotein biosynthesis is unaffected in our knockout cells.

Collectively, the transcriptome analysis uncovered the molecular changes that were

Figure 5.5: GO analysis of the genes that were significantly (a) up-regulated and (b) down-regulated in the suspension cells when compared to the adherent cells.

consistent with observed phenotypes and no strong gene signatures that might indicate alterations in protein glycosylation in the GLUL knockout cells were observed, thereby, making them suitable for the use as a platform for the manufacture of glycoproteins.

5.5 Conclusion

CRISPR-Cas9 system was employed to generate a GLUL-knockout in HEK293 cells in order to prevent endogenous GLUL protein from interfering with the gene selection as observed in a previous report [233]. Further, an analysis of the RNA-

Figure 5.6: A heatmap of the expression of genes involved in glycan biosynthesis showing the expression is similar in both wildtype and knockout cells.

Seq of the cells (as detailed in this chapter) helped identify the molecular changes associated and reveal the cells to be ideal for glycoprotein production. Thus, confirming their ability to be used as suitable bioproduction platform for glycoproteins. Following the characterization of the knockout cells, these cells were successfully demonstrated to be used for the production of EPO without non-human glycan epitopes.

Preface to Chapter 6

In order to showcase the application of CRISPR-Cas system, GLUL knockout cell lines based on HEK293 cells were developed. Using RNA-Seq data, these cells were characterized and found to be ideal for developed as a bioproduction platform for stable production of recombinant erythropoietin (EPO) with fully human glycosylation. This is of great usefulness as the fully human glycosylation provides an advantage over the other existing production platforms that produce recombinant proteins with non-human glycans that can potentially illicit immunogenic response.

In continuation to the efforts for studying and understanding the functioning of editing system, another main aspect that can be explored is regulation of the activity. In order to probe the regulation of editing, the RNA editing mediated by ADAR, was selected as an ideal study system. The next study describes the probing for RNA editing regulators, using a genome-wide computational study in combination with experiments that helped elucidate splicing factor, SRSF9, as one of the key regulators of RNA editing, thereby, giving rise to the highly complex editing landscape.

Chapter 6

Genome-wide editing analysis identified SRSF9 as selective repressor of ADAR2-mediated editing of brain-specific sites in primates

Although, the "human genome encodes only two catalytically active RNA editing enzymes (ADAR1 and ADAR2)", diverse spatial patterns of adenosine-to-inosine (A-to-I) editing is observed across different tissues. This indicates a role of other regulatory factors in shaping up the editing landscape. Here, in this study, splicing factor SRSF9 is identified to control the editing of many brain-specific sites in primates. A series of experiments enabled identification of SRSF9 as a robust repressor of A-to-I editing by ADAR2. RNA-Seq and transcriptome-wide location analysis revealed 1328 sites (predominantly brain-specific) directly regulated by SRSF9. Further, the computational analysis helped reveal that cell viability is

maintained by SRSF9 by inhibiting ADAR-2 "mediated editing of genes involved in proteostasis, energy metabolism, the cell cycle and DNA repair". Thus, highlighting "SRSF9 as an important editing regulator and suggesting potential roles for other splicing factors".

6.1 Background

Adenosine-to-inosine (A-to-I) is the most prominent RNA editing resulting in nucleotide substitution causing the RNA sequence to be altered. This alteration "can result in new protein isoforms, alteration in RNA splicing, influence RNA stability, alteration in microRNA targets,etc". Abnormal "editing in humans is known to contribute to various diseases, including multiple cancer types and neurological disorders" [237] [238] [239] [240] [241] [242] [243]. RNA editing, therefore, is not random and must be carefully orchestrated to ensure normal cell physiology. Further, two enzymes of the ADAR family (ADAR1 and ADAR2) that catalyze the deamination reaction in A-to-I RNA editing have been demonstrated to be essential for many biological processes. Previous reports confirm that "ADAR1 knockout mouse and ADAR1 editing-deficient mouse are embryonic lethal, while the ADAR2 knockout mouse dies from epileptic seizures within 20 days of birth" [244] [245] [246] [247] [248]. These results further confirm that tight spatiotemporal control of the RNA editing is essential for survival.

ADAR-mediated editing activities "are highly regulated at the transcriptional, post-transcriptional and post-translational levels" as they are important for normal physiology and contribute to diseases. Interferons induce the expression of ADAR1 p150 isoform (25,26). CREB transcription factor stimulates expression of ADAR2, while thiamine deficiency down-regulated it [249] [250]. Multiple ADAR isoforms with different editing efficiencies result from alternative splicing at the post-transcriptional level [251]. The ADAR2 pre-mRNA undergoes

editing itself that then influences the splicing process. Some regulations at the post-translational level include SUMO1 mediated sumoylation of ADAR1 lowers the activity of the enzyme and E3 ubiquitin ligase WWP2 mediated ubiquitination of ADAR2 enhances the degradation of the protein [252] [253]. Previous study has reported regulation of "degradation of both the ADAR1 and ADAR2 enzymes by AIMP2, an aminoacyl tRNA synthetase complex-interacting protein" [254]. Additionally, sub-cellular localization of the ADAR enzymes has been demonstrated in different studies as another important mechanism to regulate RNA editing activity [253] [255] [256].

"An extensive profiling of A-to-I RNA editing in thousands of human samples from the Genotype-Tissue Expression (GTEx) project and in hundreds of additional primate and mouse samples" was performed in a collaborative study in our lab recently [254]. "The study revealed a highly dynamic landscape of editing, including diverse spatial patterns of editing across tissues". With only two catalytically active adenosine deaminases performing the editing, such diverse spatial pattern might be a result of other non-ADAR regulators. High editing in brain than in non-brain tissues for many protein-coding sites was observed in all mammalian species from the study. Relating this to ADAR expression levels, it was found that in mouse the high editing correlates with high ADAR1 and/or ADAR2 expression levels but in humans this correlation was missing. Thus, in this study alternative regulatory factors to explain this pattern of high editing observed in human brain in comparison to non-brain tissues were probed for.

This study helped unravel the splicing factor SRSF9 as a key factor that regulates the editing of a number of protein-coding and non-coding sites in human brain. The expression of SRSF9 is found to be generally higher in non-brain tissues than in brain. Subsequently, through a combination of experimental and computational approaches this study demonstrates that "SRSF9 repressed editing by potentially

interfering with the dimerization of ADAR2 in the nucleus" and also showed "the direct editing regulon of SRSF9" that significantly enriched of editing for brain-specific sites. "A novel sequence motif in the ADAR2-dependent SRSF9 binding sites that may help account for some of the brain-specific editing observed in humans" was uncovered. These results highlight the role of regulatory factors, potentially splicing factors, in shaping the editing landscape in human and other primates as well.

6.2 Experiments and biological inferences

Experiments performed "on multiple tissues from a single human individual (N37) showed that the overall editing profiles from across different tissues" were similar, but many sites in the protein-coding regions exhibited a brain-specific editing. Majority of these sites were present in genes that had high expression even in non-brain tissues. Similar results were obtained when editing levels were examined in GTEx tissue samples. Collectively, these results indicate that expression level of the "genes could not adequately account for the brain-specific editing in these protein-coding sites". Examining the ADAR expression levels in N37 human individual demonstrated that the expression of both ADAR1 and ADAR2 was higher in lung than in cerebellum or frontal lobe and ADAR3 was enriched in brain. Also, "in the GTEx samples, ADAR1 and ADAR2 were expressed throughout the body and ADAR3 was enriched in brain". These patterns fail to explain the higher brain-specific editing observed and so in order to investigate the same the expression levels of non-ADAR regulators of editing was examined. Splicing factor, SRSF9, was found to be expressed in lower amounts in brain in comparison to non-brain tissues in both N37 individual and GTEx samples. Previous reports have shown SRSF9 can repress the editing in two substrates , CYFIP2 and CFLAR [257]. These results along with the expression profile obtained in the

present study enabled choosing SRSF9 as a potential candidate, regulating brain-specific editing, to investigate. Further, examining the expression levels of ADAR enzymes and SRSF9 in multiple cynomolgus monkey (*Macaca fascicularis*) tissues indicated no enrichment of ADAR1 and ADAR2 in brain tissues, while the "expression of SRSF9 was reduced in brain than in non-brain tissues". These results indicate that SRSF9 could potentially "enable brain-specific editing of selected sites in humans and other primates".

In order to investigate the hypothesis, gene perturbation experiments were performed in HEK293T cells and human neural progenitor cell line ReN-VM. In both the cells, the results highlighted SRSF9's ability "to selectively repress the editing of particular protein-coding sites in human cells". Further, to gain insights into the mechanism of repression, interactions between SRSF9 and ADAR2 were examined by co-immunoprecipitation experiments using different constructs of the two proteins with and without some regions. The results from these experiments indicated that dsRBD1 of ADAR2 interacts with SRSF9 in the nucleus and the "RRM2 domain of SRSF9 alone is sufficient for interaction with ADAR2". Co-immunoprecipitation experiments in the presence of RNA helped identify that the interaction between ADAR2 and SRSF9 is mediated by the RNA substrate and also indicated that SRSF9 could potentially repress ADAR2-mediated editing by preventing dimerization of ADAR2. Further, experiments on minigene constructs also helped identify that SRSF9 mediated regulation was a long distance one, perhaps through folding of RNA in three-dimensional space and "that SRSF9 relies on different cis-acting elements to perform its functions in RNA splicing and RNA editing".

The detailed description of the experiments, their procedures and results can be found in [258].

6.3 Pipeline developed

6.3.1 Pipeline for analyzing bulk RNA sequencing

For gaining insights into the variation observed at the gene expression and editing patterns of the different normal and perturbed cell lines they are subjected to illumina bulk RNA sequencing and a computational pipeline is developed to analyze the data from the sequencing. As a first step, the illumina sequenced reads are mapped against human genome assembly hg19 using STAR with mismatch penalty parameter, *outFilterMismatchNmax*, set to 15 [9] [10]. Following this two main kinds of analysis are designed for this study:

- Transcriptome analysis: In this analysis, a pipeline is developed to perform comparison of the gene expression profiles and understand their functional profiles.
- Editing analysis: In this analysis, a pipeline is developed to compute the RNA editing rates (conversion of A to I) observed in different samples for known editing sites deposited in RADAR.

Transcriptome analysis

The *summarizeOverlaps* function is used to obtain the raw counts from the aligned reads. Further, DESeq2 package is used to calculate the normalized gene expression counts and perform the differential gene expression analysis to compare samples from different conditions [235]. An enrichment analysis is performed on the differentially expressed genes to get insights about their functional profile using DAVID [236].

Editing analysis

The reads with mapping quality ≥ 20 corresponding to uniquely mapped reads are filtered. These reads are then reordered, sorted, assigned group information and further filtered for duplicates using Picard [160]. GATK toolkit is used to perform local realignment and recalibration [12] [13] [14]. Samtools mpileup is used for variant calling against a position list obtained from RADAR database that consists of a list of known editing sites [259] [260] [261]. Using this for each of the known editing sites, it is possible to calculate the coverage (number of times the site is covered in the different sequenced reads), number of times a G is observed at the editing site (that signifies A converted to I), number of times the editing site remains A, etc. Then, a minimum coverage threshold of 20 reads is imposed for the editing calculations. And, the editing rates at each of the sites are computed by calculating number of times G is observed by coverage (which signifies the total of A and G occurrences at the location).

The pipeline developed is illustrated in figure 6.1.

6.3.2 eCLIP data analysis pipeline

Apart from the bulk RNA sequencing performed on normal and perturbed cell lines, enhanced crosslinking and immunoprecipitation (eCLIP) experiments are performed. These experiments help to gain insights into exact binding sites of RNA binding proteins on their target RNAs and in this case it is performed to identify the exact binding sites of SRSF9 on their target RNAs. The format of the sequencing output after the eCLIP experiments is similar to that after bulk RNA sequencing performed earlier. Below, are the details of the steps employed to computationally process this sequencing data to identify the binding sites of ADAR on its target RNAs.

The reads are first trimmed to remove the adapter sequences. Subsequently, the

Figure 6.1: The pipeline developed to analyze the deep sequencing data for transcriptome and editing analysis.

reads are mapped using STAR to hg19 assembly with the maximum number of mismatches per read (*outFilterMismatchNmax*) at 15, the maximum number of mismatches per pair relative to read length (*outFilterMismatchNoverLmax*) at 0.4, the minimum number of bases that are exactly matched over the read length (*outFilterMatchNminOverLread*) at 0.3 meaning 30%, and the minimum normalized score for the alignment (*outFilterScoreMinOverLread*) at 0.3 as well [9] [10]. As with our editing pipeline, the next step is to sort for uniquely mapped reads. Further, the next few steps are performed by using the template eCLIP pipeline published by Yeo Lab in ENCODE [262]. The uniquely mapped reads are processed to remove PCR duplicates that accounts for not just exact duplicates but also slight differences due to sequencing errors in the barcodes. This file is then

sorted, indexed and then filtered to extract only the second read of each pair for processing with peak caller. Clipper program is used for peak calling. The "make_bigwig_files.py" and "bedToBigBed" scripts are used for generating visualization compatible files on the genome browser or other visualization softwares. Further, the experimental variation helping to identify relevant eCLIP peaks and how they are used in association with the RNA editing sites for finding significant sites where SRSF9 binding influences ADAR-mediated editing is explained in detail in section 6.4.2

6.4 Results and biological inferences

6.4.1 SRSF9 down-regulating ADAR2-mediated RNA editing at RADAR sites

In order to obtain a more comprehensive view of the ADAR2-mediated editing regulated by SRSF9, a computational pipeline was designed to analyze RNA-Seq data from experimental samples containing overexpression of ADAR2 or both ADAR2 and SRSF9 as shown in figure 6.2 (a). The computational pipeline enabled quantification of editing "on the extensive list of editing sites in RADAR database" [261]. "A filter of a minimum coverage of 20 reads" was imposed for measuring the editing levels with confidence. A comparison of the control vs. cells with overexpression of ADAR2 identified 18,174 sites that were differentially edited and 97% of these showed a significant increase in editing as is expected (p-value <0.05 , Fisher's test). This is depicted in figure 6.2 (b). A comparison of the cells with ADAR2 overexpression vs. those with overexpression of "both ADAR2 and SRSF9 indicate that 6,994 sites were differentially edited" (p-value <0.05 , Fisher's test) and importantly 92.5% of them were down-regulated as shown in figure 6.2 (b). This is consistent with the experimental results that SRSF9 acts as

a repressor of editing. Also, it is observed that "most of the significant sites were in 3' UTR and introns indicating the regulatory role of SRSF9 extends beyond protein-coding regions" that were tested in the experiments.

Figure 6.2: (a) "A schematic of the experimental setup to study the regulation of editing by SRSF9 using RNA-Seq where purple rings represent plasmids encoding SRSF9, while green rings represent plasmids encoding ADAR2; and (b) A comparison of editing levels between left panel: control cells and cells with ADAR2 overexpressed (OE) and right panel: cells with ADAR2 overexpressed and cells with both ADAR2 and SRSF9 overexpressed" indicating that SRSF9 down regulates the editing at majority of sites examined significantly.

6.4.2 Editing sites directly regulated by SRSF9 identified using eCLIP-Seq

In attempts to identify the ADAR2-mediated editing sites directly regulated by SRSF9, eCLIP-seq (enhanced crosslinking followed by immunoprecipitation and deep sequencing) is used to map the RNA binding sites of SRSF9 in HEK293T

cells. A computational pipeline was developed to analyze the deep sequencing data and identify the binding sites as detailed in section 6.3.2. But, the identified binding sites can potentially have confounding factors as SRSF9 performs other functions including splicing and binding at a site does not necessarily signify regulation of editing at the site. Therefore, the experimental design was done such that the binding sites were identified without and with over-expression of ADAR2 as shown in figure 6.3 (a) and enhanced binding was probed for in the presence of ADAR2.

The eCLIP "experiments yielded a total of 80,514 peaks throughout the transcriptome" and 31,772 of them corresponded to enhanced binding of SRSF9 on over-expression of ADAR2. The binding to endogenous SON transcripts is used as an example to showcase the results from the experiment. In the absence of ADAR2, SRSF9 was not bound to the transcript but upon over-expression of ADAR2, SRSF9 was found to exhibit strong binding at multiple locations as shown in figure 6.3 (b). Further, the eCLIP results were processed along with the known RNA-editing sites from RADAR and our RNA-Seq data. This enabled identification of 2,646 SRSF9-bound genes containing "at least one editing site having a minimum coverage of 20 reads in the RNA-Seq data". A comparison of this gene list "with the list of genes whose editing was identified to be significantly inhibited by SRSF9" in the previous section helped identify 3,823 sites in 683 genes that form the direct editing regulon of SRSF9 (refer figure 6.3 (c) and (d)) using details from both binding data (eCLIP-seq) and functional data (RNA-Seq).

Further, from the analysis we noted that although SRSF9 was bound to 1,963 genes in the presence of ADAR2, the editing sites in them were not repressed by SRSF9. This could be either due to very low editing at these sites in spite of over-expression of ADAR2 or due to SRSF9 alone not being able to repress the editing at these sites. In order to rule out the first possibility, the maximum editing level

Figure 6.3: (a) "A schematic of the experimental setup to study the transcriptome-wide binding of SRSF9 using eCLIP-seq where purple rings represent plasmids encoding SRSF9, while green rings represent plasmids encoding ADAR2; (b) Transcript map of SON from eCLIP-seq comprising two biological replicates each for SRSF9 binding without and with ADAR2 over-expression" along with paired size-matched inputs showing SRSF9 binds tightly with over-expression of ADAR2 to SON transcript; (c) "Venn diagram showing the overlap of the eCLIP-seq and RNA-Seq results" indicating 3823 as the number of sites whose editing is directly regulated by SRSF9; and (d) "Venn diagram showing the overlap of the eCLIP-seq and RNA-Seq results" indicating 683 as the number of genes constituting the direct editing regulon of SRSF9.

in each gene was examined (refer figure 6.4) and results indicate that though a majority of "genes were lowly edited still 860 genes were edited at a frequency of at least 5%". Thereby, indicating that SRSF9 alone was possibly not sufficient to repress editing at the editing sites in these 1,963 genes. This might be due to the existence of other unknown regulatory elements that in combination act at these sites to finally repress editing. Yet, "some of these ADAR2-dependent, SRSF9-bound genes could potentially be non-biological targets as well and a result of over-expression of SRSF9".

Figure 6.4: A plot of distribution of maximum editing at genes where SRSF9 binding does not impact editing significantly showing that editing level is greater than 5% in 860 genes.

The eCLIP-seq data enabled monitoring the distance of SRSF9 mediated regulation and thereby, investigate if the distance can affect the ability to inhibit editing. The "distance of each editing site to the closest SRSF9 binding peak" was computed for all sites (including 3,823 significant sites and 50,326 not affected sites). In order to account for the variable transcript length, the distances were normalized to gene length and the distribution of distance plotted separately for the sites where significant difference in editing was observed and those for which no impact on editing was observed (refer figure 6.5). The plots illustrate that the distributions look similar and SRSF9 binding was not closer to sites where it had

an negative impact on editing (p-value >0.1 , Kolmogorov-Smirnov test). Thus, these results from the genome-wise analysis are consistent with the experimental results obtained indicating that SRSF9 can inhibit editing at a distance [258].

Figure 6.5: A plot illustrating the distance distribution between the editing site and closest SRSF9 binding site, top panel: for sites where SRSF9 binding influences editing and bottom panel: for sites where SRSF9 binding has no impact on editing. The distributions are similar indicating that distance does not influence the impact on editing.

6.4.3 Down-regulation of ADAR2-mediated editing by SRSF9 in multiple human tissues

"GTEx project data on 53 distinct body sites (13 brain and 40 non-brain tissues) from 552 donors" was examined for their editing profile to gain further insights into regulation by SRSF9. Out of the 3,823 sites identified in our initial analysis as direct editing regulon of SRSF9, 3,173 sites were present in this data after imposing a filter that each site must contain "frequency measurements for at least one brain tissue and one non-brain tissue". 1,328 out of the 3,173 sites (41.9%)

showed "significantly higher editing levels in brain than in non-brain tissues (p-value < 0.05 , Wilcoxon test), which represented a significant enrichment for brain-specific editing" in comparison to different negative controls (p-value $< 1 * 10^{-22}$, Fisher's test) as shown in figure 6.6 (a). The negative controls used here include "(i) all editing sites in the 683 significant genes whose editing was not affected by SRSF9 binding", (ii) all sites with a coverage of at least 20 in the RNA-Seq data whose editing was not influenced by SRSF9, and (iii) all non-significant sites from the RNA-Seq data having a minimum coverage of 20 and a minimum editing frequency of 5% in at least one GTEx tissue.

Additionally, a sequence logo program, WebLogo, is used to identify the sequence conservation of nucleotides (motif) close to the ADAR2-mediated SRSF9 binding sites and ADAR2-independent eCLIP-seq peaks by extracting the sequences around the peaks. Two sequence motifs were identified for ADAR2-mediated SRSF9 binding sites, while for the ADAR2-independent eCLIP-seq peaks GA-rich motifs were identified as depicted in figure 6.6 (b) and (c). The latter "is consistent with previous results on SRSF9 and other SR proteins" [263] [264].

6.4.4 Physiological consequence of SRSF9 mediated editing repression

In an attempt to understand the reasoning behind repression of ADAR2-mediated editing by SRSF9 in non-brain tissues, experiments were performed in human neuroblastoma cell line SH-SY5Y including wildtype and genetically perturbed cells involving either over-expression of ADAR2 or knock-down of SRSF9 or both simultaneously. Analysis was performed on the RNA-Seq of these cells to gain insights into the molecular changes associated with the perturbations. Principal component analysis (PCA) on the gene expression levels clearly segregated the samples based on experimental condition as shown in figure 6.7 (a). Genes with

Figure 6.6: (a) Heatmap of the editing profile of GTEx tissues "depicting brain-specific sites regulated by SRSF9; enriched sequence motif present in (b) ADAR2-dependent SRSF9 binding peaks"; and (c) ADAR2-independent eCLIP-seq peaks.

very low expression count (less than 10) and those with low variability across samples (coefficient of variation less than 0.1) were discarded. Hierarchical clustering based on gene expression as depicted in figure 6.7 (b) helped elucidate the distinct expression patterns between the four experimental conditions. Further, differentially expressed gene sets between double perturbed (ADAR2 over-expressed and SRSF9 knock-down) cells and only single perturbed cells were probed. GO analysis on the up-regulated genes indicate enrichment "in RNA splicing, oxidative phosphorylation, ribosome and translation-related processes, proteasome, steroid biosynthesis and RNA polymerase" as shown in figure 6.7 (c). While, GO analysis on the down-regulated genes indicate enrichment in stress response and related functions as shown in figure 6.7 (d). This is consistent with the experimental results that indicate ADAR2 over-expressed, "SRSF9 depleted cells were more susceptible to apoptosis as their cellular defense mechanisms have been switched off or toned down".

Next, the editing levels were calculated using the RNA-Seq data and compared across the experimental conditions as depicted in figure 6.8 (a). A minimum coverage filter was imposed as 20 reads and editing rates were calculated for sites recorded in RADAR database [261]. Upon over-expression of ADAR2, "a total of 25,508 sites were differentially edited in comparison to control cells, of which 97.4% showed a significant increase in their editing levels" (p-value <0.05, Fisher's test). While, knock-down of SRSF9, resulted in 3,534 differentially edited sites with 86.7% of these showing significant increase in editing (p-value <0.05, Fisher's test) consistent with the role of SRSF9 as a repressor of editing. Further, simultaneously over-expression of ADAR2 and knock-down of SRSF9 resulted in "14,640 sites to be differentially edited relative to control cells" and 95.5% of them exhibited significantly increased editing levels (p-value <0.05, Fisher's test). Although, "the editing activity in these double perturbed cells were clearly stronger than

Figure 6.7: (a) PCA based on gene expression levels showing separation of samples based on experimental conditions; (b) Hierarchical clustering of the normalized gene expression counts depicting variation in samples and clustering based on experimental conditions; (c) GO analysis of genes significantly up-regulated in ADAR2 over-expression and SRSF9 depleted cells with enrichment score greater than 1; and (d) GO analysis of genes significantly down-regulated in ADAR2 over-expression and SRSF9 depleted cells with enrichment score greater than 1.

that in control cells and also in SRSF9 depleted cells, it appeared to be weaker than the editing activity in ADAR2 over-expressed cells". Later, on checking the transcript and protein level experimentally, no abnormality was found in transcript levels but the protein levels seemed to be similar to the control cells or the SRSF9 knock-down cells. In line with these experimental findings and "given that the gene expression analysis had earlier uncovered many genes related to the proteasome and ubiquitin-linked proteolysis that were significantly up-regulated upon simultaneous perturbations of ADAR2 and SRSF9", we propose a massive remodeling of the cellular proteome.

Additionally, on examining the significantly edited sites and comparing them across different experimental conditions, it was found that 2,762 sites showed significantly increased when ADAR2 was over-expressed along with SRSF9 knock-down but not in either of the single gene perturbation condition (p-value <0.05 , Fisher's test) as depicted in the Venn diagram in figure 6.8 (b). GO analysis of the genes containing these editing sites uncovered functions "related to the cell cycle and mitosis as well as terms related to cellular structures that undergo dynamic changes in cell division, such as the cytoskeleton and the chromosomes" (refer figure 6.8 (c)). These suggest "that RNA editing may be involved in the dramatic reduction in cell growth. Genes with functions in DNA damage and DNA repair were also enriched, therefore, suggesting a link between editing and the cells' susceptibility to apoptosis when stressed". Further, functions related to the ribosome, mitochondrion and spliceosome were also enriched that were earlier identified to be affected when performing GO analysis on the differentially expressed genes. Collectively, these "results suggest that SRSF9 prevents loss of cell viability by blocking ADAR2-mediated editing of key sites in genes that perform important functions in protein and energy metabolism as well as the cell cycle".

Figure 6.8: "(a) Comparison of editing levels between control cells and cells with ADAR2 over-expressed (left panel), between control cells and cells with SRSF9 knock-down (middle panel), and between control cells and cells with both ADAR2 over-expressed and SRSF9 knock-down (right panel); (b) Comparison of the sites with significant increase in editing in each of the three different genetic perturbation conditions depicted using a Venn diagram; and (c) GO analysis (top 10 functional terms listed) of the 2762 editing sites that were significantly up-regulated in the ADAR2 over-expressed and SRSF9 depleted cells but not in the single perturbed cells".

6.5 Discussion

"ADAR family of enzymes catalyzes the A-to-I RNA editing that is a fundamental post-transcriptional modification in animals". These modifications have to be carefully regulated as they are "critical in many cellular processes". In an "effort to build a comprehensive reference atlas" of the editing, a recent work in our lab helped "uncovered diverse spatiotemporal patterns of RNA editing in human" [254]. An important question that remains unanswered is how these patterns are generated given that only two catalytically active enzymes exist in the ADAR family. In particular, it is observed that many editing sites are "more highly edited in the brain than in non-brain tissues". This observation is readily accounted for in the mouse by the ADAR expression, "however, the same is not true in human because ADAR1 and ADAR2 are highly expressed in several non-brain tissues, such as the lung".

In this study, the splicing factor SRSF9 is found to "selectively represses the editing of numerous sites in non-brain tissues, thereby allowing them to be more highly edited in the brain". Experimentally it is "found that the RRM2 domain of SRSF9 alone is sufficient to biochemically interact with ADAR2 in the nucleus" and repress editing. Consistent with previous work on a related protein SRSF1 that reports the use of "only its pseudo-RRM domain (RRM2) to recognize its natural substrates in the cell", the RRM1 domain of SRSF9 is found to be dispensable for interaction and editing repression [263]. Further, our experiments identified RNA substrate to play an important role in mediating the regulatory role of SRSF9. But, this is contradictory to previous report that "RNase treatment or mutations in an RNA-binding α -helix in RRM2 domain abolishes interaction between SRSF9 and ADAR2", thereby preventing SRSF9 from inhibiting editing [263]. This variation could be due to difference in experimental setup that includes concentration of RNase use or duration of the treatment. Further, the data from both our ex-

periments and computational analysis indicate SRSF9 can regulate editing from a distance (refer figure 6.5), which again emphasizes on a more prominent role of RNA in the process resulting from its three-dimensional structure and folding. Additionally, experimental data indicates that SRSF9 represses editing by potentially "interfering with the dimerization of ADAR2", yet the details of mechanism is unclear and more work is required for gaining insights into it.

SRSF9-mediated repression of editing is target-specific and in an attempt to understand how this choice is made, the genome-wise computational analysis performed in this study helped identify two sequence motifs enriched for the sites regulated by SRSF9 (refer figure 6.6). Despite this, there are many genes regulated by SRSF9 that do not contain this sequence motif or a GA-rich motif (found for SR proteins) and this could possibly mean SRSF9 repression also depends on other motifs (structural or higher order) to perform its regulatory functions.

Experimental and computational analysis on SH-SY5Y cell lines (wildtype, single perturbed and double perturbed) enable understanding the reason for repression of editing of brain-specific sites in non-brain tissues. Experiments showed that the double perturbed cells had reduced growth rate and cell viability. The transcriptomic analysis based on RNA-Seq of these cells reveal that GO analysis on the differentially expressed genes and edited genes are consistent and relate to protein homeostasis and energy metabolism. Further, the expression of genes involved in pathways necessary for coping with stress were down-regulated in the double perturbed cells and the "differentially edited genes are enriched in functions relating to cell cycle as well as DNA damage and repair". Thus, confirming that "although the poor growth and cell viability of the double perturbed cells (ADAR2 over-expressed and SRSF9 depleted cells) may simply be an outcome compounded by two separate stressors", an important role is likely played by RNA editing as well "due to its association with multiple dysregulated biological processes". Nev-

ertheless, more work is needed to understand the "underlying mechanism linking editing with the observed phenotypes".

Also, in the current work, the focus has mainly been "on addressing brain-specific editing, but there are many other spatiotemporal editing patterns" including selective editing of many sites in only a single tissue, "an editing site's dependence on ADAR1 or ADAR2 depending on the biological context", etc [254]. And clearly, other trans-acting factors must be involved in regulating the activity of the two deaminases, thereby, resulting in the complex spatiotemporal editing patterns.

With the current work elucidating the important role of SRSF9 in shaping the editing landscape in mammals and a previous study reported "various splicing factors, including SR proteins, associated with ADAR2 in large nuclear ribonucleoprotein (lnRNP) particles, it is tempting to speculate that other spliceosome components may also regulate editing" [265]. More work and carefully designed studies with a combination of experiments and computational data can help unravel the other factors influencing the editing patterns.

6.6 Conclusion

A combination of experiments and computational analysis helped showcase SRSF9 as a repressor of editing of brain-specific sites in non-brain tissues, thereby, enable shaping the editing landscape. Experiments indicate RRM2 domain of SRSF9 mediates interaction with ADAR2 and RNA plays a prominent role in their interaction. The SRSF9 mediated repression of editing is potentially by interfering with ADAR2 dimerization. Further, the computational editing analysis on the list of known editing sites from RADAR helped identify the SRSF9 editing regulon. Analysis on eCLIP-seq data along with editing analysis on the RNA-Seq showcased the long range influence of SRSF9 regulation, thereby emphasizing the role of RNA in the interaction between SRSF9 and ADAR2. Two sequence motifs

significantly enriched in ADAR2-mediated SRSF9 binding sites were identified. Finally, RNA-Seq analysis on gene perturbation cells enabled identification of loss of cell viability by preventing ADAR2-mediated editing as the reason for SRSF9 mediated repression of editing in non-brain tissues.

Chapter 7

Conclusion: summary, and future directions

Editing at both DNA and RNA level is gaining immense interest as studies continue to reveal their huge biological significance and impact. This work describes my attempts at gaining insights into editing at both the DNA and RNA level, more specifically in human cells, using different computational techniques. Also, this work illustrates the computational techniques devised to gain these biological insights that can potentially be employed for future studies.

7.1 Main contributions and significance of work

The main contributions from this work can be highlighted based on two key contributing aspects of this work that include computational pipelines developed and the biological insights gained by employing these pipelines on biologically relevant data.

7.1.1 Computational contribution

With next generation sequencing becoming a more integral part of biological experiments for gaining precise insights, the requirement to develop pipelines that can enable analyzing and extracting the required details from the huge amount of data generated from these sequencing technologies has increased manifolds.

As a part of this work [involving chapters 3, 4, 5 and 6], the pipelines developed showcase enormous applicability in the present and for future studies. The reason being absence of standardized pipelines for these applications in spite of next generation sequencing becoming more affordable and commonly used in research. Although, individual tools for some of the preliminary steps [like cleaning of data, mapping reads to reference, gene expression quantification, realignment and recalibration, etc] in the data analysis exist separately, integrated pipelines are absent. Moreso, as downstream processing pipelines vary depending on what one is looking at it is more challenging to unify as one integrated pipeline.

In the present work, involving DNA and RNA editing, two key downstream applications are involved [identifying and quantifying editing at DNA and RNA level] along with another commonly used protocol involving gene expression quantification to compare conditions. An integrated pipeline of these applications developed in the present work is depicted in figure 7.1 with attributes that can characterize a read's alignment to reference and use this for further processing including pattern recognition; and using variant calling based methods for computing RNA editing rates at known editing sites. The figure 7.1 shows an incremental addition of features for various applications starting from the preprocessing of the sequenced reads by removal of the primers/ adaptors to generate a file containing just the read sequences [in fastq format]. Followed by mapping of these sequenced reads to the corresponding reference [in bam format].

- One possible and common direction from here involves obtaining gene ex-

Figure 7.1: An integrated schematic of the various computational pipelines developed in this thesis to analyze the NGS data. The color coded bar on the side of each step represents the chapter these steps were employed in.

pression from these mapped data files and this has application in a wide range of studies when one is interested in looking at variation between different conditions tested. This can further be processed to give insights into functional enrichment when comparing conditions.

- Another possible direction involves further processing of these mapped files including removal of non-uniquely mapped reads and local realignment, recalibration that readies the file in bam format for further downstream processing.
 - The reads are processed into categories based on how they align to their reference [matched alignment/ alignment with insertion/ alignment with deletion].
 - * Processing for NHEJ/ HDR rates
 - * Processing for pattern recognition on the indels generated
 - The reads matched with the reference are processed to identify base mapped to known RNA editing site coordinates by using variant calling methods.
 - * Processing for editing rates

Further, in future, the pipelines devised in this work can be used in addition to experimental functional assays when probing for alterations in DNA, post application of CRISPR-Cas like systems for DNA editing. As the pipeline developed here is sensitive enough to detect as small as 1bp insertion/ deletion it will enable getting a more precise picture of the editing outcome in and around the location of intended editing. Further, the pipeline's ability to compute the RNA editing rates for known sites is not only useful in comparing different experimental conditions [with knockouts and over expression] and identifying regulators [as illustrated in the current work in chapter 6], but with increasing awareness about the association of RNA editing in disease and its progression, it serves a promising starting

contribution to probe for such associations between the two.

7.1.2 Biological insights

The main biological insights gained from this work are summarized as below:

- Identification of the likely role of highly conserved residues in the active site of the HNH nuclease domain of Cas9 in maintaining the structural stability of the active site cleft and thereby, assisting catalysis.
- Evaluating the performance of five CRISPR-Cas enzymes in NHEJ-mediated and HDR-mediated genome editing to develop a set of guidelines that can be used for designing future genome engineering studies.
- Characterizing the mutations introduced by five CRISPR-Cas enzymes in NHEJ-mediated genome editing to identify prominent patterns that can potentially guide better and safe usage of the CRISPR-Cas based editing technology.
- Transcriptomic analysis on glutamine synthase knockout cell lines generated using CRISPR-Cas system revealing them to be ideal for use as a expression system for stable production of recombinant erythropoietin.
- Demonstrating splicing factor, SRSF9, as a repressor of ADAR2-mediated RNA editing of brain-specific sites in non-brain tissues in primates and possibly throwing light into its role in shaping the intricate editing patterns observed.

These insights in totally have added to our knowledge gain with respect to 3 key domains that include functioning/ activity of the enzymes involved, application of technology, and identification of regulatory factors in functioning of an enzymes in either of the two systems that were studied in this work including the DNA and RNA editing systems. And, in totality these insights add value as they can guide in designing future studies with these systems.

7.2 Summary, impact and future directions

This section gives a summary of the problems addressed in this thesis, highlighting the techniques used for performing the study, the results from them and impact along with possible future directions of the study.

7.2.1 Uncovering the role of highly conserved residues in the HNH nuclease domain of Cas9 using molecular dynamics simulations

Chapter 2 details the efforts in designing and performing a structure based study using molecular modeling and simulations techniques to gain insights into the functioning of Cas9 enzyme by probing one of its nuclease domains (HNH). This work enables identification of the likely role of the highly conserved residues in the active site of the HNH nuclease domain and interaction network between conserved residues, ion and catalytic residues that is critical in maintaining the structural motif of the active site cleft and thereby, assists catalysis.

One of the limitations of this work is that the study is performed on a model of the HNH nuclease domain in isolation and a single-stranded DNA molecule in the active site. The truncation is performed as a trade-off in order to enable better conformational space sampling as previous report on full protein fail to sample the active state. And, the absence of a proper template to build a model of the DNA-RNA heteroduplex in the active site cleft led to building a model with a single-stranded DNA molecule in the active site as a representation to track any effects on the nucleic acid. However, the identification of interaction network does illustrate the strength of using similar techniques that can enable gaining immense structural and functional knowledge about a complex system like the Cas9 protein in association with DNA and RNA when experimental structure

determination techniques fall short due to limitations and complexity. I believe using more intensive conformational sampling techniques that can help sampling of active state conformations from inactive state can add more value and increase our understanding of the system as a whole and its conformational changes during functioning. Also, in the present study simulations were performed using Mg^{2+} and Zn^{2+} as ions in the active site, more detailed studies designed with different divalent cation can shed light on the mechanism behind ion selectivity displayed by the nuclease domain. Understanding the structural and functional aspect of this huge complex can enable easier and better modification of the protein/ complex to receive desired outcome in its application and associated editing technology.

7.2.2 Developing computational pipeline for a systematic evaluation of CRISPR-Cas systems in human cells

Chapter 3 describes the efforts to evaluate the performance of five CRISPR-Cas enzymes in NHEJ-mediated and HDR-mediated genome editing pathways by using a combination of experiments and computational techniques. Here, I developed two computational pipelines to analyze data from NGS relating to functioning of five CRISPR-Cas enzymes in an attempt to quantify their performances in the two genome editing pathways.

The results from this study helps identify some factors affecting the functioning of each CRISPR-Cas system and highlights the importance of designing a comprehensive study on common platform to compare the performance of enzyme performing similar functions to be able to understand the factors influencing their activity and to leverage the best performing enzyme in a set up for different applications. Thus, the findings from this study can serve as a guide in choosing the appropriate system for utility depending on the application rather than using default systems and also enables understanding the limitations associated with

each system.

7.2.3 Characterizing non-random mutations introduced by CRISPR-Cas systems in NHEJ-mediated genome editing

Chapter 4 attempts to characterize the mutations introduced by five CRISPR-Cas enzymes in NHEJ-mediated genome editing based on the experiments performed in chapter 3. A computational program developed by me is used to identify the indel patterns generated and determine the most prominent indel classes introduced by the enzymes by performing a hypergeometric test for identifying overrepresentation. The preliminary results indicate non-randomness in the mutations introduced as is already reported in literature.

The work in this chapter can be further enhanced by including NGS data available from other published studies and subjected them to our pipeline to see any effect on the patterns determined. Also, using the patterns identified machine learning techniques can be used to develop a predictive tool for reporting the possible indel outcomes for each of the enzymes. Although such tools exist for SpCas9 at the moment, this study supplemented with additional data for other enzymes that can help develop similar predictive tools for the other four enzymes as well. And, the outcome from this study and from the predictive tools can become an integral part of designing DNA editing experiments as they will allow estimating possible outcomes of the editing planned. Prior knowledge of the outcome can aid not only in effective experimental planning such that it does not impact the outcome but can also be leveraged for introduction of desired indel patterns as a result of the experiment.

7.2.4 Characterizing glutamine synthase knock-out based on HEK293 for use as a human expression system for stable production of recombinant erythropoietin

Chapter 5 details the efforts made to characterize a glutamine synthetase knockout cell line generated using CRISPR-Cas system for utility as a bioproduction platform. I developed a computational pipeline for investigating the transcriptomic expression profiles of the comparing the wildtype and knockout cell lines as well as those grown in adherant and suspension cell cultures. The results from this study reveal the glutamate synthase knockout cells as ideal candidates for developing an expression system for production of recombinant erythropoietin with fully human glycosylation.

This work enumerates how the CRISPR-Cas technology may be utilized for a wide range of applications like in this case for building a bioproduction platform or for other investigating purposes by knocking out a gene or introducing desired changes at a genomic location. The computational pipeline developed here depicts how with a combination of sequencing technology one can actually ascertain that required changes/ properties have been acquired by the modified system upon using CRISPR-Cas editing technology on it before proceeding with the application. On the whole, this study shows how the CRISPR-Cas technology can be utilized to perform genomic modification followed by a check to double confirm that required changes have only been inflicted and the modified system has the desired properties before utilizing the modified system for a desired application.

7.2.5 Genome-wide editing analysis identified SRSF9 as selective repressor of "ADAR2-mediated editing of brain-specific sites in primates"

Chapter 6 explains the study designed using both experiments and computational techniques to identify and demonstrate splicing factor, SRSF9, as a repressor of ADAR2-mediated editing of brain-specific sites in non-brain tissues in primates. A genome-wide RNA editing analysis and transcriptomic profile analysis is performed using computational pipelines developed in order to demonstrate splicing factor, SRSF9, as a repressor, identify its editing regulon, and the probable reason for the SRSF9 mediated repression.

The close association of abnormal RNA editing pattern with diseases emphasizes on the need to better understand the basics of the RNA editing phenomenon. Fewer enzymes associated with editing and a vast complex editing landscape generated increases the complexity of their functioning and, thereby, highlights the gaps in our knowledge at the moment. The present study which utilized a combination of experiments with computational techniques enabled the identification of one of the regulatory factors that play an important role in shaping the editing landscape and highlights how designing such studies can help gain information on the editing regulations that result in generation of the complex editing landscape with just a few enzymes. Not just in terms of the regulatory factors of the editing but very little is known in terms of the actual structural and functional factors associated with the enzymes and their substrates. One such aspect being, the experiments in the present study indicate SRSF9 mediated repression is by interfering with dimerization of ADAR2, the mechanism behind this is unclear. Utilizing computational modelling and simulation techniques can help building a model of the complex, inspecting them at a structural level and can likely give

insights into the mechanism behind the repression/ functioning.

7.3 Concluding Remarks

In summary, the work presented in this thesis details my efforts at employing different computational techniques including molecular modelling and simulation, and development of pipelines to analyze huge data from NGS to address queries on DNA and RNA editing. The pipelines developed are specialized in two main downstream analysis that lack standardized computational pipelines at the moment that may be employed on any data of interest. It includes identification and quantification of editing both at the level of DNA, at a desired genomic locus, and RNA, at previously identified RNA editing sites in the genome. And, employing these techniques have enabled probing for and highlighting critical attributes in functioning of the CRISPR-Cas DNA editing system that can add value in utilizing the CRISPR technology safely to its fullest potential. Also, these computational techniques also enabled uncovering a regulatory factor for ADAR2-mediated RNA editing that is responsible for shaping the vast intricate editing landscape whose perturbation has been implicated in various diseases.

The work presented here not only contributed to enhancing our knowledge of the functioning of the two biological systems involved but also elucidated how these systems can be employed to gain information and provide solution to problems. And, I am confident that this will have a positive and significant impact in driving future research queries on DNA and RNA editing.

Author's Publications

Published Papers:

1. "Wang Y, Liu KI, Sutrisnoh NB, **Srinivasan H**, Zhang J, Li J, Zhang F, Lalith CRJ, Xing H, Shanmugam R, Foo JN, Yeo HT, Ooi KH, Bleckwehl T, Par YYR, Lee SM, Ismail NNB, Sanwari NAB, Lee STV, Lew J, Tan MH. Systematic evaluation of CRISPR-Cas systems reveals design principles for genome editing in human cells. *Genome Biol.*, 19(1):62, May 2018".
2. "Shanmugam R, Zhang F, **Srinivasan H**, Charles Richard JL, Liu KI, Zhang X, Woo CWA, Chua ZHM, Buschdorf JP, Meaney MJ, Tan MH. SRSF9 selectively represses ADAR2-mediated editing of brain-specific sites in primates. *Nucleic Acids Res.*, 46(14):7379-7395, Aug 2018".

Submitted Papers:

1. Chin C, Goh JB, **Srinivasan H**, Liu KI, Gowher A, Shanmugam R, Lee LH, Choo M, Tan AH, Nguyen-Khuong T, Tan MH, Ng SK. A human expression system based on HEK293 for the stable production of recombinant erythropoietin. Submitted for review to Scientific Reports.

Manuscript in preparation:

- Srinivasan H, and Tan MH. DNA repair outcomes at Cas9- and Cas12a-mediated double-stranded breaks.

Bibliography

- [1] W. J. Browne, A. C. North, D. C. Phillips, K. Brew, T. C. Vanaman, and R. L. Hill. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.*, 42(1):65–86, May 1969.
- [2] A. Sali and T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, 234(3):779–815, Dec 1993.
- [3] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv. Protein Chem.*, 66:27–85, 2003.
- [4] Bernard R. Brooks, Robert E. Bruccoleri, Barry D. Olafson, David J. States, S. Swaminathan, and Martin Karplus. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4:187–217, 1983.
- [5] B. R. Brooks, C. L. Brooks, A. D. Mackerell, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. CHARMM: the biomolecular simulation program. *J Comput Chem*, 30(10):1545–1614, Jul 2009.
- [6] D. Van Der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. Berendsen. GROMACS: fast, flexible, and free. *J Comput Chem*, 26(16):1701–1718, Dec 2005.
- [7] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *The Journal of Chemical Physics*, 76:637–649, 1982.
- [8] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv:1303.3997*, 2013.

- [9] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.
- [10] Alexander Dobin and Thomas R Gingeras. Mapping RNA-seq reads with STAR. *Current protocols in bioinformatics*, 51(1):11–14, 2015.
- [11] Daehwan Kim, Geo Pertea, Cole Trapnell, Harold Pimentel, Ryan Kelley, and Steven L Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.
- [12] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 2010.
- [13] Mark A DePristo, Eric Banks, Ryan Poplin, Kiran V Garimella, Jared R Maguire, Christopher Hartl, Anthony A Philippakis, Guillermo Del Angel, Manuel A Rivas, Matt Hanna, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491, 2011.
- [14] Geraldine A Van der Auwera, Mauricio O Carneiro, Christopher Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, et al. From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Current protocols in bioinformatics*, 43(1):11–10, 2013.
- [15] Kira S Makarova, Yuri I Wolf, and Eugene V Koonin. Comparative genomics of defense systems in archaea and bacteria. *Nucleic acids research*, 41(8):4360–4377, 2013.
- [16] Rodolphe Barrangou. The roles of CRISPR–Cas systems in adaptive immunity and beyond. *Current opinion in immunology*, 32:36–41, 2015.
- [17] Rodolphe Barrangou and Philippe Horvath. CRISPR: new horizons in phage resistance and strain identification. *Annual review of food science and technology*, 3:143–162, 2012.
- [18] Blake Wiedenheft, Samuel H Sternberg, and Jennifer A Doudna. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331, 2012.

- [19] John Van der Oost, Matthijs M Jore, Edze R Westra, Magnus Lundgren, and Stan JJ Brouns. CRISPR-based adaptive and heritable immunity in prokaryotes. *Trends in biochemical sciences*, 34(8):401–407, 2009.
- [20] Kira S Makarova, Daniel H Haft, Rodolphe Barrangou, Stan JJ Brouns, Emmanuelle Charpentier, Philippe Horvath, Sylvain Moineau, Francisco JM Mojica, Yuri I Wolf, Alexander F Yakunin, et al. Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, 9(6):467, 2011.
- [21] Daniel H Haft, Jeremy Selengut, Emmanuel F Mongodin, and Karen E Nelson. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS computational biology*, 1(6):e60, 2005.
- [22] Kira S Makarova, Nick V Grishin, Svetlana A Shabalina, Yuri I Wolf, and Eugene V Koonin. 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. *Biology direct*, 1(1):7, 2006.
- [23] Yoshizumi Ishino, Hideo Shinagawa, Kozo Makino, Mitsuko Amemura, and Atsuo Nakata. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product. *Journal of bacteriology*, 169(12):5429–5433, 1987.
- [24] Francisco JM Mojica, Cesar Díez-Villaseñor, Elena Soria, and Guadalupe Juez. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. *Molecular microbiology*, 36(1):244–246, 2000.
- [25] Ruud Jansen, Jan DA van Embden, Wim Gaastra, and Leo M Schouls. Identification of genes that are associated with DNA repeats in prokaryotes. *Molecular microbiology*, 43(6):1565–1575, 2002.
- [26] Kira S Makarova, L Aravind, Nick V Grishin, Igor B Rogozin, and Eugene V Koonin. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. *Nucleic acids research*, 30(2):482–496, 2002.
- [27] Francisco JM Mojica, Jesús García-Martínez, Elena Soria, et al. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *Journal of molecular evolution*, 60(2):174–182, 2005.
- [28] Christine Pourcel, Grégory Salvignol, and Gilles Vergnaud. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bac-

- terioophage DNA, and provide additional tools for evolutionary studies. *Microbiology*, 151(3):653–663, 2005.
- [29] Alexander Bolotin, Benoit Quinquis, Alexei Sorokin, and S Dusko Ehrlich. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. *Microbiology*, 151(8):2551–2561, 2005.
- [30] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.
- [31] Rodolphe Barrangou, Christophe Fremaux, Hélène Deveau, Melissa Richards, Patrick Boyaval, Sylvain Moineau, Dennis A Romero, and Philippe Horvath. CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, 315(5819):1709–1712, 2007.
- [32] Josiane E Garneau, Marie-Ève Dupuis, Manuela Villion, Dennis A Romero, Rodolphe Barrangou, Patrick Boyaval, Christophe Fremaux, Philippe Horvath, Alfonso H Magadán, and Sylvain Moineau. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. *Nature*, 468(7320):67, 2010.
- [33] Ido Yosef, Moran G Goren, and Udi Qimron. Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic acids research*, 40(12):5569–5576, 2012.
- [34] Daan C Swarts, Cas Mosterd, Mark WJ Van Passel, and Stan JJ Brouns. CRISPR interference directs strand specific spacer acquisition. *PloS one*, 7(4):e35888, 2012.
- [35] Jason Carte, Ruiying Wang, Hong Li, Rebecca M Terns, and Michael P Terns. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. *Genes & development*, 22(24):3489–3496, 2008.
- [36] Stan JJ Brouns, Matthijs M Jore, Magnus Lundgren, Edze R Westra, Rik JH Slijkhuis, Ambrosius PL Snijders, Mark J Dickman, Kira S Makarova, Eugene V Koonin, and John Van Der Oost. Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, 321(5891):960–964, 2008.
- [37] Matthijs M Jore, Magnus Lundgren, Esther Van Duijn, Jelle B Bultema, Edze R Westra, Saktharam P Waghmare, Blake Wiedenheft, Ümit Pul, Reinhild Wurm, Rolf Wagner, et al. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. *Nature structural & molecular biology*, 18(5):529, 2011.

- [38] Kira S Makarova, Yuri I Wolf, Omer S Alkhnbashi, Fabrizio Costa, Shiraz A Shah, Sita J Saunders, Rodolphe Barrangou, Stan JJ Brouns, Emmanuelle Charpentier, Daniel H Haft, et al. An updated evolutionary classification of CRISPR–Cas systems. *Nature Reviews Microbiology*, 13(11):722, 2015.
- [39] Kira S Makarova, L Aravind, Yuri I Wolf, and Eugene V Koonin. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR–Cas systems. *Biology direct*, 6(1):38, 2011.
- [40] Kira S Makarova and Eugene V Koonin. Annotation and classification of CRISPR–Cas systems. In *CRISPR*, pages 47–75. Springer, 2015.
- [41] Eugene V Koonin, Kira S Makarova, and Feng Zhang. Diversity, classification and evolution of CRISPR–Cas systems. *Current opinion in microbiology*, 37:67–78, 2017.
- [42] Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR–Cas9. *Science*, 346(6213):1258096, 2014.
- [43] Heidi Ledford. CRISPR, the disruptor. *Nature News*, 522(7554):20, 2015.
- [44] Fuguo Jiang and Jennifer A Doudna. CRISPR–Cas9 structures and mechanisms. *Annual review of biophysics*, 46:505–529, 2017.
- [45] Emmanuelle Charpentier and Luciano A Marraffini. Harnessing CRISPR–Cas9 immunity for genetic engineering. *Current opinion in microbiology*, 19:114–119, 2014.
- [46] Patrick D Hsu, Eric S Lander, and Feng Zhang. Development and applications of CRISPR–Cas9 for genome engineering. *Cell*, 157(6):1262–1278, 2014.
- [47] Haifeng Wang, Marie La Russa, and Lei S Qi. CRISPR/Cas9 in genome editing and beyond. *Annual review of biochemistry*, 85:227–264, 2016.
- [48] Jolly Basak and Chandran Nithin. Targeting Non-Coding RNAs in Plants with the CRISPR–Cas Technology is a Challenge yet Worth Accepting. *Frontiers in Plant Science*, 2015.
- [49] Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, page 1225829, 2012.
- [50] F Ann Ran, Le Cong, Winston X Yan, David A Scott, Jonathan S Gootenberg, Andrea J Kriz, Bernd Zetsche, Ophir Shalem, Xuebing Wu, Kira S Makarova, et al. In vivo genome editing using *Staphylococcus aureus* Cas9. *Nature*, 520(7546):186, 2015.

- [51] Yan Zhang, Nadja Heidrich, Biju Joseph Ampattu, Carl W Gunderson, H Steven Seifert, Christoph Schoen, Jörg Vogel, and Erik J Sontheimer. Processing-independent CRISPR RNAs limit natural transformation in *Neisseria meningitidis*. *Molecular cell*, 50(4):488–503, 2013.
- [52] Ines Fonfara, Anaïs Le Rhun, Krzysztof Chylinski, Kira S Makarova, Anne-Laure Lecrivain, Janek Bzdrenga, Eugene V Koonin, and Emmanuelle Charpentier. Phylogeny of Cas9 determines functional exchangeability of dual-RNA and Cas9 among orthologous type II CRISPR-Cas systems. *Nucleic acids research*, 42(4):2577–2590, 2013.
- [53] Martin Jinek, Fuguo Jiang, David W Taylor, Samuel H Sternberg, Emine Kaya, Enbo Ma, Carolin Anders, Michael Hauer, Kaihong Zhou, Steven Lin, et al. Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. *Science*, 343(6176):1247997, 2014.
- [54] Rimantas Saprunauskas, Giedrius Gasiunas, Christophe Fremaux, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. The *Streptococcus thermophilus* CRISPR/Cas system provides immunity in *Escherichia coli*. *Nucleic acids research*, 39(21):9275–9282, 2011.
- [55] Bernd Zetsche, Jonathan S Gootenberg, Omar O Abudayyeh, Ian M Slaymaker, Kira S Makarova, Patrick Essletzbichler, Sara E Volz, Julia Joung, John Van Der Oost, Aviv Regev, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, 163(3):759–771, 2015.
- [56] Bernd Zetsche, Jonathan S Gootenberg, Omar O Abudayyeh, Ian M Slaymaker, Kira S Makarova, Patrick Essletzbichler, Sara E Volz, Julia Joung, John Van Der Oost, Aviv Regev, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell*, 163(3):759–771, 2015.
- [57] Eszter Tóth, Bernadett C Czene, Péter I Kulcsár, Sarah L Krausz, András Tálas, Antal Nyeste, Éva Varga, Krisztina Huszár, Nóra Weinhardt, Zoltán Ligeti, et al. Mb- and FxCpf1 nucleases are active in mammalian cells: activities and PAM preferences of four wild-type Cpf1 nucleases and of their altered PAM specificity variants. *Nucleic acids research*, 46(19):10272–10285, 2018.
- [58] Alexandra East-Seletsky, Mitchell R O’Connell, David Burstein, Gavin J Knott, and Jennifer A Doudna. RNA targeting by functionally orthogonal type VI-A CRISPR-Cas enzymes. *Molecular cell*, 66(3):373–383, 2017.
- [59] Omar O Abudayyeh, Jonathan S Gootenberg, Silvana Konermann, Julia Joung, Ian M Slaymaker, David BT Cox, Sergey Shmakov, Kira S

- Makarova, Ekaterina Semenova, Leonid Minakhin, et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science*, 353(6299):aaf5573, 2016.
- [60] Alexandra East-Seletsky, Mitchell R O’Connell, Spencer C Knight, David Burstein, Jamie HD Cate, Robert Tjian, and Jennifer A Doudna. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. *Nature*, 538(7624):270, 2016.
- [61] Aaron A Smargon, David BT Cox, Neena K Pyzocha, Kaijie Zheng, Ian M Slaymaker, Jonathan S Gootenberg, Omar A Abudayyeh, Patrick Essletzbichler, Sergey Shmakov, Kira S Makarova, et al. Cas13b is a type VI-B CRISPR-associated RNA-guided RNase differentially regulated by accessory proteins Csx27 and Csx28. *Molecular cell*, 65(4):618–630, 2017.
- [62] Winston X Yan, Shaorong Chong, Huaibin Zhang, Kira S Makarova, Eugene V Koonin, David R Cheng, and David A Scott. Cas13d is a compact RNA-targeting type VI CRISPR effector positively modulated by a WYL-domain-containing accessory protein. *Molecular cell*, 70(2):327–339, 2018.
- [63] Xiao-Hui Zhang, Louis Y Tee, Xiao-Gang Wang, Qun-Shan Huang, and Shi-Hua Yang. Off-target effects in CRISPR/Cas9-mediated genome engineering. *Molecular Therapy-Nucleic Acids*, 4:e264, 2015.
- [64] Michelle L Kimberland, Wangfang Hou, Adolfo Alfonso-Pecchio, Stephen Wilson, Yanhua Rao, Shu Zhang, and Quinn Lu. Strategies for controlling off-target effects and biological variations in mammalian CRISPR/Cas9 genome editing experiments. *Journal of biotechnology*, 2018.
- [65] Yingbo Cui, Jiaming Xu, Minxia Cheng, Xiangke Liao, and Shaoliang Peng. Review of CRISPR/Cas9 sgRNA design tools. *Interdisciplinary Sciences: Computational Life Sciences*, 10(2):455–465, 2018.
- [66] Vanja Tadić, Goran Josipović, Vlatka Zoldoš, and Aleksandar Vojta. CRISPR/Cas9-based epigenome editing: an overview of dCas9-based tools with special emphasis on off-target activity. *Methods*, 2019.
- [67] Benjamin P Kleinstiver, Michelle S Prew, Shengdar Q Tsai, Ved V Topkar, Nhu T Nguyen, Zongli Zheng, Andrew PW Gonzales, Zhuyun Li, Randall T Peterson, Jing-Ruey Joanna Yeh, et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature*, 523(7561):481, 2015.
- [68] Benjamin P Kleinstiver, Michelle S Prew, Shengdar Q Tsai, Nhu T Nguyen, Ved V Topkar, Zongli Zheng, and J Keith Joung. Broadening the targeting range of *Staphylococcus aureus* CRISPR-Cas9 by modifying PAM recogni-

- tion. *Nature biotechnology*, 33(12):1293, 2015.
- [69] Ryan T Leenay, Kenneth R Maksimchuk, Rebecca A Slotkowski, Roma N Agrawal, Ahmed A Gomaa, Alexandra E Briner, Rodolphe Barrangou, and Chase L Beisel. Identifying and visualizing functional PAM diversity across CRISPR-Cas systems. *Molecular cell*, 62(1):137–147, 2016.
- [70] Carolin Anders, Katja Bargsten, and Martin Jinek. Structural plasticity of PAM recognition by engineered variants of the RNA-guided endonuclease Cas9. *Molecular cell*, 61(6):895–902, 2016.
- [71] Hiroshi Nishimasu, Takashi Yamano, Linyi Gao, Feng Zhang, Ryuichiro Ishitani, and Osamu Nureki. Structural basis for the altered PAM recognition by engineered CRISPR-Cpf1. *Molecular cell*, 67(1):139–147, 2017.
- [72] Johnny H Hu, Shannon M Miller, Maarten H Geurts, Weixin Tang, Liwei Chen, Ning Sun, Christina M Zeina, Xue Gao, Holly A Rees, Zhi Lin, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, 556(7699):57, 2018.
- [73] Chang Liu, Li Zhang, Hao Liu, and Kun Cheng. Delivery strategies of the CRISPR-Cas9 gene-editing system for therapeutic applications. *Journal of controlled release*, 266:17–26, 2017.
- [74] Rubul Mout, Moumita Ray, Yi-Wei Lee, Federica Scaletti, and Vincent M Rotello. In vivo delivery of CRISPR/Cas9 for therapeutic gene editing: progress and challenges. *Bioconjugate chemistry*, 28(4):880–884, 2017.
- [75] Christopher A Lino, Jason C Harper, James P Carney, and Jerilyn A Timlin. Delivering CRISPR: a review of the challenges and approaches. *Drug Delivery*, 25(1):1234–1257, 2018.
- [76] Jonatha M. Gott and Ronald B. Emeson. Functions and mechanisms of RNA editing. *Annual Review of Genetics*, 34:499, 2000.
- [77] R Benne, J Van den Burg, J P Brakenhoff, P Sloof, J H Van Boom, and M C Tromp. Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell*, 46(6):819 – 826, 1986.
- [78] Michael J. Palladino, Liam P. Keegan, Mary A. O’Connell, and Robert A. Reenan. A-to-I pre-mRNA editing in *Drosophila* is primarily involved in adult nervous system function and integrity. *Cell*, (4):437, 2000.
- [79] A P Gerber and W Keller. RNA editing by base deamination: more enzymes, more targets, new mysteries. *Trends In Biochemical Sciences*, 26(6):376 – 384, 2001.

- [80] Brenda L. Bass. RNA editing by adenosine deaminases that act on RNA. *Annual Review of Biochemistry*, 71(1):817, 2002.
- [81] Stefan Maas, Alexander Rich, and Kazuko Nishikura. A-to-I RNA Editing: Recent News and Residual Mysteries. *Journal of Biological Chemistry*, 278(3):1391, 2003.
- [82] Kim Unkyu, Wang Yi, Sanford Tamara, Zeng Yong, and Nishikura Kazuko. Molecular Cloning of cDNA for Double-Stranded RNA Adenosine Deaminase, a Candidate Enzyme for Nuclear RNA Editing. *Proceedings of the National Academy of Sciences of the United States of America*, (24):11457, 1994.
- [83] Myriam Schaub and Walter Keller. RNA editing by adenosine deaminases generates RNA and protein diversity. *Biochimie*, 84:791 – 803, 2002.
- [84] Kim Unkyu, Wang Yi, Sanford Tamara, Zeng Yong, and Nishikura Kazuko. Molecular Cloning of cDNA for Double-Stranded RNA Adenosine Deaminase, a Candidate Enzyme for Nuclear RNA Editing. *Proceedings of the National Academy of Sciences of the United States of America*, (24):11457, 1994.
- [85] M A O’Connell, S Krause, M Higuchi, J J Hsuan, N F Totty, A Jenny, and W Keller. Cloning of cDNAs encoding mammalian double-stranded RNA-specific adenosine deaminase. *Molecular And Cellular Biology*, 15(3):1389 – 1397, 1995.
- [86] F Lai, C X Chen, K C Carter, and K Nishikura. Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Molecular And Cellular Biology*, 17(5):2413 – 2424, 1997.
- [87] A Gerber, M A O’Connell, and W Keller. Two forms of human double-stranded RNA-specific editase 1 (hRED1) generated by the insertion of an Alu cassette. *RNA (New York, N.Y.)*, 3(5):453 – 463, 1997.
- [88] T Melcher, S Maas, A Herb, R Sprengel, M Higuchi, and P H Seeburg. RED2, a brain-specific member of the RNA-specific adenosine deaminase family. *The Journal Of Biological Chemistry*, 271(50):31795 – 31798, 1996.
- [89] C X Chen, D S Cho, Q Wang, F Lai, K C Carter, and K Nishikura. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. *RNA (New York, N.Y.)*, 6(5):755 – 767, 2000.
- [90] J B Patterson and C E Samuel. Expression and regulation by interferon

- of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Molecular And Cellular Biology*, 15(10):5376 – 5388, 1995.
- [91] Jolly Basak and Chandran Nithin. Targeting non-coding RNAs in plants with the CRISPR-Cas technology is a challenge yet worth accepting. *Frontiers in plant science*, 6:1001, 2015.
- [92] Thorsten Melcher, Stefan Maas, Anne Herb, Rolf Sprengel, Peter H. Seeburg, and Miyoko Higuchi. A mammalian RNA editing enzyme. *Nature*, (6564):460, 1996.
- [93] Kazuko Nishikura. Functions and regulation of RNA editing by ADAR deaminases. *Annual review of biochemistry*, 79:321–349, 2010.
- [94] Wahba Albert J., Miller Robert S., Basilio Carlos, Gardner Robert S., Lengyel Peter, and Speyer Joseph F. Synthetic Polynucleotides and the Amino Acid Code, IX. *Proceedings of the National Academy of Sciences of the United States of America*, (6):880, 1963.
- [95] Peter H Seeburg and Jochen Hartner. Regulation of ion channel/neurotransmitter receptor function by RNA editing. *Current Opinion in Neurobiology*, 13:279 – 283, 2003.
- [96] Susan M. Rueter, T. Renee Dawson, and Ronald B. Emeson. Regulation of alternative splicing by RNA editing. *Nature*, page 75, 1999.
- [97] Stefan Maas, Yukio Kawahara, Kristen M Tamburro, and Kazuko Nishikura. A-to-I RNA editing and human disease. *RNA Biology*, 3(1):1 – 9, 2006.
- [98] Mie Kubota-Sakashita, Kazuya Iwamoto, Miki Bundo, and Tadafumi Kato. A role of ADAR2 and RNA editing of glutamate receptors in mood disorders and schizophrenia. *Molecular Brain*, 7:5, 2014.
- [99] Yukio Kawahara, Kyoko Ito, Hui Sun, Hitoshi Aizawa, Ichiro Kanazawa, and Shin Kwak. Glutamate receptors: RNA editing and death of motor neurons. *Nature*, 427(6977):801, 2004.
- [100] Taeyoung Hwang, Chul-Kee Park, Anthony K.L. Leung, Yuan Gao, Thomas M. Hyde, Joel E. Kleinman, Anandita Rajpurohit, Ran Tao, Joo Heon Shin, and Daniel R. Weinberger. Dynamic regulation of RNA editing in human brain development and disease. *Nature Neuroscience*, (8):1093, 2016.
- [101] Leilei Chen, Yan Li, Chi Ho Lin, Tim Hon Man Chan, Raymond Kwok Kei Chow, Yangyang Song, Ming Liu, Yun-Fei Yuan, Li Fu, Kar Lok Kong, Lihua Qi, Na Zhang, Amy Hin Yan Tong, Dora Lai-Wan Kwong, Kwan Man,

- Chung Mau Lo, Si Lok, Daniel G Tenen, and Xin-Yuan Guan. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nature Medicine*, 19(2):209 – 216, 2013.
- [102] Yan-Ru Qin, Jun-Jing Qiao, Tim Hon Man Chan, Ying-Hui Zhu, Fang-Fang Li, Haibo Liu, Jing Fei, Yan Li, Xin-Yuan Guan, and Leilei Chen. Adenosine-to-inosine RNA editing mediated by ADARs in esophageal squamous cell carcinoma. *Cancer Research*, 74(3):840 – 851, 2014.
- [103] Tim Hon Man Chan, Aditi Qamra, Kar Tong Tan, Jing Guo, Henry Yang, Lihua Qi, Jaymie Siqi Lin, Vanessa Hui En Ng, Yangyang Song, Huiqi Hong, Su Ting Tay, Yujing Liu, Jeeyun Lee, Sun Yong Rha, Feng Zhu, Jimmy Bok Yan So, Bin Tean Teh, Khay Guan Yeoh, Steve Rozen, Daniel G Tenen, Patrick Tan, and Leilei Chen. ADAR-Mediated RNA Editing Predicts Progression and Prognosis of Gastric Cancer. *Gastroenterology*, 151(4):637–650.e10, 2016.
- [104] ADAR2-Dependent RNA Editing of AMPA Receptor Subunit GluR2 Determines Vulnerability of Neurons in Forebrain Ischemia. *Neuron*, (5):719, 2006.
- [105] Leng Han, Lixia Diao, Shuangxing Yu, Xiaoyan Xu, Jie Li, Rui Zhang, Yang Yang, Henrica M J Werner, A Karina Eterovic, Yuan Yuan, Jun Li, Nikitha Nair, Rosalba Minelli, Yiu Huen Tsang, Lydia W T Cheung, Kang Jin Jeong, Jason Roszik, Zhenlin Ju, Scott E Woodman, Yiling Lu, Kenneth L Scott, Jin Billy Li, Gordon B Mills, and Han Liang. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. *Cancer Cell*, 28(4):515 – 528, 2015.
- [106] Nurit Paz, Erez Y Levanon, Ninette Amariglio, Amy B Heimberger, Zvi Ram, Shlomi Constantini, Zohar S Barbash, Konstantin Adamsky, Michal Safran, Avi Hirschberg, Meir Krupsky, Issachar Ben-Dov, Simona Cazacu, Tom Mikkelsen, Chaya Brodie, Eli Eisenberg, and Gideon Rechavi. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Research*, 17(11):1586 – 1595, 2007.
- [107] Paz-Yaacov Nurit, Bazak Lily, Buchumenski Ilana, Porath Hagit T., Danan-Gotthold Miri, Knisbacher Binyamin A., Eisenberg Eli, and Levanon Erez Y. Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. *Cell Reports*, Vol 13, Iss 2, Pp 267-276 (2015), (2):267, 2015.
- [108] Xinxin Peng, Xiaoyan Xu, Yumeng Wang, David H. Hawke, Shuangxing

- Yu, Leng Han, Zhicheng Zhou, Kamalika Mojumdar, Kang Jin Jeong, Marilyne Labrie, Yiu Huen Tsang, Minying Zhang, Yiling Lu, Patrick Hwu, Kenneth L. Scott, Han Liang, and Gordon B. Mills. A-to-I RNA Editing Contributes to Proteomic Diversity in Cancer. *Cancer Cell*, (5), 2018.
- [109] Caterina Cenci, Rita Barzotti, Federica Galeano, Sandro Corbelli, Rossella Rota, Luca Massimi, Concezio Di Rocco, Mary A O’Connell, and Angela Gallo. Down-regulation of RNA editing in pediatric astrocytomas: ADAR2 editing activity inhibits cell migration and proliferation. *The Journal Of Biological Chemistry*, 283(11):7251 – 7260, 2008.
- [110] Fumagalli Debora, Gacquer David, Rothé Françoise, Lefort Anne, Libert Frederick, Brown David, Kheddoumi Naima, Shlien Adam, Konopka Tomasz, Salgado Roberto, Larsimont Denis, Polyak Kornelia, Willard-Gallo Karen, Desmedt Christine, Piccart Martine, Abramowicz Marc, Campbell Peter J., Sotiriou Christos, and Detours Vincent. Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. *Cell Reports, Vol 13, Iss 2, Pp 277-289 (2015)*, (2):277, 2015.
- [111] Deepanjan Paul, Ashis Narayan Sinha, Arjun Ray, Megha Lal, Subhashree Nayak, Anchal Sharma, Bharati Mehani, Debasish Mukherjee, Saurabh V Laddha, Ashish Suri, Chitra Sarkar, and Arijit Mukhopadhyay. A-to-I editing in human miRNAs is enriched in seed sequence, influenced by sequence contexts and significantly hypoedited in glioblastoma multiforme. *Scientific Reports*, 7(1):2466, 2017.
- [112] Gan Zhenji, Zhao Liyun, Yang Liu, Huang Ping, Zhao Feng, Li Wenjun, and Liu Yong. RNA Editing by ADAR2 Is Metabolically Regulated in Pancreatic Islets and β -Cells. *Journal of Biological Chemistry*, 281(44):33386 – 33394, 2006.
- [113] Liu Yang, Liyun Zhao, Zhenji Gan, Zixuan He, Jingyue Xu, Xiang Gao, Xiaorui Wang, Weiping Han, Liangyi Chen, Tao Xu, Wenjun Li, and Yong Liu. Deficiency in RNA editing enzyme ADAR2 impairs regulated exocytosis. *FASEB Journal: Official Publication Of The Federation Of American Societies For Experimental Biology*, 24(10):3720 – 3732, 2010.
- [114] C. E. Samuel. ADARs: viruses and innate immunity. *Curr. Top. Microbiol. Immunol.*, 353:163–195, 2012.
- [115] Elitza Deltcheva, Krzysztof Chylinski, Cynthia M. Sharma, Karine Gonzales, Zaid A. Pirzada, Maria R. Eckert, Jörg Vogel, and Emmanuelle Charpentier. CRISPR RNA maturation by trans-encoded small RNA and host factor

- RNase III. *Nature*, pages 602–607, 2011.
- [116] Kira S Makarova, Daniel H Haft, Rodolphe Barrangou, Stan JJ Brouns, Emmanuelle Charpentier, Philippe Horvath, Sylvain Moineau, Francisco JM Mojica, Yuri I Wolf, Alexander F Yakunin, et al. Evolution and classification of the CRISPR–Cas systems. *Nature Reviews Microbiology*, 9(6):467, 2011.
- [117] Giedrius Gasiunas, Rodolphe Barrangou, Philippe Horvath, and Virginijus Siksnys. Cas9–crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. *Proceedings of the National Academy of Sciences*, 109(39):E2579–E2586, 2012.
- [118] Xuelian Wang, Xiumin Huang, Xiuli Fang, Youzhong Zhang, and Wanpeng Wang. CRISPR-Cas9 system as a versatile tool for genome engineering in human cells. *Molecular Therapy-Nucleic Acids*, 5:e388, 2016.
- [119] Hiroshi Nishimasu, F Ann Ran, Patrick D Hsu, Silvana Konermann, Soraya I Shehata, Naoshi Dohmae, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell*, 156(5):935–949, 2014.
- [120] Hiroshi Nishimasu, Le Cong, Winston X Yan, F Ann Ran, Bernd Zetsche, Yinqing Li, Arisa Kurabayashi, Ryuichiro Ishitani, Feng Zhang, and Osamu Nureki. Crystal structure of *Staphylococcus aureus* Cas9. *Cell*, 162(5):1113–1126, 2015.
- [121] Hisato Hirano, Jonathan S Gootenberg, Takuro Horii, Omar O Abudayyeh, Mika Kimura, Patrick D Hsu, Takanori Nakane, Ryuichiro Ishitani, Izuho Hatada, Feng Zhang, et al. Structure and engineering of *Francisella novicida* Cas9. *Cell*, 164(5):950–961, 2016.
- [122] Carolin Anders, Ole Niewoehner, Alessia Duerst, and Martin Jinek. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature*, 513(7519):569, 2014.
- [123] Fuguo Jiang, David W Taylor, Janice S Chen, Jack E Kornfeld, Kaihong Zhou, Aubri J Thompson, Eva Nogales, and Jennifer A Doudna. Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. *Science*, 351(6275):867–871, 2016.
- [124] Karolina A Majorek, Stanislaw Dunin-Horkawicz, Kamil Steczkiewicz, Anna Muszewska, Marcin Nowotny, Krzysztof Ginalski, and Janusz M Bujnicki. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic acids research*, 42(7):4160–4179, 2014.

- [125] Giulia Palermo, Yinglong Miao, Ross C Walker, Martin Jinek, and J Andrew McCammon. Striking plasticity of CRISPR-Cas9 and key role of non-target DNA, as revealed by molecular simulations. *ACS central science*, 2(10):756–763, 2016.
- [126] Zhicheng Zuo and Jin Liu. Cas9-catalyzed DNA cleavage generates staggered ends: Evidence from molecular dynamics simulations. *Scientific reports*, 6:37584, 2016.
- [127] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O’donovan, Isabelle Phan, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic acids research*, 31(1):365–370, 2003.
- [128] F Sievers and DG Higgins. Clustal Omega. *Current Protocols in Bioinformatics*. 2002.
- [129] Fabian Sievers and Desmond G Higgins. Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods*, pages 105–116. Springer, 2014.
- [130] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Söding, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology*, 7(1):539, 2011.
- [131] Andrew M Waterhouse, James B Procter, David MA Martin, Michèle Clamp, and Geoffrey J Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.
- [132] Protein Data Bank. Protein data bank. *Nature New Biol*, 233:223, 1971.
- [133] Christian Biertümpfel, Wei Yang, and Dietrich Suck. Crystal structure of T4 endonuclease VII resolving a Holliday junction. *Nature*, 449(7162):616, 2007.
- [134] D. A. Case, V. Babin, J. T. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham, T. A. Darden, R. E. Duke, H. Gohlke, A. W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. M. Merz, F. Paesani, D. R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. L. Simmerling, W. Smith, J. Swails, Walker, J. Wang, R. M. Wolf, X. Wu, and

- P. A. Kollman. *Amber 14*. University of California, San Francisco, 2014.
- [135] J. A. Maier, C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser, and C. Simmerling. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput*, 11(8):3696–3713, Aug 2015.
- [136] Pengfei Li, Lin Frank Song, and Kenneth M Merz Jr. Parameterization of highly charged metal ions using the 12-6-4 LJ-type nonbonded model in explicit water. *The Journal of Physical Chemistry B*, 119(3):883–895, 2014.
- [137] William L Jorgensen, Jayaraman Chandrasekhar, Jeffery D Madura, Roger W Impey, and Michael L Klein. Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, 79(2):926–935, 1983.
- [138] Tom Darden, Darrin York, and Lee Pedersen. Particle mesh Ewald: An $N \log(N)$ method for Ewald sums in large systems. *The Journal of chemical physics*, 98(12):10089–10092, 1993.
- [139] Warren L DeLano. The PyMOL molecular graphics system. <http://www.pymol.org>, 2002.
- [140] William Humphrey, Andrew Dalke, and Klaus Schulten. VMD: visual molecular dynamics. *Journal of molecular graphics*, 14(1):33–38, 1996.
- [141] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*, 25(13):1605–1612, 2004.
- [142] James Hemphill, Erin K Borchardt, Kalyn Brown, Aravind Asokan, and Alexander Deiters. Optical control of CRISPR/Cas9 gene editing. *Journal of the American Chemical Society*, 137(17):5642–5645, 2015.
- [143] Cong Huai, Gan Li, Ruijie Yao, Yingyi Zhang, Mi Cao, Liangliang Kong, Chenqiang Jia, Hui Yuan, Hongyan Chen, Daru Lu, et al. Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nature communications*, 8(1):1375, 2017.
- [144] Oleg Yu Fedoroff, Miguel Salazar, and Brian R Reid. Structure of a DNA: RNA hybrid duplex: why RNase H does not cleave pure RNA. *Journal of molecular biology*, 233(3):509–523, 1993.
- [145] De Dong, Kuan Ren, Xiaolin Qiu, Jianlin Zheng, Minghui Guo, Xiaoyu Guan, Hongnan Liu, Ningning Li, Bailing Zhang, Daijun Yang, et al. The crystal structure of Cpf1 in complex with CRISPR RNA. *Nature*,

- 532(7600):522, 2016.
- [146] Takashi Yamano, Hiroshi Nishimasu, Bernd Zetsche, Hisato Hirano, Ian M Slaymaker, Yinqing Li, Iana Fedorova, Takanori Nakane, Kira S Makarova, Eugene V Koonin, et al. Crystal structure of Cpf1 in complex with guide RNA and target DNA. *Cell*, 165(4):949–962, 2016.
 - [147] Jennifer A Doudna and Emmanuelle Charpentier. The new frontier of genome engineering with CRISPR-Cas9. *Science*, 346(6213):1258096, 2014.
 - [148] Patrick D Hsu, Eric S Lander, and Feng Zhang. Development and applications of CRISPR-Cas9 for genome engineering. *Cell*, 157(6):1262–1278, 2014.
 - [149] Le Cong, F Ann Ran, David Cox, Shuailiang Lin, Robert Barretto, Naomi Habib, Patrick D Hsu, Xuebing Wu, Wenyan Jiang, Luciano Marraffini, et al. Multiplex genome engineering using CRISPR/Cas systems. *Science*, page 1231143, 2013.
 - [150] Prashant Mali, Luhan Yang, Kevin M Esvelt, John Aach, Marc Guell, James E DiCarlo, Julie E Norville, and George M Church. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826, 2013.
 - [151] Martin Jinek, Alexandra East, Aaron Cheng, Steven Lin, Enbo Ma, and Jennifer Doudna. RNA-programmed genome editing in human cells. *elife*, 2:e00471, 2013.
 - [152] Zhonggang Hou, Yan Zhang, Nicholas E Propson, Sara E Howden, Li-Fang Chu, Erik J Sontheimer, and James A Thomson. Efficient genome engineering in human pluripotent stem cells using Cas9 from *Neisseria meningitidis*. *Proceedings of the National Academy of Sciences*, 110(39):15644–15649, 2013.
 - [153] Eunji Kim, Taeyoung Koo, Sung Wook Park, Daesik Kim, Kyoungmi Kim, Hee-Yeon Cho, Dong Woo Song, Kyu Jun Lee, Min Hee Jung, Seokjoong Kim, et al. In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nature Communications*, 8:14500, 2017.
 - [154] Alexis C Komor, Ahmed H Badran, and David R Liu. CRISPR-based technologies for the manipulation of eukaryotic genomes. *Cell*, 168(1-2):20–36, 2017.
 - [155] Benjamin P Kleinstiver, Shengdar Q Tsai, Michelle S Prew, Nhu T Nguyen, Moira M Welch, Jose M Lopez, Zachary R McCaw, Martin J Aryee, and J Keith Joung. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature biotechnology*, 34(8):869, 2016.

- [156] Daesik Kim, Jungeun Kim, Junho K Hur, Kyung Wook Been, Sun-heui Yoon, and Jin-Soo Kim. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature biotechnology*, 34(8):863, 2016.
- [157] Cedric R Clapier and Bradley R Cairns. The biology of chromatin remodeling complexes. *Annual review of biochemistry*, 78:273–304, 2009.
- [158] Yuanming Wang, Kaiwen Ivy Liu, Norfala-Aliah Binte Sutrisnoh, Harini Srinivasan, Junyi Zhang, Jia Li, Fan Zhang, Charles Richard John Lalith, Heyun Xing, Raghuvaran Shanmugam, et al. Systematic evaluation of CRISPR-Cas systems reveals design principles for genome editing in human cells. *Genome biology*, 19(1):62, 2018.
- [159] Jeffry D Sander and J Keith Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature biotechnology*, 32(4):347, 2014.
- [160] Picard toolkit. <http://broadinstitute.github.io/picard/>, 2018.
- [161] John M Hinz, Marian F Laughery, and John J Wyrick. Nucleosomes inhibit Cas9 endonuclease activity in vitro. *Biochemistry*, 54(48):7063–7066, 2015.
- [162] Max A Horlbeck, Lea B Witkowsky, Benjamin Guglielmi, Joseph M Replogle, Luke A Gilbert, Jacqueline E Villalta, Sharon E Torigoe, Robert Tjian, and Jonathan S Weissman. Nucleosomes impede Cas9 access to DNA in vivo and in vitro. *Elife*, 5:e12677, 2016.
- [163] R Stefan Isaac, Fuguo Jiang, Jennifer A Doudna, Wendell A Lim, Geeta J Narlikar, and Ricardo Almeida. Nucleosome breathing and remodeling constrain CRISPR-Cas9 function. *Elife*, 5:e13450, 2016.
- [164] René M Daer, Josh P Cutts, David A Brafman, and Karmella A Haynes. The impact of chromatin dynamics on Cas9-mediated genome editing in human cells. *ACS synthetic biology*, 6(3):428–438, 2016.
- [165] Steven Lin, Brett T Staahl, Ravi K Alla, and Jennifer A Doudna. Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *elife*, 3:e04766, 2014.
- [166] Luhan Yang, Marc Guell, Susan Byrne, Joyce L Yang, Alejandro De Los Angeles, Prashant Mali, John Aach, Caroline Kim-Kiselak, Adrian W Briggs, Xavier Rios, et al. Optimization of scarless human stem cell genome editing. *Nucleic acids research*, 41(19):9049–9061, 2013.
- [167] Christopher D Richardson, Graham J Ray, Mark A DeWitt, Gemma L Curie, and Jacob E Corn. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature biotechnology*, 34(3):339, 2016.

- [168] Miguel A Moreno-Mateos, Juan P Fernandez, Romain Rouet, Charles E Vejnar, Maura A Lane, Emily Mis, Mustafa K Khokha, Jennifer A Doudna, and Antonio J Giraldez. CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing. *Nature communications*, 8(1):2024, 2017.
- [169] Luhan Yang, Marc Guell, Susan Byrne, Joyce L Yang, Alejandro De Los Angeles, Prashant Mali, John Aach, Caroline Kim-Kiselak, Adrian W Briggs, Xavier Rios, et al. Optimization of scarless human stem cell genome editing. *Nucleic acids research*, 41(19):9049–9061, 2013.
- [170] Christopher D Richardson, Graham J Ray, Mark A DeWitt, Gemma L Curie, and Jacob E Corn. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nature biotechnology*, 34(3):339, 2016.
- [171] Christopher D Richardson, Graham J Ray, NL Bray, and JE Corn. Non-homologous DNA increases gene disruption efficiency by altering DNA repair outcomes. *Nature communications*, 7:12463, 2016.
- [172] Benjamin P Kleinstiver, Shengdar Q Tsai, Michelle S Prew, Nhu T Nguyen, Moira M Welch, Jose M Lopez, Zachary R McCaw, Martin J Aryee, and J Keith Joung. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature biotechnology*, 34(8):869, 2016.
- [173] Daesik Kim, Jungeun Kim, Junho K Hur, Kyung Wook Been, Sun-heui Yoon, and Jin-Soo Kim. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. *Nature biotechnology*, 34(8):863, 2016.
- [174] Daesik Kim, Sangsu Bae, Jeongbin Park, Eunji Kim, Seokjoong Kim, Hye Ryeong Yu, Jinha Hwang, Jong-Il Kim, and Jin-Soo Kim. Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nature methods*, 12(3):237, 2015.
- [175] Daesik Kim, Sojung Kim, Sunghyun Kim, Jeongbin Park, and KIM Jin-Soo. Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome research*, pages gr-199588, 2016.
- [176] Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Iafrate, Long P Le, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nature biotechnology*, 33(2):187, 2015.
- [177] Benjamin P Kleinstiver, Shengdar Q Tsai, Michelle S Prew, Nhu T Nguyen, Moira M Welch, Jose M Lopez, Zachary R McCaw, Martin J Aryee, and

- J Keith Joung. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. *Nature biotechnology*, 34(8):869, 2016.
- [178] Van Trung Chu, Timm Weber, Benedikt Wefers, Wolfgang Wurst, Sandrine Sander, Klaus Rajewsky, and Ralf Kühn. Increasing the efficiency of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. *Nature biotechnology*, 33(5):543, 2015.
- [179] Takeshi Maruyama, Stephanie K Dougan, Matthias C Truttmann, Angelina M Bilate, Jessica R Ingram, and Hidde L Ploegh. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of non-homologous end joining. *Nature biotechnology*, 33(5):538, 2015.
- [180] Chen Yu, Yanxia Liu, Tianhua Ma, Kai Liu, Shaohua Xu, Yu Zhang, Honglei Liu, Marie La Russa, Min Xie, Sheng Ding, et al. Small molecules enhance CRISPR genome editing in pluripotent stem cells. *Cell stem cell*, 16(2):142–147, 2015.
- [181] Jordan Pinder, Jayme Salsman, and Graham Dellaire. Nuclear domain ‘knock-in’ screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic acids research*, 43(19):9379–9392, 2015.
- [182] Jun Song, Dongshan Yang, Jie Xu, Tianqing Zhu, Y Eugene Chen, and Jifeng Zhang. RS-1 enhances CRISPR/Cas9-and TALEN-mediated knock-in efficiency. *Nature communications*, 7:10548, 2016.
- [183] Tony Gutschner, Monika Haemmerle, Giannicola Genovese, Giulio F Draetta, and Lynda Chin. Post-translational regulation of Cas9 during G1 enhances homology-directed repair. *Cell reports*, 14(6):1555–1566, 2016.
- [184] Jian-Ping Zhang, Xiao-Lan Li, Guo-Hua Li, Wanqiu Chen, Cameron Arakaki, Gary D Botimer, David Baylink, Lu Zhang, Wei Wen, Ya-Wen Fu, et al. Efficient precise knockin with a double cut HDR donor after CRISPR/Cas9-mediated double-stranded DNA cleavage. *Genome biology*, 18(1):35, 2017.
- [185] Miguel A Moreno-Mateos, Juan P Fernandez, Romain Rouet, Charles E Vejnar, Maura A Lane, Emily Mis, Mustafa K Khokha, Jennifer A Doudna, and Antonio J Giraldez. CRISPR-Cpf1 mediates efficient homology-directed repair and temperature-controlled genome editing. *Nature communications*, 8(1):2024, 2017.
- [186] K. K. Chiruvella, Z. Liang, and T. E. Wilson. Repair of double-strand breaks by end joining. *Cold Spring Harb Perspect Biol*, 5(5):a012757, May 2013.

- [187] J. Her and S. F. Bunting. How cells ensure correct repair of DNA double-strand breaks. *J. Biol. Chem.*, 293(27):10502–10511, 07 2018.
- [188] L. N. Truong, Y. Li, L. Z. Shi, P. Y. Hwang, J. He, H. Wang, N. Razavian, M. W. Berns, and X. Wu. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proc. Natl. Acad. Sci. U.S.A.*, 110(19):7720–7725, May 2013.
- [189] A. Shibata. Regulation of repair pathway choice at two-ended DNA double-strand breaks. *Mutat. Res.*, 803-805:51–55, 10 2017.
- [190] T. Helleday, S. Eshtad, and S. Nik-Zainal. Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.*, 15(9):585–598, Sep 2014.
- [191] S. Bae, J. Kweon, H. S. Kim, and J. S. Kim. Microhomology-based choice of Cas9 nuclease target sites. *Nat. Methods*, 11(7):705–706, Jul 2014.
- [192] Y. Kim, J. Kweon, and J. S. Kim. TALENs and ZFNs are associated with different mutation signatures. *Nat. Methods*, 10(3):185, Mar 2013.
- [193] F. Robert, M. Barbeau, S. Ethier, J. Dostie, and J. Pelletier. Pharmacological inhibition of DNA-PK stimulates Cas9-mediated genome editing. *Genome Med*, 7:93, Aug 2015.
- [194] M. van Overbeek, D. Capurso, M. M. Carter, M. S. Thompson, E. Frias, C. Russ, J. S. Reece-Hoyes, C. Nye, S. Gradia, B. Vidal, J. Zheng, G. R. Hoffman, C. K. Fuller, and A. P. May. DNA Repair Profiling Reveals Non-random Outcomes at Cas9-Mediated Breaks. *Mol. Cell*, 63(4):633–646, 08 2016.
- [195] Raj Chari, Prashant Mali, Mark Moosburner, and George M Church. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nature Methods*, 12(9):823 – 826, 2015.
- [196] T. Gutschner, M. Haemmerle, G. Genovese, G. F. Draetta, and L. Chin. Post-translational Regulation of Cas9 during G1 Enhances Homology-Directed Repair. *Cell Rep*, 14(6):1555–1566, Feb 2016.
- [197] V. Y. Khodaverdian, T. Hanscom, A. M. Yu, T. L. Yu, V. Mak, A. J. Brown, S. A. Roberts, and M. McVey. Secondary structure forming sequences drive SD-MMEJ repair of DNA double-strand breaks. *Nucleic Acids Res.*, 45(22):12848–12861, Dec 2017.
- [198] A. Taheri-Ghahfarokhi, B. J. M. Taylor, R. Nitsch, A. Lundin, A. L. Cavallo, K. Madeyski-Bengtson, F. Karlsson, M. Clausen, R. Hicks, L. M. Mayr,

- M. Bohlooly-Y, and M. Maresca. Decoding non-random mutational signatures at Cas9 targeted sites. *Nucleic Acids Res.*, 46(16):8417–8434, Sep 2018.
- [199] P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, and G. M. Church. RNA-guided human genome engineering via Cas9. *Science*, 339(6121):823–826, Feb 2013.
- [200] T. Wang, K. Birsoy, N. W. Hughes, K. M. Krupczak, Y. Post, J. J. Wei, E. S. Lander, and D. M. Sabatini. Identification and characterization of essential genes in the human genome. *Science*, 350(6264):1096–1101, Nov 2015.
- [201] M. Gasperini, G. M. Findlay, A. McKenna, J. H. Milbank, C. Lee, M. D. Zhang, D. A. Cusanovich, and J. Shendure. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *Am. J. Hum. Genet.*, 101(2):192–205, Aug 2017.
- [202] J. C. Klein, W. Chen, M. Gasperini, and J. Shendure. Identifying Novel Enhancer Elements with CRISPR-Based Screens. *ACS Chem. Biol.*, 13(2):326–332, 02 2018.
- [203] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole-organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 353(6298):aaf7907, Jul 2016.
- [204] K. L. Frieda, J. M. Linton, S. Hormoz, J. Choi, K. K. Chow, Z. S. Singer, M. W. Budde, M. B. Elowitz, and L. Cai. Synthetic recording and in situ readout of lineage information in single cells. *Nature*, 541(7635):107–111, 01 2017.
- [205] R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, and G. M. Church. Developmental barcoding of whole mouse via homing CRISPR. *Science*, 361(6405), 08 2018.
- [206] M M Purugganan, S Shah, J F Kearney, and D B Roth. Ku80 is required for addition of N nucleotides to V(D)J recombination junctions by terminal deoxynucleotidyl transferase. *Nucleic Acids Research*, 29(7):1638 – 1646, 2001.
- [207] B. Corneo, R. L. Wendland, L. Deriano, X. Cui, I. A. Klein, S. Y. Wong, S. Arnal, A. J. Holub, G. R. Weller, B. A. Pancake, S. Shah, V. L. Brandt, K. Meek, and D. B. Roth. Rag mutations reveal robust alternative end joining. *Nature*, 449(7161):483–486, Sep 2007.

- [208] P. Rouet, F. Smih, and M. Jasin. Introduction of double-strand breaks into the genome of mouse cells by expression of a rare-cutting endonuclease. *Mol. Cell. Biol.*, 14(12):8096–8106, Dec 1994.
- [209] A. H. Lloyd, D. Wang, and J. N. Timmis. Single molecule PCR reveals similar patterns of non-homologous DSB repair in tobacco and Arabidopsis. *PLoS ONE*, 7(2):e32255, 2012.
- [210] R. S. Bindra, A. G. Goglia, M. Jasin, and S. N. Powell. Development of an assay to measure mutagenic non-homologous end-joining repair activity in mammalian cells. *Nucleic Acids Res.*, 41(11):e115, Jun 2013.
- [211] C. P. Soong, G. A. Breuer, R. A. Hannon, S. D. Kim, A. F. Salem, G. Wang, R. Yu, N. J. Carriero, R. Bjornson, R. K. Sundaram, and R. S. Bindra. Development of a novel method to create double-strand break repair fingerprints using next-generation sequencing. *DNA Repair (Amst.)*, 26:44–53, Feb 2015.
- [212] J. K. Moore and J. E. Haber. Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, 16(5):2164–2173, May 1996.
- [213] P. Li, J. Li, M. Li, K. Dou, M. J. Zhang, F. Suo, and L. L. Du. Multiple end joining mechanisms repair a chromosomal DNA break in fission yeast. *DNA Repair (Amst.)*, 11(2):120–130, Feb 2012.
- [214] M. W. Shen, M. Arbab, J. Y. Hsu, D. Worstell, S. J. Culbertson, O. Krabbe, C. A. Cassa, D. R. Liu, D. K. Gifford, and R. I. Sherwood. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, 563(7733):646–651, 11 2018.
- [215] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001–. [Online; accessed <today>].
- [216] Felicity Allen, Luca Crepaldi, Clara Alsinet, Alexander J Strong, Vitalii Kleshchevnikov, Pietro De Angeli, Petra Páleníková, Anton Khodak, Vladimir Kiselev, Michael Kosicki, et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nature biotechnology*, 37(1):64, 2019.
- [217] Gary Walsh. Biopharmaceutical benchmarks 2014. *Nature biotechnology*, 32(10):992, 2014.
- [218] Jennifer Dumont, Don Euwart, Baisong Mei, Scott Estes, and Rashmi Kshirsagar. Human cell lines for biopharmaceutical manufacturing: history, status, and future perspectives. *Critical reviews in biotechnology*, 36(6):1110–

- 1122, 2016.
- [219] Marie-Eve Lalonde and Yves Durocher. Therapeutic glycoprotein production in mammalian cells. *Journal of biotechnology*, 251:128–140, 2017.
 - [220] Jianwei Zhu. Mammalian cell protein expression for biopharmaceutical production. *Biotechnology advances*, 30(5):1158–1170, 2012.
 - [221] Tingfeng Lai, Yuansheng Yang, and Say Ng. Advances in mammalian cell line development technologies for recombinant protein production. *Pharmaceuticals*, 6(5):579–603, 2013.
 - [222] Justin Bryan Goh and Say Kong Ng. Impact of host cell line choice on glycan profile. *Critical reviews in biotechnology*, 38(6):851–867, 2018.
 - [223] Darius Ghaderi, Rachel E Taylor, Vered Padler-Karavani, Sandra Diaz, and Ajit Varki. Implications of the presence of N-glycolylneuraminic acid in recombinant therapeutic glycoproteins. *Nature biotechnology*, 28(8):863, 2010.
 - [224] Lukas Fliedl, Johannes Grillari, and Regina Grillari-Voglauer. Human cell lines for the production of recombinant proteins: on the horizon. *New biotechnology*, 32(6):673–679, 2015.
 - [225] Kathleen L Berkner. Expression of recombinant vitamin K-dependent proteins in mammalian cells: Factors IX and VII. In *Methods in enzymology*, volume 222, pages 450–477. Elsevier, 1993.
 - [226] Christoph Kannicht, Margareta Ramström, Guido Kohla, Maya Tiemeyer, Elisabeth Casademunt, Olaf Walter, and Helena Sandberg. Characterisation of the post-translational modifications of a novel, human cell line-derived recombinant human factor VIII. *Thrombosis research*, 131(1):78–88, 2013.
 - [227] RT Peters, G Toby, Q Lu, T Liu, JD Kulman, SC Low, AJ Bitonti, and GF Pierce. Biochemical and functional characterization of a recombinant monomeric factor VIII–Fc fusion protein. *Journal of Thrombosis and Haemostasis*, 11(1):132–141, 2013.
 - [228] Philip Thomas and Trevor G Smart. HEK293 cell line: a vehicle for the expression of recombinant proteins. *Journal of pharmacological and toxicological methods*, 51(3):187–200, 2005.
 - [229] Petra Meissner, Horst Pick, Alexandra Kulangara, Philippe Chatellard, Kirstin Friedrich, and Florian M Wurm. Transient gene expression: recombinant protein production with suspension-adapted HEK293-EBNA cells. *Biotechnology and bioengineering*, 75(2):197–203, 2001.
 - [230] Martin Loignon, Sylvie Perret, John Kelly, Denise Boulais, Brian Cass, Louis Bisson, Fatemeh Afkhamizarreh, and Yves Durocher. Stable high volumetric

- production of glycosylated human recombinant IFN α 2b in HEK293 cells. *BMC biotechnology*, 8(1):65, 2008.
- [231] Jared L Spidel, Benjamin Vaessen, Yin Yin Chan, Luigi Grasso, and J Bradford Kline. Rapid high-throughput cloning and stable expression of antibodies in HEK293 cells. *Journal of immunological methods*, 439:50–58, 2016.
- [232] Lianchun Fan, Ibrahim Kadura, Lara E Krebs, Christopher C Hatfield, Margaret M Shaw, and Christopher C Frye. Improving the efficiency of CHO cell line generation using glutamine synthetase gene knockout cells. *Biotechnology and bioengineering*, 109(4):1007–1015, 2012.
- [233] Soo Min Noh, Gyun Min Lee, et al. Limitations to the development of recombinant human embryonic kidney 293E cells using glutamine synthetase-mediated gene amplification: methionine sulfoximine resistance. *Journal of biotechnology*, 231:136–140, 2016.
- [234] Da Young Yu, Sang Yoon Lee, and Gyun Min Lee. Glutamine synthetase gene knockout-human embryonic kidney 293E cells for stable production of monoclonal antibodies. *Biotechnology and bioengineering*, 115(5):1367–1372, 2018.
- [235] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology*, 15(12):550, 2014.
- [236] Da Wei Huang, Brad T Sherman, Qina Tan, Joseph Kir, David Liu, David Bryant, Yongjian Guo, Robert Stephens, Michael W Baseler, H Clifford Lane, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic acids research*, 35(suppl_2):W169–W175, 2007.
- [237] Leilei Chen, Yan Li, Chi Ho Lin, Tim Hon Man Chan, Raymond Kwok Kei Chow, Yangyang Song, Ming Liu, Yun-Fei Yuan, Li Fu, Kar Lok Kong, et al. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. *Nature medicine*, 19(2):209, 2013.
- [238] Yan-Ru Qin, Jun-Jing Qiao, Tim Hon Man Chan, Ying-Hui Zhu, Fang-Fang Li, Haibo Liu, Jing Fei, Yan Li, Xin-Yuan Guan, and Leilei Chen. Adenosine-to-inosine RNA editing mediated by ADARs in esophageal squamous cell carcinoma. *Cancer research*, 74(3):840–851, 2014.
- [239] Tim Hon Man Chan, Aditi Qamra, Kar Tong Tan, Jing Guo, Henry Yang, Lihua Qi, Jaymie Siqi Lin, Vanessa Hui En Ng, Yangyang Song, Huiqi Hong, et al. ADAR-mediated RNA editing predicts progression and prognosis of

- gastric cancer. *Gastroenterology*, 151(4):637–650, 2016.
- [240] Mie Kubota-Sakashita, Kazuya Iwamoto, Miki Bundo, and Tadafumi Kato. A role of ADAR2 and RNA editing of glutamate receptors in mood disorders and schizophrenia. *Molecular brain*, 7(1):5, 2014.
- [241] Yukio Kawahara, Kyoko Ito, Hui Sun, Hitoshi Aizawa, Ichiro Kanazawa, and Shin Kwak. Glutamate receptors: RNA editing and death of motor neurons. *Nature*, 427(6977):801, 2004.
- [242] Taeyoung Hwang, Chul-Kee Park, Anthony KL Leung, Yuan Gao, Thomas M Hyde, Joel E Kleinman, Anandita Rajpurohit, Ran Tao, Joo Heon Shin, and Daniel R Weinberger. Dynamic regulation of RNA editing in human brain development and disease. *Nature neuroscience*, 19(8):1093, 2016.
- [243] Khen Khmermesh, Anna Maria D’Erchia, Michal Barak, Anita Annese, Chaim Wachtel, Erez Y Levanon, Ernesto Picardi, and Eli Eisenberg. Reduced levels of protein recoding by A-to-I RNA editing in Alzheimer’s disease. *Rna*, 22(2):290–302, 2016.
- [244] Q Wang, J Khillan, P Gadue, and K Nishikura. Requirement of the RNA editing deaminase ADAR1 gene for embryonic erythropoiesis. *Science*, 290(5497):1765–1768, 2000.
- [245] Qingde Wang, Mana Miyakoda, Weidong Yang, Jaspal Khillan, David L Stachura, Mitchell J Weiss, and Kazuko Nishikura. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. *Journal of Biological Chemistry*, 279(6):4952–4961, 2004.
- [246] Jochen C Hartner, Carolin Schmittwolf, Andreas Kispert, Albrecht M Müller, Miyoko Higuchi, and Peter H Seeburg. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. *Journal of Biological Chemistry*, 279(6):4894–4902, 2004.
- [247] Brian J Liddicoat, Robert Piskol, Alistair M Chalk, Gokul Ramaswami, Miyoko Higuchi, Jochen C Hartner, Jin Billy Li, Peter H Seeburg, and Carl R Walkley. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science*, 349(6252):1115–1120, 2015.
- [248] Miyoko Higuchi, Stefan Maas, Frank N Single, Jochen Hartner, Andrei Rozov, Nail Burnashev, Dirk Feldmeyer, Rolf Sprengel, and Peter H Seeburg. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. *Nature*, 406(6791):78, 2000.
- [249] Peter L Peng, Xiafen Zhong, Weihong Tu, Mangala M Soundarapandian,

- Peter Molner, Dongya Zhu, Lorraine Lau, Shuhong Liu, Fang Liu, and YouMing Lu. ADAR2-dependent RNA editing of AMPA receptor subunit GluR2 determines vulnerability of neurons in forebrain ischemia. *Neuron*, 49(5):719–733, 2006.
- [250] Shuchen Lee, Guang Yang, Yue Yong, Ying Liu, Liyun Zhao, Jing Xu, Xiaomin Zhang, Yanjie Wan, Chun Feng, Zhiqin Fan, et al. ADAR2-dependent RNA editing of GluR2 is involved in thiamine deficiency-induced alteration of calcium dynamics. *Molecular neurodegeneration*, 5(1):54, 2010.
- [251] Susan M Rueter, T Renee Dawson, and Ronald B Emeson. Regulation of alternative splicing by RNA editing. *Nature*, 399(6731):75, 1999.
- [252] Joana MP Desterro, Liam P Keegan, Ellis Jaffray, Ron T Hay, Mary A O’Connell, and Maria Carmo-Fonseca. SUMO-1 modification alters ADAR1 editing activity. *Molecular biology of the cell*, 16(11):5115–5126, 2005.
- [253] Roberto Marcucci, James Brindle, Simona Paro, Angela Casadio, Sophie Hempel, Nicholas Morrice, Andrea Bisso, Liam P Keegan, Giannino Del Sal, and Mary A O’connell. Pin1 and WWP2 regulate GluR2 Q/R site RNA editing by ADAR2 with opposing effects. *The EMBO journal*, 30(20):4211–4222, 2011.
- [254] Meng How Tan, Qin Li, Raghuvaran Shanmugam, Robert Piskol, Jennefer Kohler, Amy N Young, Kaiwen Ivy Liu, Rui Zhang, Gokul Ramaswami, Kentaro Ariyoshi, et al. Dynamic landscape and regulation of RNA editing in mammals. *Nature*, 550(7675):249, 2017.
- [255] Christopher L Sansam, K Sam Wells, and Ronald B Emeson. Modulation of RNA editing by functional nucleolar sequestration of ADAR2. *Proceedings of the National Academy of Sciences*, 100(24):14018–14023, 2003.
- [256] Joana MP Desterro, Liam P Keegan, Miguel Lafarga, Maria Teresa Berciano, Mary O’Connell, and Maria Carmo-Fonseca. Dynamic association of RNA-editing enzymes with the nucleolus. *Journal of cell science*, 116(9):1805–1818, 2003.
- [257] Aamira Tariq, Wojciech Garncarz, Cornelia Handl, Ales Balik, Oliver Pusch, and Michael F Jantsch. RNA-interacting proteins act as site-specific repressors of ADAR2-mediated RNA editing and fluctuate upon neuronal stimulation. *Nucleic acids research*, 41(4):2581–2593, 2012.
- [258] Raghuvaran Shanmugam, Fan Zhang, Harini Srinivasan, John Lalith Charles Richard, Kaiwen I Liu, Xiujun Zhang, Cheok Wei A Woo, Zi Hao M Chua, Jan Paul Buschdorf, Michael J Meaney, et al. SRSF9 selectively re-

- presses ADAR2-mediated editing of brain-specific sites in primates. *Nucleic acids research*, 46(14):7379–7395, 2018.
- [259] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.
- [260] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.
- [261] Gokul Ramaswami and Jin Billy Li. RADAR: a rigorously annotated database of A-to-I RNA editing. *Nucleic acids research*, 42(D1):D109–D113, 2013.
- [262] Eric L Van Nostrand, Gabriel A Pratt, Alexander A Shishkin, Chelsea Gelboin-Burkhart, Mark Y Fang, Balaji Sundararaman, Steven M Blue, Thai B Nguyen, Christine Surka, Keri Elkins, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, 13(6):508, 2016.
- [263] Antoine Cléry, Rahul Sinha, Olga Anczuków, Anna Corrionero, Ahmed Moursy, Gerrit M Daubner, Juan Valcárcel, Adrian R Krainer, and Frédéric H-T Allain. Isolated pseudo-RNA-recognition motifs of SR proteins can regulate splicing using a noncanonical mode of RNA recognition. *Proceedings of the National Academy of Sciences*, 110(30):E2802–E2811, 2013.
- [264] Debashish Ray, Hilal Kazan, Kate B Cook, Matthew T Weirauch, Hamed S Najafabadi, Xiao Li, Serge Gueroussov, Mihai Albu, Hong Zheng, Ally Yang, et al. A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172, 2013.
- [265] Oleg Raitskin, Dan-Sung C Cho, Joseph Sperling, Kazuko Nishikura, and Ruth Sperling. RNA editing activity is associated with splicing factors in hnRNP particles: The nuclear pre-mRNA processing machinery. *Proceedings of the National Academy of Sciences*, 98(12):6571–6576, 2001.