UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Investigating roles of RNA editing in virus host interaction

by Shuquan Su

Thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

under the supervision of Prof. Jinyan Li
Prof Gyorgy Hutvagner

University of Technology Sydney
Faculty of Engineering and Information Technology

August 2024

Certificate of Original Authorship

I, ***Shuquan Su***, declare that this thesis is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Computer Science, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: August 20th, 2024

Acknowledgements

I would like to express my deepest appreciation to all those who provided me the possibility to complete this thesis. A special gratitude I give to my supervisors, Prof. Jinyan Li and Prof. Gyorgy Hutvagner, whose contribution in stimulating suggestions and encouragement, helped me to coordinate my research project especially in writing this thesis. Besides, I also thank China Scholarship Council Scholarship (CSC scholarship), Sydney Consulate General of China, and also Graduate Research School (GRS) of UTS to financially support my PhD study.

My grateful thanks are also extended to all the colleagues I work with, especially Dr. Tian Lan, Mr. Pengyao Ping of Prof. Jinyan Li's research group in UTS, and Dr. Eileen McGowan, Dr. Pattarasiri (Eve) Rangsrikitphoti, Ms. Sara Terer, Mr. Raul Alvarado Bachmann, Dr. Nattamolphan (Bank) Wittayavimol, and Ms. Anila Hashmi from Prof. Gyorgy Hutvagner's research group.

A special thanks goes to my teammates and friends, Dr. Zhongran (Edward) Ni, Ms. Tao Xie, Ms Yi (Zoe) Su, Mr. Yue (Daniel) Ma, and Dr. Weiye Liang whose insight and expertise greatly enhanced my research. I sincerely appreciate all the great help and supports from my best friend Dr. Zhongran (Edward) Ni who had been working with me in various projects throughout these years.

Last but not the least, I would like to thank my family: my parents Mr. Jiapei Su, Mrs. Haiyan Chen and my sibling Mr. Shuhui Su for supporting me spiritually throughout my Ph.D. journey and my life in general. I would also like to extend my heartfelt thanks to my partner, Ms. Zhe (Stella) Su, for her unwavering support, patience, and encouragement throughout the duration of my PhD studies and the process of writing this thesis.

Impact of the COVID-19 Pandemic on this Thesis

In January 2020, I returned to China to visit my parents and renew my visa before beginning my PhD in Australia. At the same time, SARS-CoV-2 was emerging from its initial outbreak in Wuhan, spreading rapidly across China and beyond. Within weeks, the pandemic had escalated into a global crisis, leading to widespread international travel restrictions. The Australian government implemented a strict travel ban, disrupting my original plan to start my PhD. This unexpected long-distance arrangement presented significant challenges because all my skills in scientific research are laboratory-based. Not having lab access meant I could not perform any kinds of research experiments and gather any kinds of lab-based research data. More importantly, the absence of direct supervision and face-to-face collaboration with peers and faculty further compounded the difficulty.

To adapt, I had to embrace remote learning methods, focusing on developing computational skills and bioinformatics knowledge from a complete zero background. This required not only learning programming skills from scratch but also finding innovative ways to analyse data without traditional lab-based resources. The steep learning curve demanded self-discipline and creativity in problem-solving, pushing me to leverage online resources, remote communication tools, and virtual collaborations. The travel ban persisted longer than anticipated, extending until December 2021. This prolonged separation had a profound impact on my PhD progress and research direction. By the time I could finally return to Australia, significant changes had occurred within my university. The supervisor panel had shifted, leading to a complete reorganisation of my research project. This transition disrupted my research momentum and necessitated rebuilding relationships and re-establishing my PhD focus.

Overall, the COVID-19 pandemic significantly impacted my PhD journey, requiring resilience, flexibility, and adaptability in the face of unprecedented challenges. Despite these obstacles, I have persisted in my studies and look forward to completing my PhD with a renewed sense of determination and a broader skill set.

List of Participated Publications

Su S, Ni Z, Lan T, Ping P, Tang J, Hutvagner G*, Li J*. Viral host range and path shifting are predictable through tree-based learning on codon usage biases and genomic characteristics. *Scientific Reports*. (2025). (Manuscript under-review)

Lan T, **Su S**, Ping P, Hutvagner G, Liu T, Pan Y, Li J*. Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning. *Nature Machine Intelligence*. (2024). 1-11. https://doi.org/10.1038/s42256-024-00803-z

Ping P, **Su S**, Cai X, Lan T, Zhang X, Peng H, Pan Y, Liu W, Li J*. Turn 'noise' to signal: accurately rectify millions of erroneous short reads through graph learning on edit distances. *Genomics, Proteomics & Bioinformatics*. (2025). (Manuscript accepted)

Lan T, **Su S**, Ping P, Li J*. Novel design for multi-epitope vaccines of COVID-19 and critical in-silico assessment steps. *Current Bioinformatics*. (2024). DOI: 10.2174/0115748936303461240827061629

Ping P, Lan T, **Su S**, Li J*. How error correction affects PCR deduplication: a survey based on UMI datasets of short reads. *Quantitative Biology*. (2025). (Manuscript under-review)

Liu M, Liang W*, Su Y, Qi J, Wen Y, Wang L, **Su S**, Zhao J, Shan J, Wang J*. COL8A1 is a potential prognostic biomarker associated with migration, proliferation, and tumor microenvironment in glioma. *Experimental Cell Research*. (2024). 439(1):114076. 10.1016/j.yexcr.2024.114076.

Table of Contents

List of Figures

List of Tables

List of Appendix

List of Abbreviations

[A]

aaRS	Amino-acyl tRNA synthetase
ACE2	Angiotensin-converting enzyme 2
ACE2tr	ACE2 truncated form
ACTB	Actin beta
ADAR	Adenosine deaminase on RNA
ADAR1	Adenosine deaminase on RNA 1
ADAR1L	Adenosine deaminase on RNA 1 reference isoform
ADAR1S	Adenosine deaminase on RNA 1 truncated isoform
ADAR2	Adenosine deaminase on RNA 2
ADAR3	Adenosine deaminase on RNA 3
ADAT	Adenosine deaminase on tRNA
ADAT1	Adenosine deaminase on tRNA 1
ADAT2	Adenosine deaminase on tRNA 2
ADAT3	Adenosine deaminase on tRNA 3
AGRF	Australian Genome Research Facility
AGS	Aicardi-Goutieres Syndrome
ASL	Anti-codon stem-loop of tRNA

[B]

BA	Balanced accuracy
BCA assay	Bicinchoninic acid assay
BH adjustment	Benjamini-Hochberg adjustment

[C]

CAG promoter	Chicken beta-actin promoter

CAI	Codon Adaptation Index
Cas9	CRISPR associated protein 9
CDS	Coding sequence
CMV	Cytomegalovirus
COVID	Coronavirus disease
CRISPR	Clustered regularly interspaced short palindromic repeats
CSL	Codon stretch length
CUBE	Cytometry Utilities Box Expansion

[D]

DBSCAN	Density-based spatial clustering of applications with noise
DMEM	Dulbecco's Modified Eagle medium
DSH	Dyschromatosis Symmetrica Hereditaria
dsRBD	Double-stranded RNA binding domain
dsRNA	Double-stranded RNA
DT	Decision Tree classifier

[E]

EBOV	Ebola virus
EBV	Epstein–Barr virus
eCFP	Enhanced cyan fluorescent protein
EDTA	Ethylenediaminetetraacetic acid
eGFP	Enhanced green fluorescent protein
eYFP	Enhanced yellow fluorescent protein

[F]

FBS	Fetal bovine serum
FKPM	Fragments per kilobase of transcript per million
FRET	Fluorescence resonance energy transfer

[G]

GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
Gau	Gaussian process classifier
GP	Glycoprotein
GRCh38	Genome Reference Consortium Human Build 38
GTF	Gene transfer format
GtRNAdb	Genomic tRNA Database

[H]

HBsAg	Hepatitis B virus surface antigen
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HDV	Hepatitis D virus
HIV-1	Human immunodeficiency viruses 1
hTLV-2	Human T-lymphotropic virus 2
HVCF	Human virus codon fitness score

[I]

IDE	Integrated development environment
IFN	Interferon
iHPC	Interactive High-Performance Computing
IP6	Inositol hexakisphosphate
IRES	Internal ribosome entry site

[K]

KPNA2	Karyopherin Subunit Alpha 2

[L]

LCMV	Lymphocytic choriomeningitis virus
LOO	Leave-One-Out

[M]

m6A	N6-methyladenosine
MCA	MultiCodonAnalyser
MDA5	Melanoma Differentiation-Associated protein 5
MERS-CoV	Middle East respiratory syndrome coronavirus
MPV	Metapneumovirus
mRFP1	Monomeric red fluorescent protein 1
MuV	Mumps virus
MV	Measles virus

[N]

NCBI	National Center for Biotechnology Information
NES	Nuclear export signal
NGS	Next-Generation Sequencing
NLP	Natural language processing
NLS	Nuclear localisation signal
NN	Neural Network
NoV	Norovirus
NZP	Non-zeroes percentages

[O]

ORF	Open reading frame

[P]

PBS	Phosphate buffered saline
PCA	Principal Component Analysis
PDB	Protein data bank
PIV	Parainfluenza
PKR	Protein kinase R
PPI	Protein-protein interactions

PPIA	Peptidylprolyl isomerase A
PV	Polyoma virus

[R]

RANSAC	Random sample consensus
RefSeq	Reference sequence
RF	Random Forest classifier
RIPA	Radioimmunoprecipitation assay
RNAseq	RNA sequencing
ROC curve	Receiver operating characteristic curve
ROC-AUC	Area under the ROC curve
RSCU	Relative Synonymous Codon Usage
RSMCU	Relative Synonymous Multi-Codon Usage

[S]

SARS-CoV	Severe Acute Respiratory Syndrome-related Coronavirus
SARS-CoV-2	Severe Acute Respiratory Syndrome-related Coronavirus 2
SCC	Spearman correlation coefficient
SDHA	Succinate dehydrogenase complex flavoprotein subunit A
SDM	Site-directed mutagenesis
ssRNA	Single-stranded RNA
STAR	Spliced Transcripts Alignment to a Reference
SVC	Support Vector classifier

[T]

TAD1	tRNA-specific adenosine deaminase 1
tAI	tRNA Adaptation Index
TMPRSS2	Transmembrane Serine Protease 2
TMRPSS2tr	TMRPSS2 truncated form
TNFR	Tumor necrosis factor receptor

TRMT5	tRNA methyltransferase 5
TRN1	Transportin 1

[U]

UMAP	Uniform Manifold Approximation and Projection
UTR	Untranslated region

[V]

VCF	Virus codon fitness
VV	Vaccinia virus

[X]

XPO1	Exportin 1
XPO5	Exportin 5
yW	Wybutosine

[Z]

Zα	z-DNA binding domains α
Zβ	z-DNA binding domains β
ZIKV	Zika virus

Abstract

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, which is believed to have originated from bats, emerged as a global health crisis in late 2019. The virus rapidly disseminated across the globe, resulting in widespread illness, significant mortality, and unprecedented societal disruption. According to the WHO there have been over 775 million reported cases of infection and over 7 million cumulative deaths attributed to COVID-19. Fortunately, the pandemic has come under control thanks to the dedicated efforts of medical professionals and innovative scientists. A substantial amount of virus genome sequence data has been generated, facilitated by advancements in sequencing technology, particularly through the use of on-site sequencers that enable the sequencing of virus genomes from a wide range of sources. Consequently, it is crucial to study the relationship between virus genome sequences and their host ranges, as this knowledge can provide essential insights for preventing and anticipating zoonotic transmission of novel viruses.

The virus infection cycle within host cells is well-characterised and comprises several major steps: virus entry, viral gene expression, viral genome replication, virion assembly, and release of new virions, among others. These steps exhibit variability based on the inherent nature and sequences of various viral genomes, as the efficiency of virus infection is closely linked to the interactions among diverse viral proteins. The viral proteins are synthesised based on virus genome sequences and their interplay with various host cell regulatory mechanisms. Therefore, studying the relationships between virus genome sequences and the steps of the virus infection cycle can enhance our understanding and prediction of virus host adaptation. In this thesis, particular emphasis will be placed on viral gene expression and virus entry.

Regarding host regulation and its potential impact on virus adaptation, RNA editing represents a crucial regulatory mechanism. RNA editing is a post-transcriptional molecular process that alters the nucleotide sequence of an RNA molecule, resulting in divergence between the RNA sequence and its corresponding DNA template. These modifications can profoundly affect RNA function, including alterations in mRNA and tRNA, leading to modified amino acid

sequences and inefficient protein translation. Thus, studying the roles of RNA editing in virus infections, particularly in viral gene expression and virus entry, can enhance our understanding of host immune responses and virus host adaptation. One of the most well-studied forms of RNA editing is adenosine-to-inosine (A-to-I) editing. This specific type of editing is catalysed by the enzyme family known as adenosine deaminases acting on RNA (ADAR). During A-to-I editing, an adenosine (A) nucleotide is deaminated to form inosine (I), which is recognised as guanosine (G) by the cellular machinery during translation. Another recently discovered A-to-I editing enzyme is adenosine deaminases acting on tRNA (ADAT), which converts adenosine (A) to inosine (I) on tRNA anti-codons, thereby expanding the tRNA pairing codons. In the initial part of this thesis, I summarise all findings to date regarding RNA editing in the context of virus infections as literature review sections.

Virus codon usage biases within viral genome sequences are a major determinant of viral gene expression, as viral gene expression relies on the host's translational machinery to produce viral proteins. In this thesis, I employed statistical analysis to compare the Relative Synonymous Codon Usage (RSCU), an important metric for codon usage biases, between human and non-human viruses. This analysis identified codons with significantly different usage biases in human viruses. Correlation analysis revealed that A/U-rich codons are more abundant in human viruses, suggesting that human viruses are naturally more susceptible to ADAR editing. Additionally, statistical comparisons indicated that human viruses are more enriched in ADAT-related codons, highlighting the potential roles of ADAT editing in virus infections.

Subsequently, I evaluated the virus codon fitness (VCF) in human hosts by training a Random Forest (RF) model to predict host labels based on virus RSCU matrixes and other genomic features. Utilising RSCU to predict host labels yielded high accuracy (~0.8) and demonstrated high reliability with >0.6 accuracy when training the model with only 5% of the data. The prediction accuracy further improved when additional features, such as Taxonomy data and CDS length, were incorporated. To further assess the reliability of this prediction, I employed the Leave-One-Out (LOO) method to predict the host of one virus using a model trained on all other viruses. This approach successfully predicted the human host label for all historically dangerous pandemic-causing viruses. The prediction probability of the trained model serves as the readout of the human virus codon fitness score (HVCF), which is used to evaluate virus

codon fitness in human hosts. I also applied the HVCF method to various scenarios, including evaluating HVCF from virus genomes sequenced from human and non-human sources, and assessing the HVCF of SARS-CoV-2 genomes sequenced during the COVID-19 pandemic. These applications provided insights into the codon adaptations of different viruses for different reality purposes. The findings suggest that it is practical to use machine learning methods to evaluate codon fitness in hosts, ultimately contributing to the prediction of virus hosts for newly emerged viruses based on sequenced virus genomes.

Another critical step in the virus infection cycle, virus entry, fundamentally relies on protein-protein interactions (PPI) between viral Spike proteins and host receptors. For instance, the entry of SARS-CoV-2 into host cells requires PPI between its Spike protein and human ACE2 and TMPRSS2. Until now, many deep learning models can evaluate PPI based on protein sequences. However, methods to efficiently examine PPI changes caused by mutations are lacking for proving the PPI prediction. Here, I employed fluorescence resonance energy transfer (FRET) to examine PPI between proteins of interest fused with fluorescence reporter tags. Plasmids were constructed for intracellular transfection to express these fusion proteins. The fluorescence profiles detected using flow cytometry were converted into single-cell FRET efficiency using a previously established bioinformatics tool. The FRET-based PPI detection assay successfully detected PPI between SARS-CoV-2 Spike protein and human ACE2 and TMPRSS2. Additionally, RNA editing events on ACE2 and TMPRSS2 were identified using next-generation RNA sequencing from samples extracted from cells expressing ADAR, ACE2, and TMPRSS2. Although I planned to mutate the fusion protein plasmids according to detect the alterations in FRET efficiency caused by RNA editing events, I was unable to complete this task.

In this thesis, I have successfully investigated the codon characteristics of human viruses and evaluated their codon fitness in the human host using a machine learning model. Additionally, I established a FRET-based PPI detection assay to assess interactions between virus Spike proteins and human receptors. These findings advance our understanding of viral gene expression and virus entry mechanisms, thereby enhancing our ability to predict the host range of unknown viruses.

CHAPTER 1

Literature Review of RNA Editing in Virus Infections

by Shuquan (Steve) Su

Keywords:

Literature review, RNA editing, Virus infections, Adenosine deaminase

Chapter 1. Literature Review of RNA Editing in Virus Infections

1.1. Introduction

Virus mutation is a natural process whereby changes occur in the genetic material of a virus. These changes can happen through various mechanisms, such as errors during replication, recombination, or through the influence of external factors. Mutations can lead to changes in the viral characteristics, affecting its transmissibility, virulence, and resistance to vaccines and treatments. While most mutations are neutral or even detrimental to the virus, occasionally, a mutation can provide a competitive advantage, allowing the virus to spread more effectively or evade the immune response. This evolutionary process is a critical aspect of virology, influencing disease outbreak patterns and the development of effective vaccines and therapeutic strategies.

Surprisingly, hosts can also induce viral mutations through specific cellular mechanisms, one of which is RNA editing, a process that can lead to virus mutations. RNA editing is a post-transcriptional molecular process through which the nucleotide sequence of an RNA molecule is altered, leading to a difference between the RNA sequence and its corresponding DNA template. The modifications can have profound effects on the RNA's function, including alterations in the coding potential of messenger RNAs (mRNAs), which can lead to the production of proteins with modified amino acid sequences. RNA editing has great impacts on expanding the genetic diversity of RNA molecules.

Before going into research chapters, this chapter will earnestly review research discoveries and literatures in this research field of RNA editing (A-to-I editing to be specific) and its impacts on virus infections. The detailed literature review will help understand the field of A-to-I editing in virus infections, which leads to my interests in this field then proposing potential knowledge gaps.

1.2. Deaminase: an enzyme that creates special mutations

1.2.1. History of adenosine deaminase discovery

The major type of RNA editing, A-to-I editing, as a primary form of RNA modification, is one of the major contributors of nucleotides variations for either host or virus. Inosine (I), recognised as a nucleotide, was initially identified in the wobble position of tRNA by Francis Crick, co-discoverer of the DNA double helix, in 1966. This finding revealed that inosine functionally mimics guanosine (G) in Watson-Crick base pairing[5]. The enzyme responsible for inosine formation, ADAR (Adenosine Deaminase Acting on RNA), was not discovered until 1987 through studies on Xenopus laevis embryos. These studies linked ADAR to activities involved in unwinding double-stranded RNA, and it was later identified that ADAR catalyses the conversion of adenosine to inosine via deamination, a process known as A-to-I editing[6-8]. A-to-I editing is the most abundant form of found RNA editing event in Metazoans[9, 10], with humans exhibiting the highest levels of this editing, particularly in non-coding regions, among primates[11]. This widespread occurrence underscores the significant biological functions and evolutionary implications of A-to-I RNA editing.

The human genes responsible for the biochemical reaction of adenosine deamination were subsequently named Adenosine Deaminase Acting on RNA, or ADAR. The human genes responsible for the biochemical reaction of adenosine deamination were subsequently named Adenosine Deaminase Acting on RNA, or ADAR. ADAR genes are extensively distributed across a broad range of organisms, excluding protozoa, yeast, and plants[12], and display high conservation across different species[13]. Notably, ADAR3, another member within the ADAR family, was identified in 1996 by examining sequences homologous to the deaminase domain of ADAR1. ADAR3 is found as homolog to rat RED2 and exclusively expressed in the human brain[14].

The genes responsible for adenosine deamination at the wobble position of tRNA were later identified in 1999 and are analogous to the yeast protein tRNA-specific adenosine deaminase 1 (TAD1)[15]. These genes are later designated as Adenosine Deaminase Acting on tRNA, or ADAT, highlighting their unique characteristics and functional divergence from ADAR genes.

Moreover, deaminase on tRNA is believed to evolve from cytidine deaminases (CDAs) acting on mono-nucleotides, thus ADAT gene is technically the evolutionary ancestors of ADAR[16]. Despite its earlier discovery, ADAT has garnered limited attention in the field of RNA editing research, particularly in virological studies, compared to ADAR.

1.2.2. Different members in ADAR and ADAT gene family

To date, there are three members in the ADAR gene family ADAR1, ADAR2, and ADAR3, and three ADAT genes ADAT1, ADAT2, and ADAT3. Various transcripts of ADAR1 and ADAR2 have been recognised, arising from diverse mechanisms such as different promoters, alternative splicing, and varying ATG start codons [17-20]. According to the NCBI database, comprehensive gene information for both the ADAR and ADAT families is summarised in Table 1, with details on the various transcripts and isoforms summarised in Appendix 1. Research indicates that ADAR1 and ADAR2 are highly conserved with distinct functions throughout evolutionary history, suggesting the ADAR gene family are vital for various organisms[21].

Table 1. Summarised general information of ADAR and ADAT genes from NCBI Gene database.

Gene	NCBI Gene ID	Location	Exon count	Chromosome	NCBI Annotation location (GRCh38.p14)	Verified transcripts count	Reference transcript accession ID
ADAR1	103	1q21.3	20	1	NC_000001.11 (154582057..154627997, complement)	10	NM_001111.5
ADAR2	104	21q22.3	21	21	NC_000021.9 (45074578..45226563)	7	NM_001112.4
ADAR3	105	10p15.3	10	10	NC_000010.11 (1177313..1737525, complement)	1	NM_018702.4
ADAT1	23536	16q23.1	12	16	NC_000016.10 (75596868..75623281, complement)	10	NM_012091.5
ADAT2	134637	6q24.2	10	6	NC_000006.12 (143422832..143450695, complement)	2	NM_182503.3
ADAT3	113179	19p13.3	2	19	NC_000019.10 (1905399..1913447)	2	NM_138422.4

The ADAR1 gene expresses two primary isoforms: a longer isoform sized 150 kDa (referred to as the reference isoform of ADAR) and a shorter, truncated isoform sized 110 kDa, designated as ADAR1L (or ADAR1-p150) and ADAR1S (or ADAR1-p110), respectively[17, 20, 22-26]. These isoforms are encoded by the same ADAR1 gene but differ in their translational start sites (ATG) and promoter sequences, leading to their distinct

functionalities[17, 20, 22, 23]. Notably, compared to ADAR1S, ADAR1L has an additional z-DNA binding domain, which may lead to distinct substrate selections. These differences cause the distinct biological functions between ADAR1L and ADAR1S, where ADAR1L is believed to be mainly interferon-inducible while ADAR1S is mostly expressed constitutively and vital in human developmental process[20, 22-26]. Although ADAR1L can be significantly induced by interferon (IFN), it also maintains a level of basal expression[20, 22, 24-26]. The interferon-inducibility of ADAR1L renders it particularly significant during viral infections, a topic that will be explored in greater detail later in this thesis.

In contrast to the ADAR1 gene, only a single dominant isoform has been identified for both ADAR2 and ADAR3, as is the case for all ADAT genes[27]. Although multiple transcripts of ADAR2 have been detected, only one is the predominant[19, 27, 28].

1.2.3. The general domain structures characterised in ADAR and ADAT genes

For a comprehensive understanding of the structural attributes and biological roles of ADARs and ADATs, Figure 1 illustrates their domain compositions. ADAR1L, the primary predominant isoform of ADAR1, is equipped with two z-DNA binding domains, three double-stranded RNA binding domains (dsRBDs), and a deaminase (or A-to-I editase) domain. As previously noted, ADAR1S is differentiated from ADAR1L by the absence of complete z-DNA binding domains yet retains identical configurations for all other domains. ADAR2 and ADAR3 exhibit similar domain compositions, each featuring two dsRBDs and a single deaminase domain. Uniquely, ADAR3 incorporates an additional single-stranded RNA binding domain, known as the R-domain, which is not present in other ADAR genes. Compared to ADAR family, the domains composition of members in ADAT family are simpler, which only one deaminase is found respectively in ADAT1, ADAT2, and ADAT3. In general, the critical functional domain in both ADARs and ADATs is the deaminase domain, which conducting deamination reaction on RNA molecules as their major biological function.

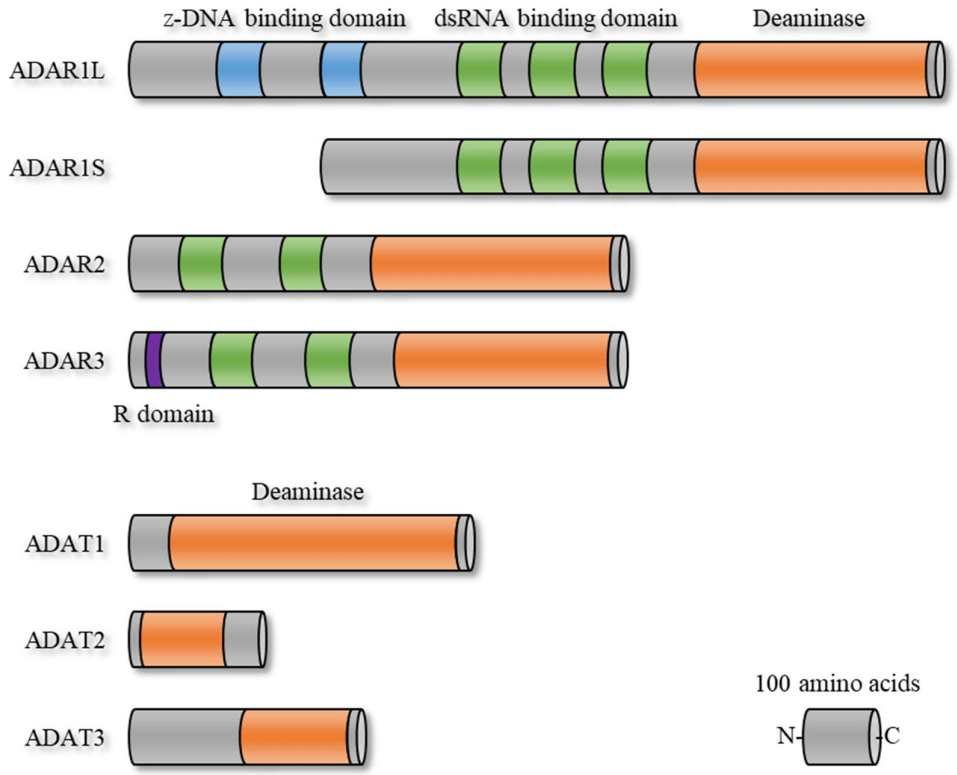


Figure 1. Domain structure features of human ADAR and ADAT gene families. ADAR family contains three members: ADAR1, ADAR2 and ADAR3. Two ADAR1 isoforms, generated by alternative start sites, are studied in this project because both play important roles in host biology, and they are ADAR1L (or ADAR1-p150) and ADAR1S (or ADAR1-p110). ADAT family contains three members: ADAT1, ADAT2 and ADAT3.

1.2.4. The biochemistry conducting to A-to-I editing: Adenosine deamination

As previously indicated, the deaminase domains of ADARs and ADATs are evolutionarily homologous. Consequently, it is not surprising that the biochemical reactions of converting adenosine into inosine, or deamination reactions, between ADARs and ADATs are basically identical. The deaminase domain, or A-to-I editase, of both ADARs and ADATs catalyses the deamination of an adenosine by replacing the amine group ($-NH_2$) on the sixth carbon (C_6) with a carbonyl group (=O), thereby transforming adenosine into inosine (Figure 2). Although the deaminase domain of ADAR3 closely resembles that of ADAR2, with 50% sequence identity and a similar targeted substrate pool, the ADAR3 deaminase domain is nonetheless

catalytically inactive[14, 29, 30]. The molecular characteristics behind ADAR3 deaminase will be discussed later.

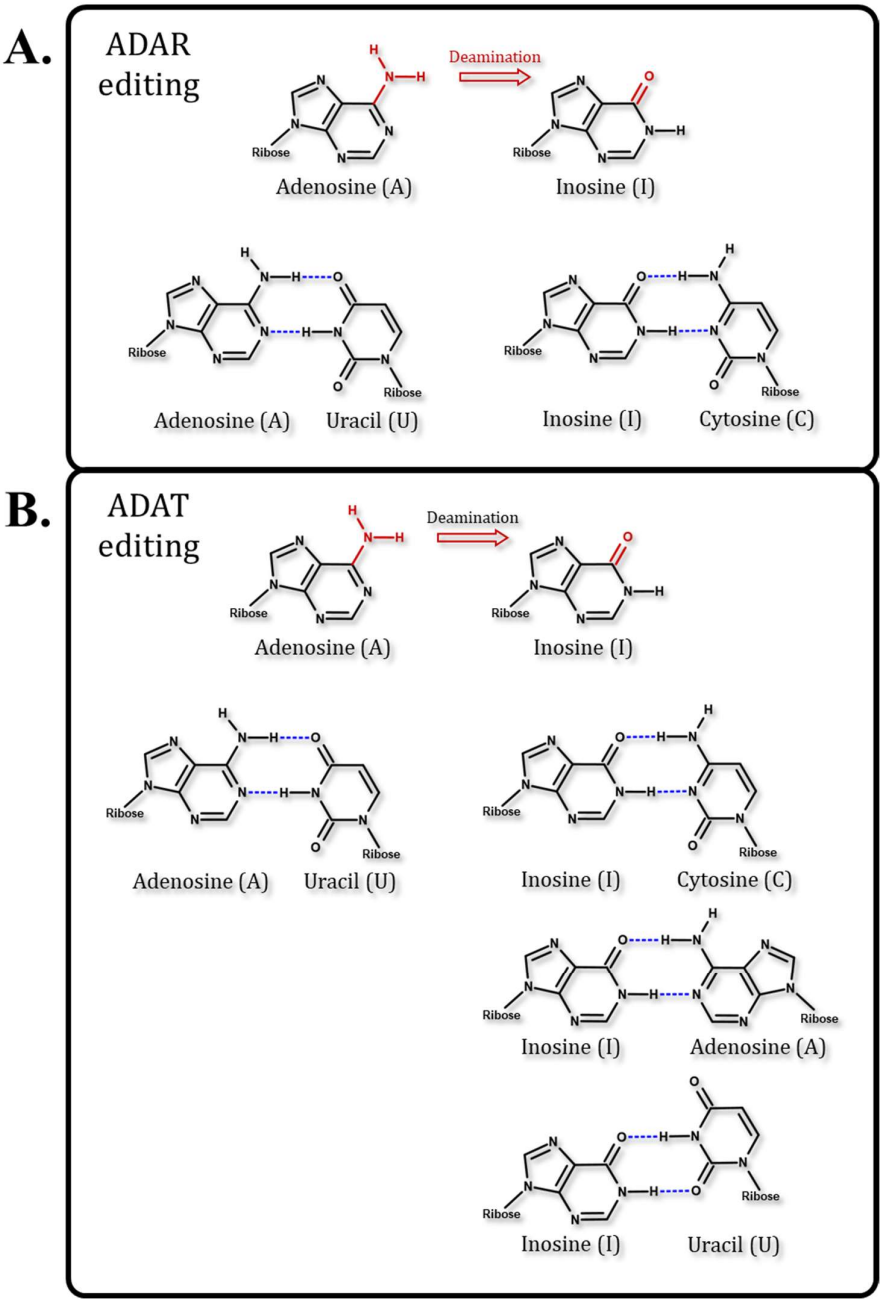


Figure 2. Biochemical reaction of A-to-I editing and pairing of post-edited inosines. (A) A-to-I editing of ADAR genes. Before ADAR editing, the adenosine/A is paired up with uridine/U, whereas the inosine/I is paired up with the cytidine/C after ADAR editing. (B) A-to-I editing of ADAT genes. Before ADAT editing, the adenosine/A is paired up with uridine/U on tRNA anticodon, whereas the inosine/I is paired up with the cytidine/C, adenosine/A, and uridine/U after ADAT editing.

Moreover, ADARs function independently of accessory proteins or RNA factors for the deamination reaction, which purified ADAR proteins and substrate are found sufficient for A-to-I conversion[19, 31-33]. Although the chemical reaction is nearly identical for both ADARs and ADATs, the biological implications differ significantly because different types of RNAs are edited. ADARs are capable to edit most RNA molecules with double-stranded structures like mRNA, pre-mRNA, miRNA et al. Inosine within these RNAs typically pairs with cytosine (C), resulting in I-C pairings that functionally mimic guanosine (G) in G-C pairings[5-8].

Conversely, the inosines on tRNA edited by ADATs, particularly edited by ADAT2 and ADAT3, are more flexible in pairing and found capable to pair up with cytosine (C), adenosine (A) and uracil (U) in translation process, thus forming either I-C, I-A, or I-U pairing[34-36] (Figure 2.B). Interestingly, while both ADAR and ADAT editing convert adenosine to inosine, their resulting base-pairing options differ, likely due to the high flexibility of the third nucleotide in the tRNA anticodon. The distinctions in the editing substrates and the editing outcomes indicate the distinct molecular biology of ADARs and ADATs, which will be explored further in subsequent sections dedicated to the molecular biology of these enzymes.

1.3. In-depth introduction of ADAR gene domain structures and their functionalities

To enhance our understanding of the functionality of RNA editing enzymes such as ADARs and ADATs, it is crucial to delve deeper into their molecular structures and characteristics. By gaining more detailed insights into their molecular configurations, we can better comprehend how these enzymes are regulated and how they edit their substrates. This knowledge is vital for exploring the potential roles that ADARs and ADATs may play in viral infections. Such insights are instrumental in unravelling the complex interactions between these RNA editing enzymes and their substrates, which could ultimately influence the pathogenesis of viral diseases.

1.3.1. The double-stranded RNA binding domains of ADAR gene and their recruitment of dsRNA substrates

Double-stranded RNA binding domains (dsRBDs) are crucial components of ADAR enzymes, as they enable the capture of substrate RNA for A-to-I editing, significantly influencing the selection of RNA substrates by ADAR[37-39]. Typically, approximately 65 amino acids in length, dsRBDs exhibit an α-β-β-β-α folding configuration, similar to dsRBDs found in other molecules[39]. The presence of bulges and loops within the RNA substrates makes them particularly attractive to dsRBDs, markedly affecting the specificity of ADAR across different RNA substrates[40]. Once an ADAR enzyme binds to a dsRNA substrate, the proximity of the dsRBD-targeted internal loop structures to the editing site plays a critical role in determining editing efficiency. However, this proximity does not influence the selectivity of the editing site. The choice of specific adenosine targets for deamination is also determined by other factors such as the nucleotide mismatches of RNA substrate[41]. This interplay between structural attraction and sequence specificity underlines the complex mechanisms by which ADAR enzymes select and modify their RNA substrates.

Both ADAR1L and ADAR1S have three dsRBDs (dsRBD1, dsRBD2, dsRBD3), whereas ADAR2 and ADAR3 each have only two (dsRBD1, dsRBD2)[39]. The slight variances within

coding sequence of each dsRBDs and the distance in between different dsRBDs contribute to the differences in substrate preferences. Additionally, the high structural similarities in dsRBDs between ADAR2 and ADAR3 may the reason to similarities in their substrate selection, potentially resulting in competitive substrate binding between these two. Functionally, the three dsRBDs in ADAR1 exhibit distinct characteristics, each providing varying levels of dsRNA binding affinity. Notably, dsRBD3 is considered the most crucial for binding dsRNA substrates, while dsRBD2 is the least significant[42]. According to previous studies, the molecular mechanism of dsRBDs-dsRNA binding is well-studied with researches to ADAR2[4]. Although ADAR1 may be very different than ADAR2 where ADAR2 has two dsRBDs but ADAR1 has three dsRBDs and additional z-DNA binding domain[4], this knowledge from ADAR2 study helps to understand how dsRBDs recruit and interact dsRNA substrates.

Once ADARs recognize dsRNA substrates through binding affinity to the favoured spatial structures of dsRNA, the dsRBDs initiate contact and binding with the dsRNA. This interaction involves the dsRBDs contacting adjacent two minor grooves and intervening in the major groove between them[39]. The distance between ADAR2 dsRBD1 and dsRBD2 is 93 amino acids, but it is only 22 amino acids between dsRBD2 and the deaminase domain[4]. It semes that the dsRBD2 can bind a portion of the 15bp duplex and provide sufficient binding affinity to facilitate the distortion of the RNA backbone and base flipping required for editing, because dsRBD2 is too close to deaminase domain and it is difficult to allow dsRBD2 and deaminase bind the same location concurrently[4]. However, the binding of the deaminase domain causes the RNA to distort, particularly widening and expanding the 5’ major groove[43]. Then, the dsRBDs bind across expanded major grooves[44-46], and the dsRBD2 can bind either the face of editing site or opposite face of the editing site[4]. The binding of either dsRBD2 or deaminase domain can remodel or reshape the dsRNA conformation and may facilitate binding of the other component[4]. The presence of defects in the RNA duplex, such as mismatches, enhances the binding affinity of the deaminase domain, but these mismatches can sometimes interfere with the simultaneous binding of dsRBD2 and the deaminase domain[4]. Additionally, 6 amino acids were found crucial for ADAR2 binding the substrate: F457, D469, H471, P472, R474, R477[47]. These residues likely play key roles in stabilising the interaction between the dsRBD and the RNA substrate, influencing the enzyme's editing efficiency and specificity.

Although some dsRNA substrates may have insufficient binding affinity with only catalytic deaminase domain due to their conformational features, the dsRBD binding to the adjacent duplex structure can also provide additional binding force to stabilise the binding[4]. Thus, it is not surprising that editing efficiency can be enhanced when multiple dsRBDs are employed to bind the dsRNA[48]. This multipoint binding not only increases the overall stability of the enzyme-substrate complex but also enhances the precision with which the enzyme engages its target. Furthermore, by introducing mutations into the dsRNA binding domain sequences in ADAR transcripts, the substrate selectivity of the resulting mutated ADARs can be altered. This change in substrate specificity can lead to variations in the editing target pool among different individuals, which could be driven by potential single nucleotide polymorphisms (SNPs) within the dsRBD sequences[49]. Such genetic variability can influence the functional diversity of RNA editing across populations, contributing to differences in gene expression profiles and potentially impacting phenotypic traits and susceptibility to diseases.

1.3.2. The z-DNA binding domain of ADAR gene and their contribution to substrate specificity

ADAR1 is unique among the ADAR family in that it contains two z-DNA binding domains, named Zα (closer to the N-terminus) and Zβ[50, 51]. Unlike the more common right-handed B-DNA conformation, z-DNA adopts a left-handed configuration with a zig-zag pattern along its backbone[52, 53] (Appendix 2). The Zα domain is exclusively present in ADAR1L, while the Zβ domain is found in both ADAR1L and ADAR1S, albeit slightly truncated in ADAR1S[54]. These z-DNA domains in ADAR1 share a common binding motif with other z-DNA binding domains found in various molecules, indicating a homology with other common z-DNA binding domains from other genes[50]. Interestingly, the molecular interaction mechanisms between Zα and Zβ are slightly different[55], with the Zα domain demonstrating a higher affinity for z-DNA, suggesting it plays a primary role in recruiting z-DNA[42]. ADAR1 Zα domain recognises left-handed z-DNA in a conformation-specific manner instead of a sequence-specific manner, similar to dsRBDs binding dsRNA[50, 56]. In fact, Zα domain binds to z-DNA even without base-specific contacts, which residues interact with the phosphate backbones[56]. Although Zα domain binding is not sequence-specific, some

sequence-specific features can facilitate corresponding spatial structures preferred by Zα domain[57, 58]. Once the Zα domain binds z-DNA, it is difficult to dissociate[59]. The binding of the Zα domain to z-DNA may influence transcription by altering the dynamics of mRNA folding. Due to the relatively slow dissociation of the Zα domain, more time is available for mRNA to form secondary structures, which can ultimately enhance ADAR editing[60, 61]. Additionally, the Zα domain helps position the deaminase domain closer to its substrates, further facilitating the editing process[60, 61]. The editing on DNA should not be happened due to the necessity of Hydroxyl group (-OH) within ribose in deamination reaction of ADARs[2]. Additionally, Zα domain can also bind left-handed z-RNA, which is a RNA molecules with z-conformation[60]. Surprisingly, it has been reported that the Zα domain recruits DNA/RNA hybrids more efficiently than double-stranded z-DNA[62]. However, whether such DNA/RNA hybrids are subject to editing by ADAR's deaminase domain remains unknown, and the implications of such interactions in cellular activities have yet to be elucidated[63].

Mutating individual dsRBDs can alter substrate selectivity without impacting the ability of the z-DNA binding domains to bind z-DNA, indicating that the z-DNA binding domains operate functionally independent of the dsRBDs[42]. However, it is reported that the mutations in the Zα domain that weaken its z-DNA binding capacity can also lead to a reduced dsRNA binding capacity of the dsRBDs[64]. This suggests a potential interplay or cooperative structural dynamics between the Zα domain and the dsRBDs, even though they generally recruit substrates independently[60].

1.3.3. The deaminase domain of ADAR gene responsible for adenosine deamination and editing sites specificity

After reviewing the substrates recruiting domains including dsRNA binding domains and z-DNA binding domains, here is the most crucial domain of ADARs catalytic deaminase domain, or A-to-I editase, endowing the unique functionality of A-to-I editing. The interactions between dsRBDs and dsRNA, as well as between z-DNA binding domains and double-stranded z-RNA or DNA/RNA hybrids, facilitate the recruitment of the deaminase domain closer to the target

adenosine, although deamination catalysis and substrate binding are considered independent events[2, 41, 65]. Both the binding interactions of dsRBDs and the deaminase domain, along with potential interactions involving z-DNA binding domains, are thought to provide sufficient biophysical energy for conformational changes that expose the adenosine for editing within the catalytic deaminase domain[4]. Interestingly, ADAR editing conducted by the deaminase domain can occur without the recruitment help from dsRBDs and z-DNA binding domains, especially if the substrates are particularly favourable with enough substrate binding force[48]. This highlights the inherent binding affinity and functional autonomy of the deaminase domain. The various deaminase domains encoded by different ADAR genes are homologous and share the same deamination mechanism, which has been extensively studied with ADAR2 and serves as an excellent model for understanding ADAR-dsRNA interactions[2, 4]. Editing of DNA/RNA hybrids, although not as well-studied as conventional dsRNA editing, presents an area of ongoing research[66]. However, detailed discussion of this type of editing falls outside the scope of this thesis, and they won't be discussed further.

In fact, besides binding to the target-adenosine-containing strand, ADAR also requires binding to the ribose on the complementary unedited strand opposite to the editing strand in the dsRNA helix[67]. Thus, the ADAR-dsRNA interaction is more complex than it initially appears. For better understanding of ADAR-dsRNA interaction, three important internal regions (Region 1, Region 2, and Region 3) inside the deaminase domain is highlighted of protein-RNA complex[4] (Figure 3, Appendix 3). These regions approximately covers 20 bp of dsRNA substrate along the single side of double-helix (same site as target adenosine), and includes two adjacent major groove and the minor groove where the target adenosine is located[4]. Actually, ADAR does not necessarily need to bind all three regions in the deaminase domain simultaneously, nor does it require a perfect alignment with the dsRNA structure (i.e. GluR-B Q/R and R/G editing)[4]. The primary requirement is simply to maintain sufficient structural stability to keep ADAR and the dsRNA adequately bound. This flexibility in binding allows ADAR to adapt to a variety of RNA structures and sequences, thereby enabling efficient editing activity across a diverse range of substrates.

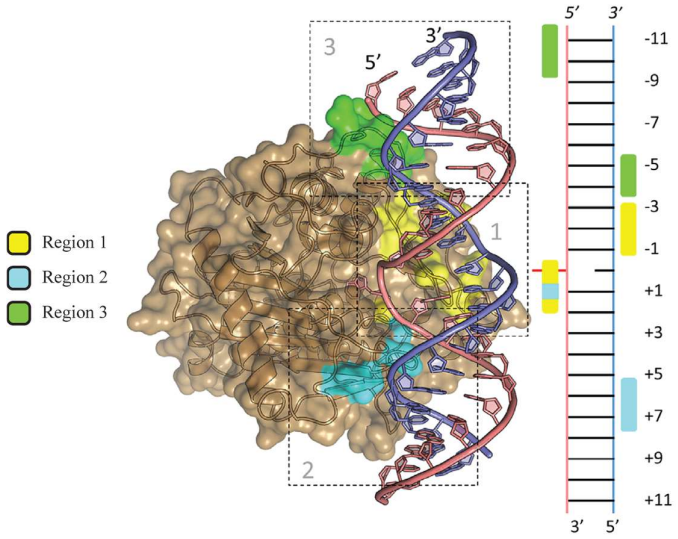


Figure 3. Overview of three critical regions in ADAR deaminase domain. The three critical regions include Region 1 (yellow), Region 2 (cyan), Region 3 (green). This figure is adapted from previous publication[4].

Region 1 encompasses the minor groove of dsRNA substrates where the edited adenosine is located, acting as the core site for the deamination reaction. The recruitment of dsRNA by the dsRBDs toward the deaminase domain allows for the residues 486-491 of the deaminase to promote the base flipping mechanism essential for editing. This 486-491 loop, by penetrating into the dsRNA helix, distorts the RNA backbone on the complementary strand, effectively setting the stage for editing[4]. The ADAR flipping loop, which is involved in this base-flipping mechanism, typically covers three base pairs, linking to a 5' U and a 3' G[2]. A crucial component of this mechanism is the residue R510, which forms hydrogen bonds with the 3′-phosphodiester of the 5'-nearest-neighbour nucleotide (typically U at the -1 position). Both ADAR1 and ADAR2 share this arginine (R510), whereas in ADAR3, this residue is replaced by an alanine, leading to ADAR3's inactivity in A-to-I editing[2]. In addition to R510, hydrogen bonds and salt bridges formed by S495 with phosphates of the (-2)-position nucleotide and R481 with phosphates of the (-3)-position nucleotides further stabilise the dsRNA distortion and facilitate the base flipping of the target adenosine[4]. Thus, the target adenosine is flipped out from the dsRNA helix and join into the zinc-containing reaction

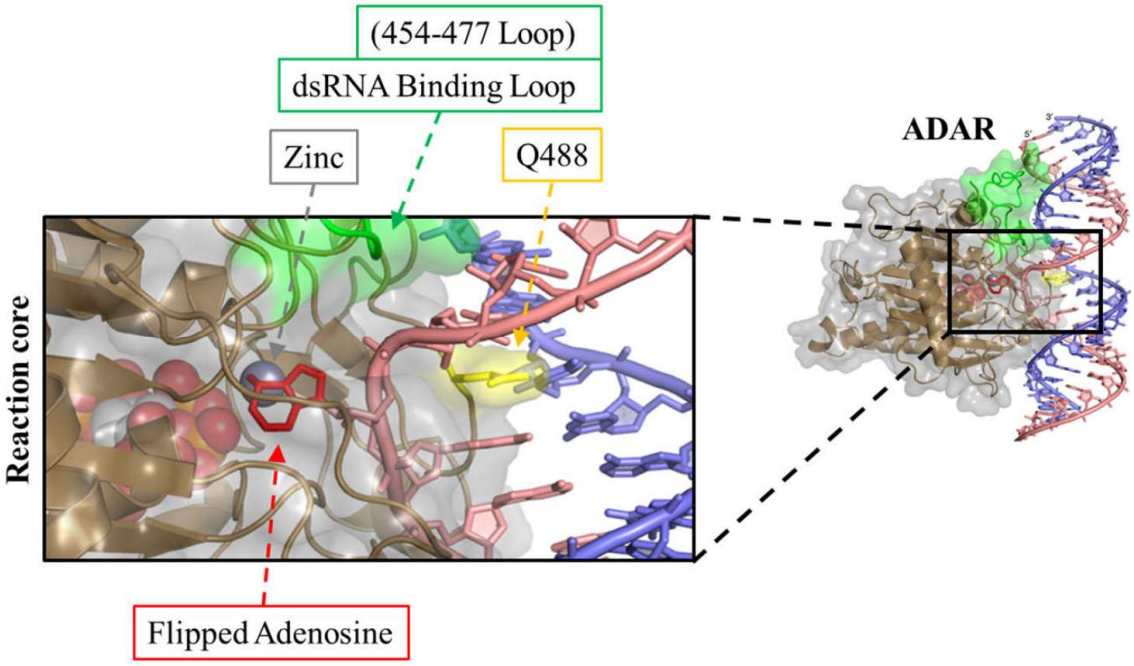


Figure 4. 3D structure of the A-to-I editase domain of ADARs and composition of its reaction centre. The 3D structure of binding interaction between the A-to-I editase domain of ADARs and its dsRNA substrate is shown on the right. Its reaction centre and the ADAR-dsRNA interaction site is zoomed in, which is shown on the left. The location of the minor groove and major groove are shown. The green structure of ADARs is the 454-477 loop. The grey sphere is a zinc ion, and the red structure beside it is the target adenosine, The yellow structure is Q488, which is inserted into the double helix structure. The R510 is hidden behind the Q488, which is outside the dsRNA double helix. This figure is adapted from previous publication[2].

core[2, 68] (Figure 4). Base-flipping allows ADAR to access C6 atom of the adenosine[69]. Within the catalytic core, a zinc ion ($Zn^{2+)}$ in the active site and it is proved essential for the deamination reaction[68, 70]. It is assisted by an inositol hexaphosphate (IP6) molecule, which contributes to the protein's folding and is essential for the enzymatic activity[70]. The deamination reaction itself is a Nucleophilic Aromatic Substitution (S_NAr-type reaction), which the zinc-bound hydroxide ion and N1 protonation generates the covalent hydrate of the adenine ring (the Messenheimer intermeidate)[68]. The proton is transferred from the hydroxyl group to the leaving amine group is followed by departure of ammonia and inosine product formation[68].

Taking a closer look to the reaction core of deaminase domain with study of ADAR2 reaction core, the reaction core is stabilised with multiple H-bonds formed between different ADAR2

residues and nucleotides on this distorted dsRNA substrate. One of the critical aspects of this interaction is the intercalation of residue 488 into the dsRNA helix at the minor groove, which occupies the space created by the flipped-out edited adenosine. This residue forms hydrogen bonds with the 2'-hydroxyl group of the opposite, orphaned nucleotide on the complementary strand and its adjacent 5' nucleotide (the -1 position)[2]. Previous research reported that ADAR prefer editing sites with A-C and A-U mismatch over A-A and A-G mismatch[71]. Only three residue-orphaned nucleotide combination were reported: E488-U, E488-C and Q488-C, which may explain why ADAR prefers editing adenosine in A-C and A-U mismatch over A-A and A-G mismatch[2]. The preference for editing at A-C and A-U mismatches is likely due to the structural incompatibility of a purine in the opposite position, which would clash with the insertion of residue 488, thereby impacting editing efficiency[41]. Interestingly, a glutamine (Q) in residue 488 results in enhancing base flipping and dramatically enhancing editing efficiency, and other mutations in Q488 will leads to decrease of editing efficiency[69]. The inability of ADAR to edit dsDNA is attributed to structural differences in the helix, the groove widths and depths to be specific, which ADAR residues are difficult to intercalate[2]. Additionally, the lack of 2'-hydroxyl groups in DNA, which are crucial for ADAR recognition, further limits its activity to RNA substrates[2]. In addition, the 2'-hydroxyl group of the edited adenosine also form H-bonds with T375, which is also crucial for editing activity[4, 72]. Other interactions were also reported in region 1. T490 and I456 also form hydrogen bonds with the 2'-hydroxyl groups of the (-2)- and (-3)-position nucleotides, respectively, on the complementary strand[4]. Additionally, residues G593, K594 and R348 also contact the complementary strand and are conserved in ADAR molecules, where mutations in such location can cause decrease in editing efficiency[2]. Overall, region 1 of the ADAR2 deaminase domain is critical for efficient deamination of most substrates, with all contacts to the target strand contributing to the stabilisation of the ADAR2 reaction core and the successful intercalation of the flipped adenosine.

Several previous articles suggested that ADAR2 molecule prefer to edit adenosine with 5' nearest-neighbour U or A and 3' nearest-neighbour G (5'-U<u>A</u>G-3')[37, 49, 69, 73]. With X-crystal structure study, this preference can be explained at a molecular level. The preference for a 5'-nearest-neighbour U or A (at the -1 position) is influenced by the spatial arrangement within the ADAR2 molecule, particularly the interaction of residue G489. This residue G489 would clash with a 5' nearest-neighbour G or C, which would destabilize the reaction structure.

Substituting a U/A with G/C at this position leads to a significant reduction in editing efficiency, by approximately 80%[2]. Furthermore, the residues 486-491 loop penetration also contributes to the 5'-nearest-neighbour-U preference[2]. Moreover, the S486 will also form H-bonds with 2'-hydroxyl group of the 3'-nearest-neighbour G (+1 position) and the +2 position nucleotide to stabilise the reaction structure[2, 4]. Replacing 3' G (+1 position) with other nucleotides will partially decrease ADAR editing efficiency[2]. This illustrates how specific nucleotide sequences adjacent to the target adenosine are crucial not just for binding affinity but also for the structural integrity and effectiveness of the editing mechanism.

Region 2 of the ADAR deaminase domain plays a critical role in stabilising the interaction with the RNA substrate by engaging with the adjacent major groove directed 3' to the editing site[4]. This deaminase region covers 20 bp of RNA, which required stable dsRNA helix (less mis-pairing)[4]. Key interactions within this region involve two specific loops located near residues R348 and K594. These loops interact with the RNA strand at the +6 and +7 positions in the complementary strand, 3' of the editing site. Notably, the charged side chain of K594 is strategically inserted into this major groove, playing a vital role in stabilising the RNA structure at these positions[4]. These detailed molecular interactions within Region 2 are integral for the proper function of ADAR, as they help to maintain the necessary structural integrity of the RNA during the editing process. This ensures that only the correct adenosines are targeted and modified, which is vital for the biological outcomes of RNA editing.

Region 3 of the ADAR deaminase domain interacts with the adjacent major groove that is 5' directed to the editing site[4]. Besides the critical residues, other residues have relatively more flexible sequences, which may contribute to the flexibility of dsRNA structure with different sequences[4]. A key structural component within this region is the 454-477 loop, which intercalates into the dsRNA helix, thereby stabilising the ADAR-dsRNA complex[2]. The intercalation of this loop not only stabilizes the complex but also introduces a kink in the dsRNA helix. This distortion can potentially affect interactions within the region 5' to the editing site, influencing the overall dynamics of RNA editing[74, 75]. The 454-477 loop is disordered before binding dsRNA substrate and only functional after dsRNA substrates are recruited[70]. In terms of molecular interactions, Region 3 forms hydrogen bonds and salt bridges with the phosphate backbone of both strands, spanning from -4 to -11 base pairs relative to the editing site. This region consists of residues 454 to 479, and some residues are highly

conserved across different species, reflecting their functional importance. The sequence between residues R470 and Q479, in particular, shows higher similarity across species, indicating its critical role in the enzyme's activity[2, 4]. This residue conservation suggests a functional importance of this region and ensures the adaptation against alternative substrates with different 5' spatial structures. The side chains of R470, R474, and R477 make contacts with the phosphates of -6 and -4 nucleotides on the complementary strand, while K475 binds the phosphates of -9 and -10 nucleotides, and H471 binds the phosphate of the -11 nucleotide[4].

As it is mentioned above, the bio-chemical studies of deaminase domain of ADAR are mostly based on the studies on ADAR2. Although ADAR1 has not been studied as the same level of structural detail, research has identified significant features and differences between ADAR1 and ADAR2. For instance, residue E1008 in ADAR1, corresponding to E488 in ADAR2, is implicated in the base-flipping mechanism essential for RNA editing[76], which the E1008 is flanked with two glycine (G) including G1007[76]. The G1007R mutation of ADAR1 is found relating to severe diseases like Aicardi-Goutieres Syndrome (AGS) and Dyschromatosis Symmetrica Hereditaria (DSH) because this mutation makes ADAR1 catalytically inactive[77, 78]. Moreover, the mutated ADAR1 becomes a competitive inhibitor of wild-type ADAR1[77, 78]. If E1008 is mutated into a large, polar residue like glutamic acid (E), glutamine (Q), arginine (R), histidine (H) and lysine (K), the catalytic function of ADAR1 can be still maintained[76]. Interestingly, some E1008 mutations can lead to increased catalytic efficiency to some specific substrates[76]. Conversely, the catalytic function of ADAR1 can be still maintained if G1007 is mutated into a small, non-polar residue like alanine (A), glycine (G) and valine (V)[76]. But G1007R mutation leads to catalytically inactive ADAR1, which leads to Aicardi-Goutières syndrome (AGS) and dyschromatosis symmetrica hereditaria (DSH)[76]. Interestingly, wild-type ADAR1 doesn't edit A-G mismatch efficiently[41, 71], but E1008R mutation provide a better editing efficiency against A-G mismatch. Further quantitative experiments have revealed that mutations E1008Q and E1008H result in higher catalytic activity compared to the wild-type, suggesting potential positive regulatory mechanisms of ADAR1 under specific conditions or even disorders[76]. Moreover, mutations within the ADAR coding sequences have been shown to both decrease and enhance ADAR's editing capacity, implying the existence of intricate regulatory mechanisms. In term of editing efficiency, the editing capability of ADAR1L is significantly higher than ADAR1S[79],

whereas ADAR1 deaminase domain is more active than ADAR2 deaminase domain[80]. Additionally, ADAR2 seems to have a more accurate editing site selectivity compared with ADAR1 selectivity making ADAR1 editing is more random than ADAR2 editing[41]. The homolog of ADAR2 454-477 loop sequence is different between ADAR1 and ADAR2, which might contribute to the different substrate selectivity between ADAR1 and ADAR2[2]. Comparative studies across species have shown that in Region 1, residues E488, R481, T490, S495, and R510 are highly conserved between ADAR1 and ADAR2[2]. Similarly, in Region 2, residue K594 is conserved[2]. However, Region 3 displays significant differences; ADAR1's 5' binding loop is longer, contains fewer basic residues, and includes a conserved phenylalanine (F), which contrasts with ADAR2 interaction within the major groove[2]. These observations suggest that the variations in Region 3 could account for the differences in substrate preferences between ADAR1 and ADAR2[4]. Furthermore, the editing activities of ADAR1 and ADAR2 are not always directly correlated with their expression levels, indicating variations in chemical editing efficiencies or the influence of post-transcriptional/translational regulatory mechanisms[81-83].

1.3.4. The R-domain of ADAR gene and their contribution to substrate selection

The structural characteristics of ADAR3 is less studied compared to those of ADAR1 and ADAR2. Unique from ADAR1 and ADAR2, ADAR3 possesses an Arginine (R)-rich single-stranded RNA (ssRNA) binding domain, referred to as the R-domain, located near its N-terminus[84]. However, the specific cellular functions of the R-domain and the molecular mechanisms by which it recruits ssRNA remain largely undefined due to the limited research on ADAR3. Given that ADAR3 expression is predominantly in brain and that it uniquely contains a functional R-domain within the ADAR family, it raises important questions about whether this domain confers specialised substrate selectivity and plays a distinct role in neural activities[14, 84].

Interestingly, a structure similar to the R-domain of ADAR3 has been identified in ADAR2, specifically within exon 1 at the 5' end, suggesting a new variant of ADAR2, here referred to as ADAR2R (not an official name)[14]. This R-domain is highly conserved across several species, including humans, mice, and rats[116], and functions as the 5' UTR in the predominant

ADAR2 isoform[85]. Expression of ADAR2R varies across different tissues and is notably enriched in the hippocampus and colon, possibly due to a distinct promoter or alternative splicing mechanisms within ADAR2 pre-mRNAs[85]. Moreover, ADAR2R does not appear to be regulated by interferon, suggesting it may not play a significant role in virus infections[85]. Moreover, overexpression of ADAR2R does not seem to alter the general adenosine deaminase activity, indicating that the R-domain in ADAR2R might not significantly influence substrate selection[85].

1.4. The ADAT gene domain structures and their functionalities

The molecular structures and mechanisms of ADATs are less understood compared to those of ADARs, with only the deaminase domain clearly identified in ADATs to date. Furthermore, ADAT2 and ADAT3 are known for their ability to modify the anticodon of tRNA, which has more pronounced biological implications compared to ADAT1. Consequently, ADAT1 has not been as extensively studied as ADAT2 and ADAT3.

1.4.1. The deaminase domain of ADAT gene responsible for adenosine deamination on tRNA and substrate selections

The catalytic deaminase domain of ADAT1 is actually closely related to the ADAR catalytic domain, differing significantly from ADAT2 or ADAT3 because their deaminase domain is evolutionally related to cytidine deaminases[15]. Generally, the biochemical reaction of A-to-I editing mediated by ADATs is believed to be similar to that of ADARs, relying on a base-flipping mechanism that flips out and exposes the adenosine for editing. Unlike ADARs, however, the editing activity of ADATs is heavily dependent on the dimerisation of ADAT molecules. In contrast to ADAT2 and ADAT3, ADAT1 typically forms homodimers, whereas ADAT2 primarily dimerises with ADAT3. The C-terminus of ADAT2 interacts with the N-terminal domain of ADAT3 through protein-protein interactions, which forms a tRNA-binding surface[86]. This tRNA binding surface of ADAT dimers will wrap around the flipped A34 and deaminate it into I34[86]. Because ADAT2/3 dimers recognise A34 substrates mainly by conformational structural features, which is believed as not sequence-specific[86-89]. Consequently, ADAT2/3 exhibit different preferences for various tRNA substrates, influenced by the distinct conformational structures of these tRNA molecules[87]. An adenosine at the 34^{th} position of the tRNA can alter the anticodon's structural conformation, disrupting the hydrogen bond between U33 and the phosphate of nucleotide 36, which stabilizes the U-turn conformation of the tRNA anticodon stem-loop (ASL)[36, 90]. This disruption leads to significant structural distortion, causing the adenosine to flip out and become accessible for deamination by ADAT2/3 heterodimers[89]. Interestingly, studies have shown that tRNA

substrates with a purine (either cytosine or uracil) at the 35th position are the most favourable substrates for ADAT editing[88, 91].

Moreover, ADAT2/3 dimers extend their functional interaction beyond the tRNA anticodon stem-loop; they also competitively bind to the acceptor stem of tRNA[87]. This capability allows ADAT2/3 to potentially edit precursor tRNA both before and after tRNA maturation[88, 92]. But the efficiency of ADAT editing varies with different tRNAs, which exhibit distinct sequence signatures. Furthermore, ADAT2/3 dimers demonstrate higher editing efficiency in tRNA fragments that lack the 3' CCA end[87]. The 3' CCA end is essential for the charging of pre-mature tRNA with amino acids by Amino-acyl tRNA Synthetase (aaRS)[93], highlighting the complex interplay between tRNA maturation processes and the editing activities of ADAT2/3.

1.5. The dimerisation of ADAR and ADAT genes and the subsequential impacts on RNA editing

As previously mentioned, dimerisation is crucial for the editing functionality of ADATs. However, this phenomenon is not exclusive to ADATs, and it is also observed in ADARs, which requires further discussion to elucidate its potential impacts on biological functions. Understanding the role of dimerisation in these enzyme families could reveal new insights into their regulatory mechanisms and the broader implications for RNA editing and cellular processes.

Dimerisation plays a critical role in the functionality of both ADARs and ADATs, where these enzymes predominantly form multimers, usually as dimers, within the cell. Similar phenomenon is also observed with other molecules having editing functions such as apolipoprotein B mRNA editing enzyme, catalytic polypeptide, or APOBEC, for dC-to-dU editing, which won't be discussed here[94-100] (Appendix 4).

1.5.1. The dimerisation of ADAR genes

ADAR1 and ADAR2 are known to form homodimers within cells, primarily within the nucleus, through protein-protein affinity (ADAR1-ADAR1 homodimers and ADAR2-ADAR2 homodimer)[101-104]. This phenomenon is also reported in other species that possess genes encoding ADAR enzymes. Interestingly, it has also been documented that ADAR1 and ADAR2 can form heterodimers (ADAR1-ADAR2 heterodimers), but ADAR3 cannot form any dimers and ADAR3 remains catalytically inactive[103, 105]. This inability to dimerise may contribute to ADAR3's deamination incapacity, alongside the destabilised reaction core caused by the A510 residue. However, there are reports suggesting that ADAR3 can form homodimers specifically within the brain, hinting at a specialised regulatory function of ADAR3 in neural contexts[105]. The dimerisation of ADAR is RNA independent, but only with bound RNA, the ADAR dimers can be translocated into nucleus[103, 106]. The dimerisation interface structure is highly conserved, and mutations in the corresponding coding sequences can lead to failure in forming dimers or imperfect dimerisation, resulting in a loss of deamination capacity[101]. It has been proven that the editing capacity of ADARs is maintained only if dimerisation is

unaffected and the dsRNA binding capability of ADAR1 and ADAR2 remains intact[107]. Interestingly, while both dimerisation and dsRNA binding are essential for ADAR editing activity, these functions are separate and do not interfere with each other[107]. Moreover, variations in dimerisation outcomes and efficiency do not appear to be the cause of the differing editing efficiencies observed among various ADARs[28].

In the double-stranded RNA binding domains (dsRBDs) of ADAR proteins, there exists a specific motif known as the KKxxK motif, which is essential for binding to the major groove of dsRNA[108, 109]. Mutations within the KKxxK motif in dsRBDs result in a loss of capability for both binding dsRNA substrates and catalysing deamination reactions[107]. Dimers of KKxxK-mutated ADAR2 and hybrid of KKxxK-mutated ADAR2 with wild-type ADAR2 both lost the editing activity[107]. Interestingly, the mutation in deaminase-critical residue E396 doesn't affect the editing activity of mutated/wild-type hybrid dimer (E396/WT), but KKxxK-mutated hybrid does. This suggests that dsRNA may be bound by one dsRBD from each ADAR monomer within a dimer, with the deamination reaction being catalysed by one of the deaminase domains[107]. Furthermore, heterodimers comprising a mutated monomer (with a mutated deaminase domain that lacks deaminating capacity) paired with a wild-type monomer show significantly reduced editing capacity[105].

1.5.2. The dimerisation formation of ADAT genes

As it mentioned above, the ADAT1 forms homodimers and ADAT2 and ADAT3 form heterodimers in general. The ADAT1 homodimers turn itself into catalytically active form whereas the ADAT2 and ADAT3 form heterodimers with each other to create catalytically active form[34, 110]. Therefore, the dimerisations of ADATs are crucial for their editing capacity. Within the ADAT2/3 heterodimers, ADAT2 serves as the catalytic subunit, while ADAT3 remains catalytically inactive, regardless of its dimerisation status[91, 110]. ADAT3 is catalytically inactive because the essential catalytic glutamate residue (E) is absent in the ADAT3 deaminase domain. This residue is crucial for mediating proton transfer during the deamination reaction[91, 110]. Despite its catalytic inactivity, ADAT3 still plays a significant role in recognising tRNA substrates. Both ADAT2 and ADAT3 are necessary for maintaining

the catalytic function of the ADAT2/3 heterodimers, highlighting the importance of their collaborative interaction in the editing process[110].

1.6. Intracellular localisation of ADAR and ADAT genes affect substrate selections of RNA editing

Having reviewed the molecular structures, domain characteristics, and dimerisation activities, we have addressed the mechanisms through how RNA is edited. The next critical question is to explore where RNA editing occurs. Understanding the specific cellular locations of RNA editing catalysed by ADARs and ADATs is essential for further elucidating how these modifications impact various cellular processes. Figure 5 shows the intracellular localisation enrichment of members in ADAR or ADAT family, which is acquired from GeneCard database[1], which is based on predictions of confidence levels generated from COMPARTMENTS database[111].

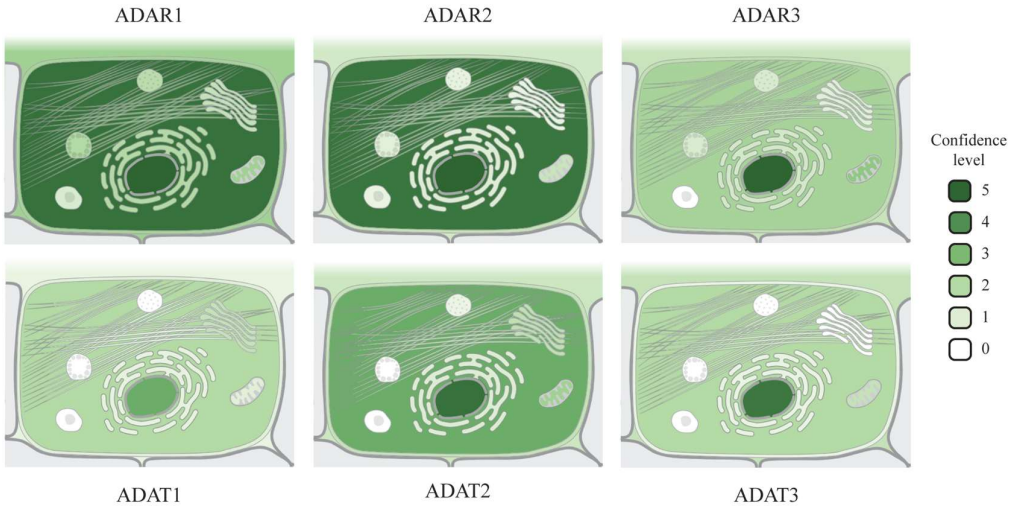


Figure 5. Intracellular localisation enrichment of ADARs and ADATs. All the members in ADAR and ADAT families are shown including ADAR1, ADAR2, ADAR3, ADAT1, ADAT2, ADAT3. The data is inherited from GeneCard database[1]. The confidence levels demonstrated is the metrics score generated the predictions indicating how reliable is the protein presented in the subcellular location. These scores are derived from multiple sources, including experimental data, text mining, and computational predictions.

1.6.1. Localisation and transport of ADAR genes and subsequential impacts on substrate selections

As previously noted, ADAR1L is found in both the cytoplasm and the nucleus, whereas ADAR1S predominantly resides in the nucleus, each employing distinct mechanisms for shuttling across the nuclear membrane[20, 112]. This difference is due to ADAR1L possessing a functional Nuclear export signal (NES, leucine-rich) but ADAR1S does not[113, 114]. Consequently, ADAR1L can more readily traverse the nuclear membrane, often in conjunction with other accessory proteins[113]. Both ADAR1L and ADAR1S are capable of transport between the nucleus and cytoplasm, though they utilise different molecular mechanisms to achieve it[112, 114, 115]. Notably, an enhanced exportin 1 (XPO1) dependent NES is identified in ADAR1L but is absent in ADAR1S, and this NES overlaps with the first z-DNA binding domain[114]. XPO1 and Ran-GTP (RAs-related Nuclear protein bound with guanosine triphosphate) interact with ADAR1L NES to form a complex facilitating ADAR1L transporting into cytoplasm[113]. Additionally, the binding between XPO1 and NES interferes the binding of Transportin 1 (TRN1) into Nuclear localisation signal (NLS) located inside the third dsRBD[114]. Binding of other factors to the first dsRBD can also interfere with the TRN1-NLS interaction[114]. Therefore, RNA substrates binding to either dsRBD1 or dsRBD3 can interfere with the translocation and accumulation of ADAR1 within the cell[114]. Although RNA binding is not required for the NLS activity of dsRBD3, it is essential for the cytoplasmic accumulation mediated by dsRBD1[114]. The cytoplasmic localisation induced by dsRBD1 might depend on a common RNA bound by both dsRBD1 and dsRBD3, where the RNA binding masks the NLS within dsRBD3[114]. Interestingly, while many nucleus-cytoplasm shuttling proteins are transcription-dependent, the shuttling of ADAR1L is transcription-independent[114, 116, 117]. A leucine-zipper-like structure is identified within the deaminase domain interfering nucleus-cytoplasm transportation, which may be responsible to dimerisation[114].

Both ADAR1L and ADAR1S possess a nuclear localisation signal (NLS), essential for their import into the nucleus from the cytoplasm[113]. The NLS of ADAR1L and ADAR1S overlaps the third dsRBD, and the third dsRBD, which plays a crucial role in regulating the nuclear import of ADAR[118]. Notably, ADAR1L is equipped with more than one NLS, though the

implications of this redundancy remain unclear (i.e. extra capability to shuttle across nuclear membrane)[113]. The translocation of ADAR proteins between cellular compartments is a potential regulatory mechanism, as the location of ADAR within the cell influences its substrate interactions. For instance, pre-mRNA is typically edited in the nucleus, whereas mature mRNA is mostly edited in the cytoplasm[113]. This suggests that the cellular distribution of ADARs is strategically significant, potentially affecting the scope and specificity of RNA editing. However, it has been suggested that ADAR molecules must be unbound to RNA to traverse the nuclear membrane, as RNA-bound dsRBDs can interfere with the recruitment of essential transport proteins like TRN1 and XPO5 (Exportin-5)[115]. TRN1, in particular, binds to the NLS and mediates a Ran-GTP-dependent import mechanism that facilitates the translocation of ADAR1 into the nucleus[115].

When it comes to ADAR2, it predominantly accumulates in the nucleolus within the nucleus[112]. Unlike ADAR1, ADAR2 lacks a nuclear export signal (NES) but possesses a non-canonical nuclear localisation signal (NLS) located between residues 75 and 132[112]. Consequently, both ADAR1 and ADAR2 are dynamically involved in continuous movement in and out of the nucleolus[112]. Although ADAR1 and ADAR2 localise to the nucleolus, they engage with RNA substrates primarily in the nucleoplasm, not within the nucleolus. This suggests that the nucleolus serves more as a transient storage location, with ADAR1 and ADAR2 shuttling dynamically in and out[112]. This shuttling movement of ADAR2 is influenced by its capacity to bind ribosomal RNA (rRNA), which can enhance the ability of ADAR2 to edit substrates within the nucleus[106]. Thus, the nucleolus appears to function as a temporary site for functional sequestration of ADAR2, potentially regulating its editing activity by modulating its shuttling between the nucleolus and nucleoplasm.

Similar to ADAR2, ADAR3 predominantly accumulates inside the nucleus[119]. ADAR3 contains a functional NLS that overlaps with its arginine-rich R-domain, and Karyopherin Subunit Alpha 2 (KPNA2) can bind to this NLS, facilitating the entry of ADAR3 into the nucleus[119]. Interestingly, a similar NLS is also identified in ADAR2R, a variant of ADAR2[119], thus it may has the similar localisation of ADAR2.

1.7. Substrates selections of ADAR and ADAT proteins

Now that we have elucidated the mechanisms underlying the deamination reaction, the next important question will be, where does the deamination happen. In another word, what types of substrates (RNA) are susceptible to editing by ADARs and ADATs.

1.7.1. Substrates selection of ADAR proteins

ADAR recognising substrates is initiated through molecular attraction of the RNA spatial structure, which initially must be A-form helix double-stranded RNA (dsRNA) (hint. dsDNA structure is a B-form helix instead, Appendix 2)[2, 120, 121]. Regarding this property of ADAR recognition, spatial changes of RNA structures caused by RNA sequences alterations will eventually either enhancing or diminishing ADAR recognition and editing efficiency at certain levels[40, 122, 123]. Consequently, RNA molecules featuring structural elements such as hairpins, bulges, and loops are more attractive to ADAR binding and subsequent editing[40, 49, 122-124] (Figure 6). It follows that ADARs prefer to edit relatively longer RNA molecules, which typically contain more potential binding-facilitating structures[4, 48, 122, 123]. When it comes to shorter length limits, a small molecule like a 15-bp dsRNA stem with a single base mis-match is already sufficient for ADAR binding and editing if the ADAR molecules accept its spatial structure[48]. Therefore, ADARs have the capacity to edit adenosines in various RNA molecules beyond mRNA or pre-mRNA, provided these substrates possess the preferred spatial structures. Indeed, significant impacts of ADAR editing on miRNA biology have been documented, including the editing of miRNA or pre-miRNA (i.e. pri-mir-142)[4, 123]. This highlights the fact that ADAR recognises substrates purely on spatial structure affinity towards RNA molecules, indicating a preference for certain sequences, though not to the extent of sequence-specificity seen in mechanisms like the CRISPR-Cas9 small-guided RNA system.

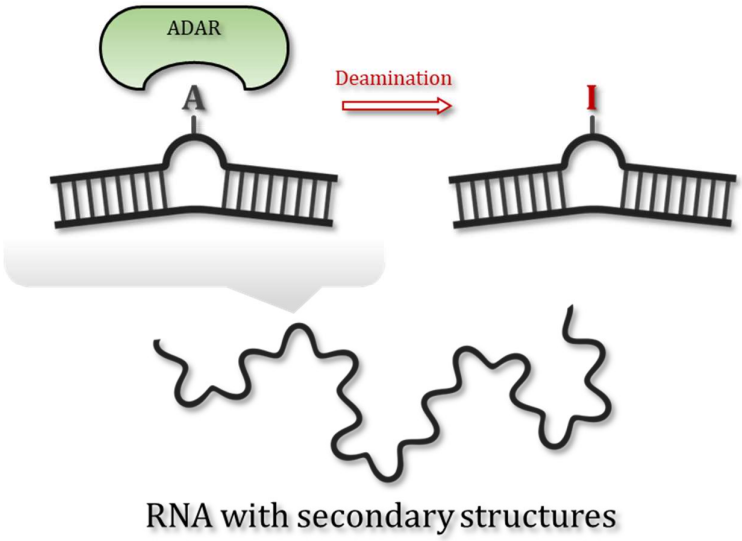


Figure 6. ADAR editing RNA substrates. ADAR prefers RNA molecules with secondary structures such as double-stranded features like bulges, hairpins etc. The adenosine/A is deaminated and converted into the inosine/I.

If the double-stranded structures are favoured by ADARs, this means that the adenosine being edited will pair up with another nucleotide when forming double-stranded structures. Notably, adenosine is more readily edited in the presence of mismatches against the target nucleotide. The nature of these mismatches, specifically their position and type, plays a critical role in determining the efficiency of adenosine editing. For instance, adenosine paired with cytosine (A-C mismatch) exhibits the highest editing efficiency among mismatch types[2, 48]. Analysis of sites that undergo extensive editing reveals that the optimal configuration for editing typically involves a 5'-UAG-3' sequence with an A-C mismatch at the editing site, flanked by at least eight base pairs on the 5' side[4]. Differences in molecular structure among various ADAR isoforms lead to distinct preferences for the spatial structure of dsRNA substrates. Consequently, ADAR1L, ADAR1S, and ADAR2 each exhibit unique substrate preferences[49, 120, 123, 125]. Notably, the overlap of ADAR1L, ADAR1S, ADAR2 targeted substrates is minimal, where ADAR1 and ADAR2 substrates range is clearly divided[125, 126]. Studies on RNA editing events reveal that ADAR1S and ADAR2 tend to dominate the editing landscape, while ADAR1L is less dominant, suggesting that ADAR1L functions more

as an inducible protein, requiring regulatory activation by cytokines such as interferons (IFN)[125, 126]. In contrast, the catalytically inactive ADAR3 primarily acts as a negative regulator of ADAR2 by competing for the same substrates, given their similar substrate preferences[29]. This dynamic interplay among ADAR members implies the complexity of RNA editing mechanisms and their regulation within cellular processes.

Interestingly, ADAR2 exhibits a particular behaviour known as self-editing, where it can edit its own mRNA transcripts[28]. This process, termed ADAR2 self-editing, can influence the alternative splicing of its transcripts. Multiple self-editing sites on ADAR2 transcripts are identified, but most of those editing creates nonsense transcripts disrupting transcripts' normal function, which will be quickly degraded through nonsense-mediated decay (nonsense transcripts deletion mediated by UPFs)[17]. The implications and regulatory mechanisms of this sub-dominant and sub-optimal isoform of ADAR2 remain unclear. However, its expression level has been observed to increase slightly during neuronal development, suggesting that self-editing may serve as a potential regulatory mechanism modulating ADAR2 function[28]. Unlike ADAR2, no self-editing events have been identified for ADAR1 yet, nor for the enzymatically inactive ADAR3[17].

The mRNA containing inosine after a deamination reaction remains fully functional and is presented on translating ribosomes, which later get translated into polypeptides[127]. Thus, A-to-I editing primarily alters RNA function based on sequence alterations. ADAR is verified capable of binding and editing more adenosines on the inosine-containing post-edited RNA, demonstrating that inosines do not inhibit subsequent A-to-I editing events[40]. Moreover, the increased inosine composition in mRNAs potentially alter their structural stability by decreasing A-C mismatch compositions[128]. Inosine within edited transcripts can distort mRNA secondary structures, which may prevent miRNA binding or affect the accessibility of other regulatory molecules[129]. Additionally, hyper-edited mRNA containing multiple consecutive I-U base pairs can be cleaved by a subunit of the RNA-induced silencing complex (RISC), although the biological consequences of this cleavage are not fully understood[130].

Furthermore, A-to-I editing are not the only one simple RNA modification, and it is beneficial to understand how ADAR editing correlate to other RNA modifications. For instance, ADARs are known to struggle with editing transcripts containing N^6-methyladenosine (m^6A), which are adenosines methylated by a large m^6A methyltransferase complex[131, 132]. Surprisingly,

despite being difficult substrates, N6-methyladenosines in mRNA are still editable by ADAR, albeit at a significantly reduced rate of approximately 2% compared to normal adenosines[133]. N^6-methyladenosine also plays an important role in virus infection and virus-related cellular regulations, such as those involving the PKR associated pathway[134]. Moreover, other synthetic adenosine analogs such as 8-azanebularine, thieno[3,4-d]-6-aminopyrimidine ((th)A), 2'-deoxyadenosine and 2'-deoxy-2'-fluoroadenosine have been shown to be editable, though their editing rates vary compared to traditional adenosine[65, 133, 135, 136]. This suggests that inosines in edited RNAs could have more profound biological functions than previously appreciated.

For a comprehensive discussion of substrate selection by ADAR enzymes, I have compiled a detailed summary of the potential editing substrates in human cells alongside their possible biological outcomes (Figure 7). This overview is designed to provide readers with a clearer, more integral understanding of ADAR editing impacts of ADAR editing. The identified substrates include: 1) mRNA exons. Editing within exons can lead to amino acid substitutions in the encoded proteins. If these substitutions are non-synonymous, they might alter protein function or lead to protein misfolding and subsequent functional disruption; 2) mRNA intron-exon junctions. Editing at these junctions can interfere with normal splicing mechanisms, potentially resulting in alternative splicing errors or the creation of novel isoforms with distinct functions; 3) 5' or 3' UTRs of mRNA or pre-mRNA. Editing in these untranslated regions can modify the regulatory sequences that control mRNA stability, localisation, or translational efficiency. Such alterations may affect the binding sites for miRNAs or other regulatory molecules, leading to changes in gene expression; 4) miRNA. Editing of miRNAs can affect their maturation process, alter their target specificity, or disrupt their regulatory capacity, which can have broad implications for cellular gene regulation networks.

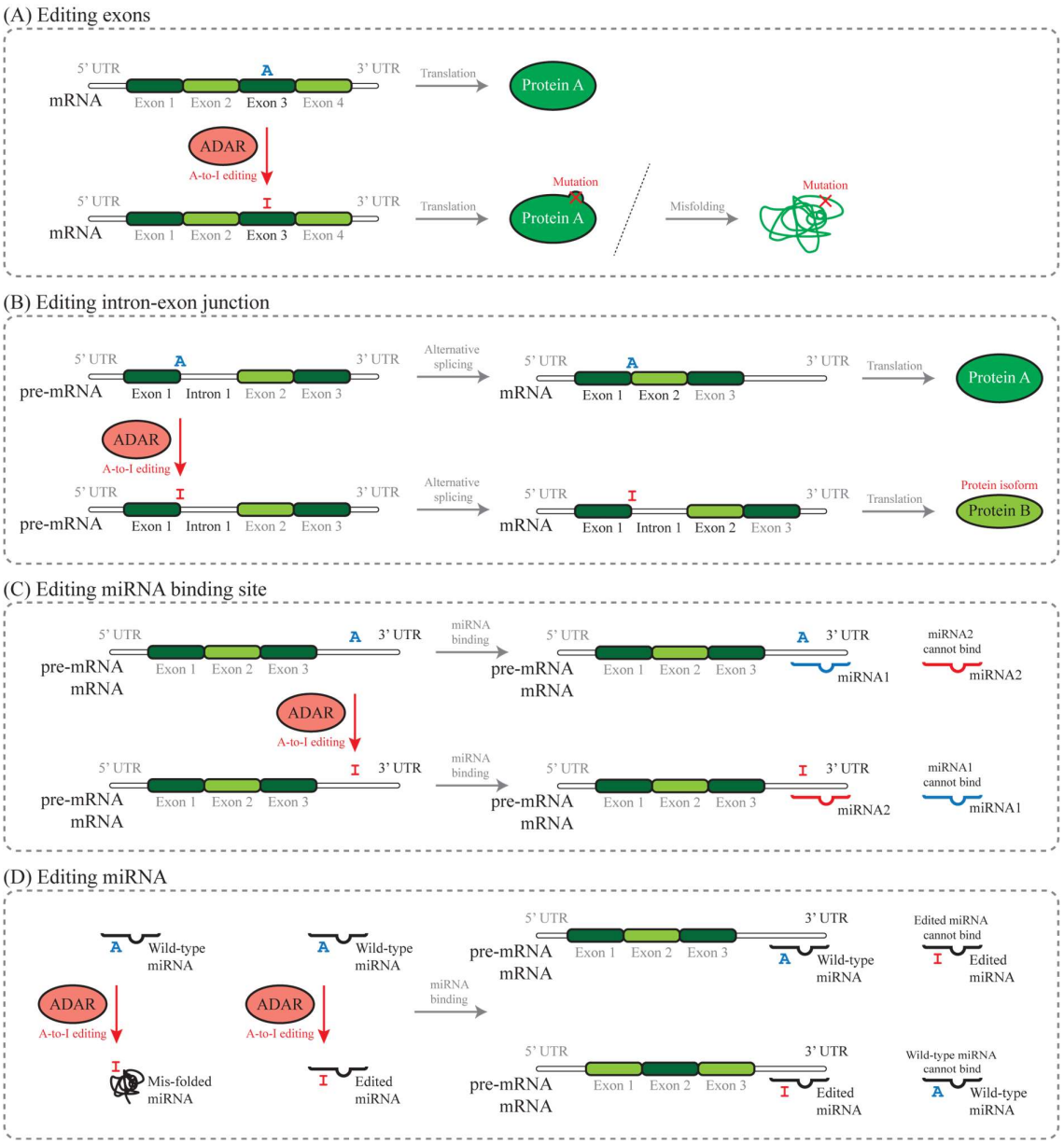


Figure 7. Possible outcomes of A-to-I editing on host's transcript. Four conditions are listed, which are (A) editing on mRNA exon, (B) editing on mRNA intron-exon junction, (C) editing on mRNA 5' or 3' UTR, and (D) editing on miRNA.

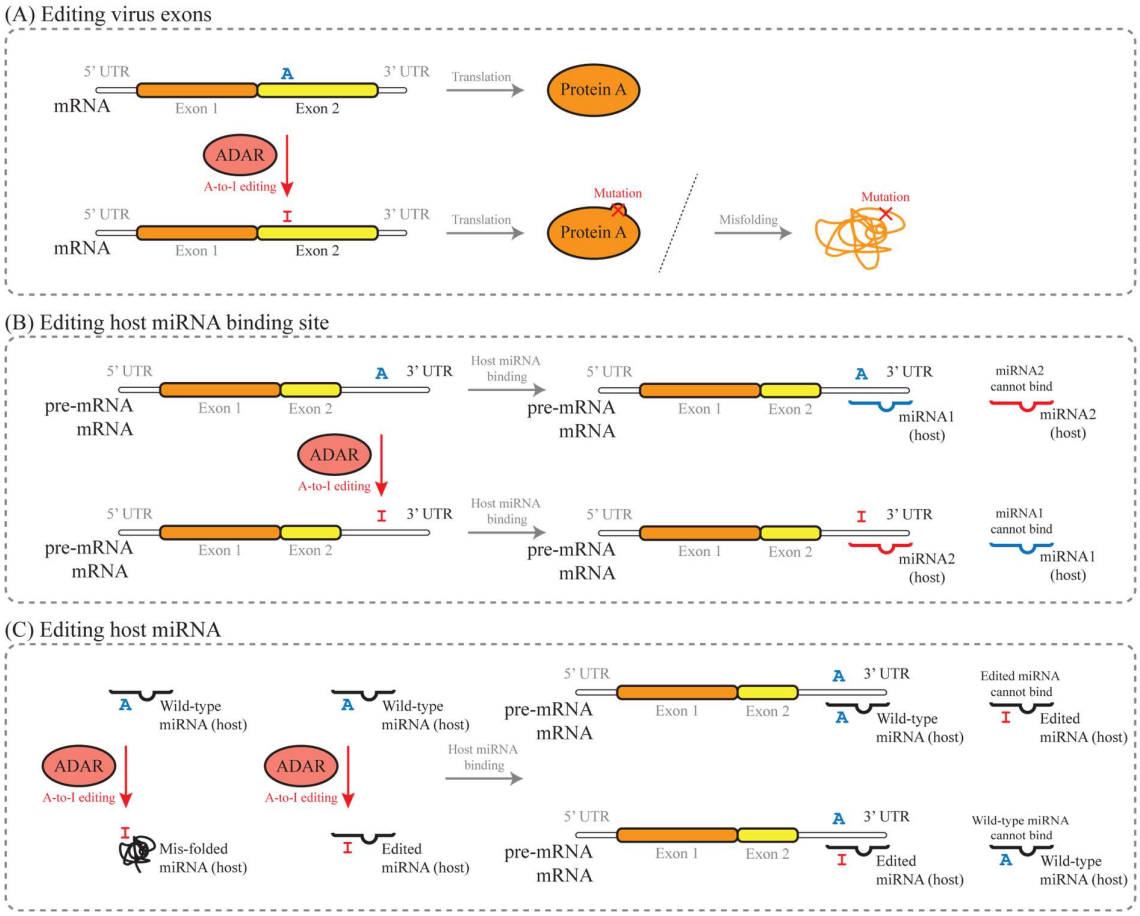


Figure 8. Possible outcomes of A-to-I editing on viral transcript. Three conditions are listed, which are (A) editing on viral coding mRNA, (B) editing on host's miRNA, (C) editing on viral mRNA 5' or 3' UTR (predicted) and other UTR region. The editing on viral UTR can cause other outcomes according to the regional functions.

To better understand ADAR-mediated RNA editing in the context of viral infections, I have also compiled a summary of potential editing events on virus-originated substrates, alongside their anticipated outcomes (Figure 8). This analysis extends the findings from host substrate editing to include viral RNA, providing a broader perspective on how ADAR activity may influence viral pathogenesis and host-virus interactions. The editing on viral RNA includes: 1) viral mRNA exons. Editing within these regions can result in amino acid substitutions in viral proteins. If these mutations are non-synonymous, they might lead to functional changes in the viral proteins or potentially cause protein misfolding, which can impact the viral ability to

replicate or interact with host cells effectively; 2) 5' or 3' UTRs of viral mRNA or pre-mRNA. Editing in these untranslated regions of viral genomes may alter sequences that are crucial for the virus life cycle, such as those involved in mRNA stability and translation. Modifications here can affect the interaction of viral RNAs with host miRNAs, potentially disrupting normal viral gene expression and evasion strategies; 3) host miRNA targeting viral transcripts: Editing of host miRNAs that target viral RNA can change the specificity and efficiency of these miRNAs. Such alterations might lead to changes in the regulation of viral transcripts, impacting viral replication and the host's antiviral response. Additionally, editing could cause errors in miRNA maturation, further influencing the host's capacity to regulate virus-associated gene expression.

1.7.2. Functional consequences of tRNA editing by ADAT protein

While ADAR editing has been extensively studied, the ADAT editing mechanism remains less understood. To date, the only confirmed A-to-I editing event by ADAT1 is the deamination of adenosine 37 (A37) in eukaryotic $tRNA^{Ala}_{AGC}$ to inosine (I37)[15]. The post-edited I37 is subsequently methylated to m^1I37 by tRNA methyltransferase 5 (TRMT5), which methylated inosine diminishes the impacts on tRNA stability caused by ADAT1 editing[15, 92, 137] (Figure 9.A). However, no significant biological impacts on cellular activities are found with either ADAT1 editing A37 into I37 and TRMT5 methylating I37 into m^1I37 until now[35].

To date, the major research focus of ADAT gene is on ADAT2 and ADAT3 due to their distinct editing outcomes on tRNA. ADAT2 and ADAT3 primarily convert adenosine 34 (A34), the wobble position of the tRNA anticodon, into inosine (I34)[15, 35]. Unlike I37 being methylated into m^1I37, post-edited I34 is not subsequently methylated and remains biologically functional. The conversion of A34 to I34 enhances tRNA's pairing flexibility during translation other than traditional Watson-Crick A-U pairing, allowing it to form base pairs with uracil (U), adenosine (A), or cytosine (C) on the mRNA codon's third nucleotide (3^{rd} nt, wobble position), resulting in I-U, I-A, and I-C pairings, respectively[34-36] (Figure 9.B). However, I-A pairing is relatively inefficient compared to I-U and I-C pairings[91]. This increased pairing flexibility of I34 significantly counters the redundancies of synonymous codons during translation[89]. Additionally, previous studies shown that knockdown of ADAT2 or ADAT3 will significantly

lower I34-tRNA level for all potential substrates, but it does not lead to cellular lethality[92]. And the knockout of ADAT2 will lead to cellular lethality, but not ADAT3, which proves ADAT2 is an essential gene[92].

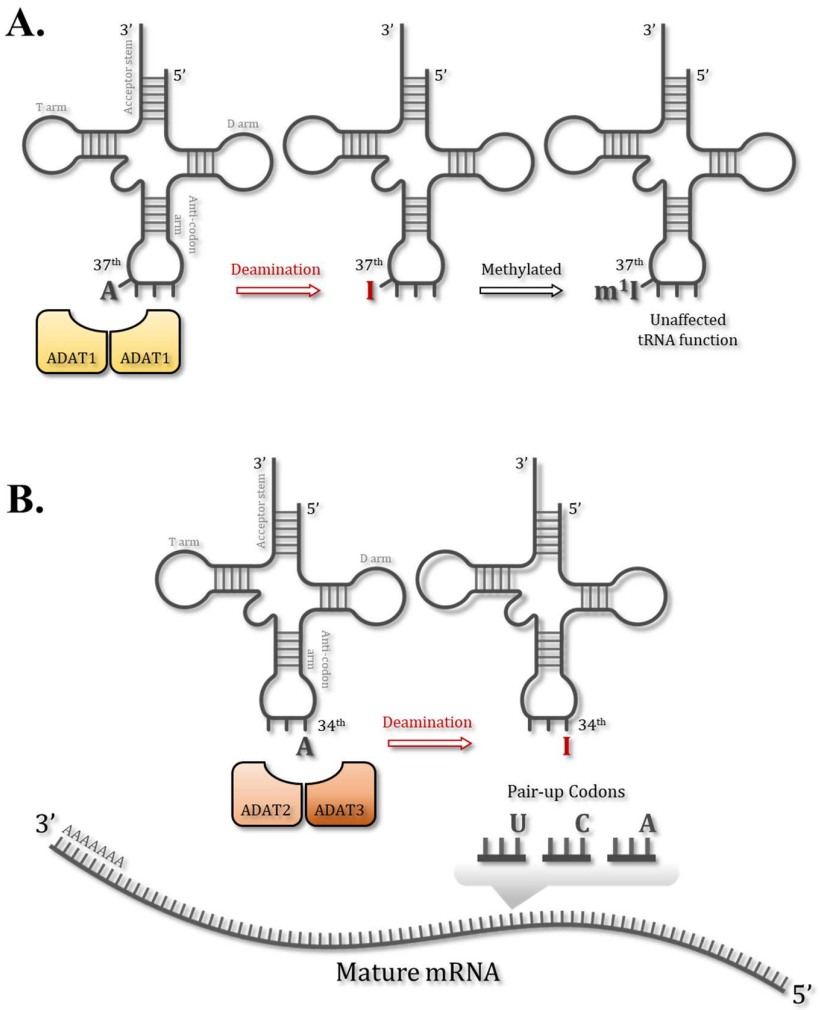


Figure 9. Schematic drawing of ADAT editing on tRNA substrates. (A) Editing of ADAT1 homodimer converts the 37th adenosine/A into inosine/I on tRNA substrates. The inosine/I is later methylated, maintaining the tRNA structure. (B) Editing of ADAT2 and ADAT3 heterodimer converts the 34th (wobble position) adenosine/A into inosine/I on tRNA substrates. The inosine/I is able to pair up with uracil/U, adenosine/A, or cytosine/C, which increase the flexibility in decoding.

According to previous studies, human ADAT2 and ADAT3 are proven to edit A34 of total eight tRNAs: $tRNA^{Thr}_{AGU}$, $tRNA^{Ala}_{AGC}$, $tRNA^{Pro}_{AGG}$, $tRNA^{Ser}_{AGA}$, $tRNA^{Leu}_{AAG}$, $tRNA^{Ile}_{AAU}$, $tRNA^{Val}_{AAC}$, $tRNA^{Arg}_{ACG}$[87, 89, 92, 138]. This editing endows these tRNAs with enhanced decoding capacities during mRNA translation, allowing the inosine at the wobble position to pair with C, A, and U. The corresponding 8 amino acids including Threonine (Thr/T), Alanine (Ala/A), Proline (Pro/P), Serine (Ser/S), Leucine (Leu/L), Isoleucine (Ile/I) and Valine (Val/V) are ADAT-related amino acids, which will be further investigated on later research. Those eight amino acids are sometimes notated as TAPSLIVR. Interestingly, ADAT2 and ADAT3 are found incapable to edit $tRNA^{Gly}_{ACC}$, which $tRNA^{Gly}_{ACC}$ is absent in eukaryotic genomes and $tRNA^{Gly}_{GCC}$ is most abundant $tRNA^{Gly}$ iso-acceptor in human[89]. To be accurate, ADAT2 and ADAT3 can bind $tRNA^{Gly}_{ACC}$ but cannot deaminase it because structural feature of $tRNA^{Gly}$ is incompatible with I34 causing significant structural instability[89]. Surprisingly, other tRNAs with G34 are often either absent of corresponding coding genes in human genomes or very low in expression level[92]. Thus, this raises an intriguing question whether only edited I34-containing tRNA can theoretically decode codons with cytosine (C) in 3rd position.

To provide a comprehensive overview of ADAT2/3 editing outcomes, I highlighted the ADAT affected codons in codon table (Figure 10). Additionally, I have included data of the predicted tRNA gene count and the expression levels of corresponding mature tRNAs for each codon. The predicted tRNA gene count was derived from the GtRNAdb database, based on tRNAscan-SE analysis of the complete human genome (Homo sapiens GRCh38/hg38)[138]. The expression levels of mature tRNAs were obtained from a previously published study that determined the abundances of mature tRNAs in human HEK293 cells using a novel RNAseq-based method[139]. Before being assigned to their respective codons, the expression levels of different tRNA iso-acceptors were summed for corresponding codons.

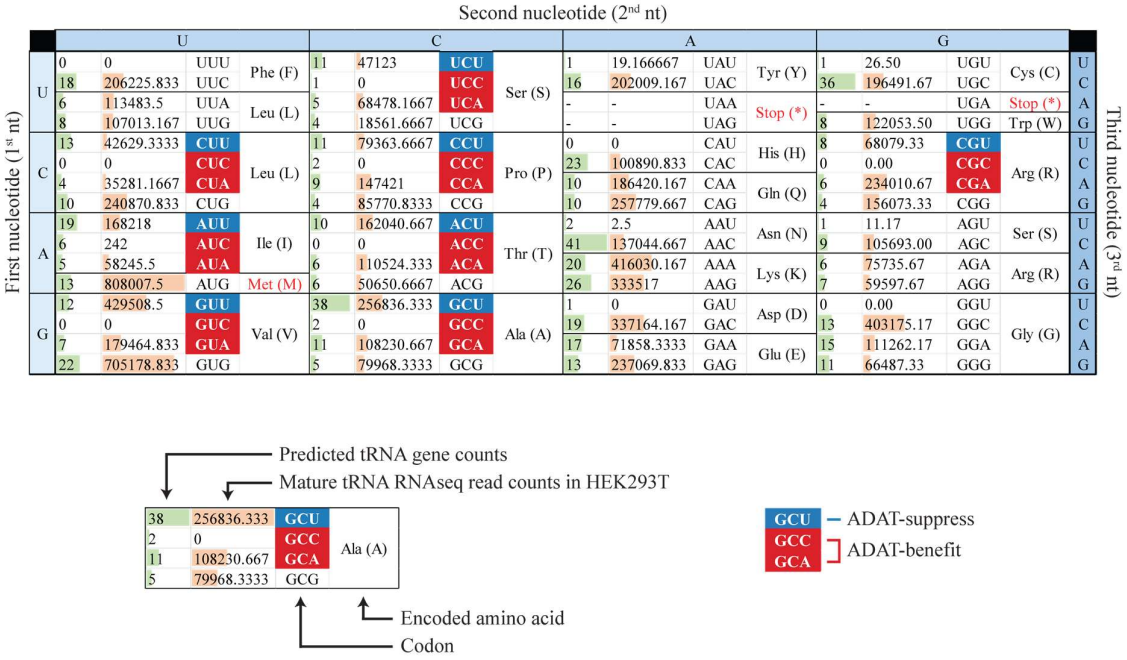


Figure 10. Codon table with human tRNA supplies and ADAT editing relations. The predicted tRNA gene counts and the mature tRNA RNAseq read counts in HEK293 are provided, which are both from previous publications. The predicted tRNA gene counts is from GtRNAdb database, which is computed with tRNAscan-SE tool. The mature tRNA RNAseq read counts are from raw data of previous publication regarding tRNA sequencing methods. The ADAT relations of the codons are also shown including ADAT suppress (red) and ADAT benefit (blue). ADAT benefit codons are the codons that have more decoding tRNAs under ADAT expression, whereas ADAT suppress codons are the codons that have less decoding tRNAs under ADAT expression.

According to the codon table, the eight ADAT2/3-editable tRNAs do not lead to amino acid substitutions nor the creation of start/stop codons, suggesting that the primary function of I34-containing tRNAs is to decode synonymous codons. Research has confirmed that I34-containing tRNAs can enhance the decoding efficiency of translation[91, 140, 141]. However, the impact of this enhancement still depends on the abundance of these edited tRNAs, or their corresponding mature tRNA levels[142]. For instance, ADAT2/3-editable $tRNA^{Ala}_{AGC}$ represents approximately 68% of the total $tRNA^{Ala}$ pool[142]. Additionally, unedited tRNAs (containing A34 not I34) are found to be less efficiently charged with their cognate amino acids[91]. The tRNA modification, which increases decoding capability of modified tRNAs, is predicted positively selected during evolution, because ADAT2/3 editable codons are relatively enriched in eukaryotic codon pool[142, 143]. Therefore, the abundance of ADAT2/3-editable tRNAs' preferred codons suggest a mechanism for translational control, where the translational efficiency of genes enriched with such codons could be regulated by modulating

ADAT2 and ADAR3 expression levels, thus influencing the abundance of edited or enhanced decoding tRNAs. From an amino-acid-wise perspective, the proteins enriched with highly enriched in TAPSLIVR amino acids, which are amino acids corresponding to ADAT2/3-editable codons, can benefit from ADAT2/3 editing increasing their synthesis efficiency[140, 143]. Surprisingly, eukaryotic proteomes are significantly abundant in TAPSLIVR-rich proteins, which also tend to be longer compared to those in other species[143]. This suggests that ADAT2/3 editing may have played a significant role in the evolution of eukaryotic genomes[142]. In summary, ADAT2 and ADAT3 editing can increase both the efficiency and fidelity of translation, unlike ADAT1[91, 142, 143]. However, a reduction in I34-containing tRNA levels does not necessarily equate to a significant decrease in translation efficiency or biogenesis rates because the overall tRNA expression level might not be decreased. Ultimately, the specific contribution of I34-containing tRNAs created by ADAT2/3 editing depends on the codon usages of the actual mRNA sequences being translated, whether these are from host or viral origins.

The translation process involving tRNA can be prone to errors when ribosomes decode mRNA with tRNA recruitment[144, 145]. One explanation why ADAT2/3 editing does not typically introduce errors during translation involves the nature of I34-containing tRNAs. Theoretically, the ADAT2/3 edited tRNA could cause mis-decoding due to ribosome-derived errors (i.e. $tRNA^{Phe}_{\underline{I}AA}$ against $Leu_{(UU\underline{G})}$, $tRNA^{His}_{\underline{I}UG}$ against $Gln_{(CA\underline{G})}$, $tRNA^{Cys}_{\underline{I}CA}$ against $Trp_{(UG\underline{G})}$) or even affect the stop codons decoding (i.e. $tRNA^{Tyr}_{\underline{I}UA}$ against $Stop_{(UA\underline{C})}$, $tRNA^{Cys}_{\underline{I}CA}$ against $Stop_{(UG\underline{A})}$)[34]. The human tRNAome includes a large number of G34-containing tRNA coding sequences corresponding to these potentially erroneous codons, but has limited copies of A34-containing tRNA coding sequences (i.e. 12 copies of $tRNA^{Phe}_{\underline{G}AA}$, 11 copies of $tRNA^{His}_{\underline{G}UG}$, 30 copies of $tRNA^{Cys}_{\underline{G}CA}$, but limited copies number of encoding sequences for $tRNA^{Phe}_{\underline{A}AA}$, $tRNA^{His}_{\underline{A}UG}$, $tRNA^{Cys}_{\underline{A}CA}$)[34, 138]. There are theoretically seven A34-tRNAs, each associated with two-synonymous codon boxes, which could potentially generate translation errors or mis-decoding after being edited by ADAT2/3: $tRNA^{Phe}_{\underline{A}AA}$, $tRNA^{Ile}_{\underline{A}AU}$, $tRNA^{His}_{\underline{A}UG}$, $tRNA^{Asn}_{\underline{A}UU}$, $tRNA^{Asp}_{\underline{A}UC}$, $tRNA^{Cys}_{\underline{A}CC}$, $tRNA^{Ser}_{\underline{A}CU}$[146]. Except for $tRNA^{Phe}_{\underline{A}AA}$, all these tRNAs have an ADAT2/3-unfavorable pyrimidine at the 35th position, making them poor substrates for ADAT editing[34]. Furthermore, the biosynthesis of wybutosine (yW37) on guanosine 37 (G37) by TRMT5 is critical for producing stable mature tRNAs with accurate decoding capacity[147]. The ADAT2/3 editable $tRNA^{Phe}_{\underline{A}AA}$ is a particularly poor substate of TRMT5 compared to

$tRNA^{Phe}_{\underline{G}AA}$, making $tRNA^{Phe}_{\underline{A}AA}$ very unstable without yW37 and thus incapable of affecting the translation process significantly[34, 88, 148-150]. Consequently, I34-containing tRNAs that could potentially create errors are unlikely to be present intracellularly in significant quantities. This built-in safeguard helps maintain the fidelity of the translation process despite the flexibility introduced by ADAT2/3 editing.

1.8. Editing on stop codon resulting in alternative viral protein isoform

Research into A-to-I editing in the context of viral infections has intensified recently, largely due to two pivotal observations: (1) ADAR1L is upregulated by interferon (IFN) signalling and other cellular pathways related to viral infections; (2) ADAR1L targets a broad range of RNA substrates, including viral RNA genomes, viral transcripts, and host RNAs implicated in viral infections. However, there is still no consensus on the roles of ADAR editing in virus infections and virus-host interactions, with ongoing debates about whether ADAR-mediated editing serves an antiviral or proviral function. The antiviral perspective argues that ADAR-induced hypermutations in viral RNA molecules (including genomes and transcripts) disrupt the normal biological activities and functionalities of viral genes, thereby impeding viral replication and progression. Conversely, the proviral viewpoint argues that hypermutations facilitated by ADAR editing increase virus mutation rates, potentially accelerating viral evolution and adaptation. These sections will conclude research findings on ADAR editing within the context of virus infections to clarify this research field, supporting my later proposed research ideas. Because there is currently a lack of research on ADAT editing in the context of viral infections, this discussion will primarily focus on summarising findings related to ADAR editing. However, the potential for exploring ADAT editing in viral contexts remains a promising avenue for future research, highlighting an underexplored area that could yield significant insights.

Regarding the various outcomes that ADAR editing can have on virus infections, one particularly intriguing phenomenon is the creation of new viral protein isoforms as a result of ADAR editing. This has significant implications for virus infection, as it can dramatically alter the functionality of specific viral proteins. In this section, I will discuss several noteworthy examples that illustrate how ADAR editing can generate new isoforms of viral proteins by editing the stop codon of viral genes or the linker sequences between two viral ORFs.

1.8.1. Editing of a stop codon of HDV's HDAg gene

Hepatitis D virus (HDV) is an enveloped, negative-sense, circularly single-stranded RNA virus classified within the family Kolmioviridae, which does not belong to any Groups in Baltimore classification but belong to the Group named Circular single strand RNA viruses[151-153] [154, 155]. The most well-studied example of ADAR impacting viral infection and virus life cycles is the A-to-I editing on the adenosine of the Hepatitis delta antigen (HDAg) gene 'UAG' stop codon. Originally, the HDAg gene expresses the smaller isoform protein product named HDAg-S. ADAR editing converts 'UAG' stop codon of HDAg-S encoding sequence in HDV genome into 'UIG' Tryptophan (W) codon, leading to the production of a longer protein isoform named HDAg-L[33, 156-163] (Figure 11). This stop/W codon editing enables HDV to produce two distinct protein isoforms from a single open reading frame (ORF). These two proteins are fully identical in amino acid sequence except that HDAg-L includes an additional 19 amino acids at its C-terminus[33, 157, 159, 164-166]. Remarkably, HDAg-S and HDAg-L perform divergent functions, yet both are crucial for the virus life cycle and replication. Specifically, HDAg-S trans-activates HDV RNA replication and is essential for the viral replication. In contrast, HDAg-L suppresses HDV replication and assists in virion assembly by interacting with HBsAg proteins[33, 157, 159, 161, 166-172]. Typically, HDAg-S is predominantly expressed in the early stages of infection, while HDAg-L is predominantly expressed during the later stages. The differential expression of these two proteins encoded by the same ORF is regulated by the level of ADAR editing, which is itself induced by antiviral IFN signalling[163].

Additionally, HDAg-S and HDAg-L interact with each other and with both genomic and antigenomic RNAs to form a ribonucleoprotein complex (RNP)[169, 173, 174]. The RNP shuttles between the nucleus and the cytoplasm, ensuring that antigenomic RNA is accessible for editing in both compartments[175-177]. Interestingly, all of ADAR1L, ADAR1S, and ADAR2 have been found to edit the HDAg stop/W site with similar efficiencies, indicating that the expression levels of the ADAR molecules and the location of the substrate are the primary determinants of editing activity[156, 178]. And it ensures the editing on the stop/W codon because ADAR1S and ADAR2 are mildly but constitutively expressing. Surprisingly, knocking down ADAR1L does not reduce RNA editing during HDV replication, whereas knocking down ADAR1S does reduce editing, likely because both ADAR1S and HDV

antigenomic RNA are predominantly enriched in the nucleus[79]. Therefore, the editing activities of ADAR1S in the nucleus and ADAR1L in the cytoplasm, combined with dynamically changing expression levels of ADAR1L/S and HDAg-S/L, likely create a controllable dynamic environment. This environment enables a switch between the production of either HDAg-S or HDAg-L, or between early and late stages of infection, providing a regulatory mechanism for the progression and management of HDV infection.

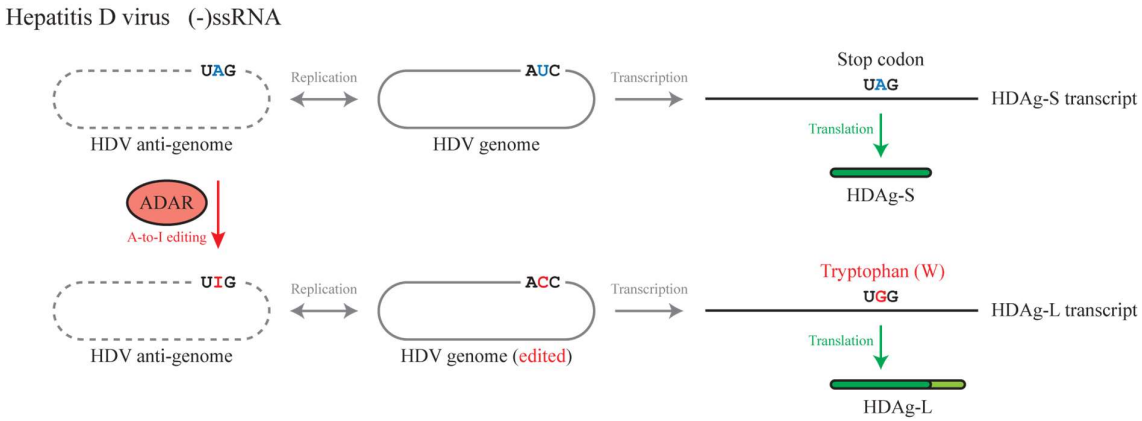


Figure 11. ADAR editing on stop codons of HDAg gene leading to expression switch of two isoforms, HDAg-S and HDAg-L. ADAR edits HDV anti-genome with UAG-to-UIG mutations, which the edited codon derives AUC-to-ACC mutation in HDV genome. Due to this mutation, the stop codon UAG of HDAg gene in the transcripts becomes UGG, resulting in Tryptophan (W) during translation. Thus, the HDAg gene expresses longer isoform HDAg-L instead of shorter isoform HDAg-S.

1.8.2. Editing of a stop codon of PIV's P gene

Parainfluenza virus (PIV) is an enveloped, monopartite, negative-sense, linear single-stranded RNA virus of the family Paramyxoviridae, classified under Group V in the Baltimore classification[179]. Different subtypes of PIV are spotted with ADAR related activities, such as PIV1, PIV3 and PIV5. Notably, similar to the HDAg gene of HDV and the GP gene of EBOV, editing events have been identified in the P gene of PIV1. ADAR editing at the stop codon of the P gene leads to a switch between two different proteins, P and L proteins, which are encoded by the same open reading frame (ORF) of the P gene[180]. Additionally, the P

gene of PIV3 undergoes extensive hypermutation due to RNA editing, resulting in a mutated P protein, although the implications of this hypermutated P protein remain uncertain[181]. Similarly, PIV5 encodes a P gene that contains two separate open reading frames (ORFs). Interestingly, neither ORF is sufficiently long to independently translate into the P protein[182]. It is proposed that specific RNA editing events may alter the linking sequence between these two ORFs, effectively merging them into one continuous ORF capable of producing the P protein[182]. Although it is suspected that these mutations may be facilitated by ADAR, this hypothesis has not yet been experimentally confirmed.

Besides the P gene, the SH gene also exhibits significant RNA editing activity. Mutations at the start codon of the SH gene can lead to a failure in SH protein expression[183], which may explain why some PIV5 strains fail to express the SH protein after infecting human cells[183, 184]. This mutation could potentially act as a regulatory switch in the PIV5 life cycle. Interestingly, when the mutated PIV5 genome that fails to express certain proteins is passaged, it often rapidly mutates back under selection pressures, indicating a strong adaptability of PIV5[183]. In addition to the P and SH genes, multiple regions in the PIV genome are characterised as hypermutated by ADAR, including the 3' UTR/C-terminus of the N gene, 5' UTR/N-terminus of the V gene, 5' UTR/N-terminus of the P gene, 3' UTR of the M gene, 5' UTR/N-terminus/3' UTR of the F gene, 5' UTR of the HN gene, and the entirety of the SH gene[183]. However, the specific outcomes and functional implications of these hypermutation events remain to be fully elucidated.

PIV5, a subtype of parainfluenza virus, exhibits significant diversity in its viral genome across different hosts. Despite extensive study, the natural hosts of PIV5 remain unidentified, although it has been confirmed to infect a wide range of species, including monkeys, humans, dogs, pigs, cats, and hamsters[185-189]. Furthermore, the diversity of PIV5 in different hosts is notable, particularly in terms of gene functions and expression patterns[179, 183]. The variations are distributed unevenly across the PIV5 genome, indicating specific regions of higher mutational activity[179, 183]. This biased hypermutation observed in the PIV5 genome may be influenced by ADAR editing, which exhibits specific affinity towards certain sequences or structures within the RNA. This editing mechanism potentially contributes to the unique patterns of genetic variation seen in PIV5 across different hosts.

1.8.3. Editing in the linker sequence between EBOV's ORFs of GP gene

Ebola virus (EBOV), or Zaire ebolavirus, is an enveloped single-stranded negative-sense RNA virus belonging to the family Filoviridae[190, 191]. A key feature of EBOV is its primary glycoprotein (GP) coding sequences, which contains two closely situated open reading frames (ORFs). The primary glycoprotein (GP) coding sequences of EBOV, containing two open reading frames (ORFs) close to each other, produces one smaller, non-structural (unglycosylated), secreted glycoprotein (SGP)[192]. The SGP lacks a transmembrane anchor sequence, which is encoded in the second ORF of the GP gene. Conversely, the GP gene also generates a larger, surface-bound, highly glycosylated glycoprotein (wild-type GP), the function of which remains unclear[192, 193]. It is suspected that the sequence between the two ORFs may be modified by ADAR editing, as guanosines are enriched in the linker sequence of the transcripts, but not in the EBOV genome. This suggests a potential ADAR-editing-mediated linkage of the two ORFs into a single complex ORF that produces GP rather than SGP[192]. Moreover, the glycan cap region (GC) and mucin-like domain (MLD) of the EBOV GP encoding sequences are also reported to be heavily edited by ADAR1[194]. While it has not been definitively proven that the switching between GP and SGP expression is triggered by ADAR editing, the A-enriched GP coding sequence supports this possibility due to its susceptibility to hyper-editing. The diversity in the coding sequences of different EBOV strains leads to significant variations in the expressed glycoprotein (GP) products[192]. The mature virion surface glycoprotein (GP) forms multimers of a single structural GP, which are crucial for the virion's binding to cell receptors and subsequent entry into the host cytoplasm[190, 191, 193]. Consequently, the extent of hyper-editing in the GP coding sequence may correlate with the severity of EBOV infections, as GP plays a vital role in the initial stages of EBOV host infection, potentially controlled by ADAR editing.

Several other coding genes in the EBOV genome are also A-enriched, experiencing substantial A-mutation pressure, especially within the coding sequences for the matrix protein (VP40), nucleoprotein (NP), glycoprotein (GP), and polymerase (Pol)[195]. Surprisingly, while EBOV infection does not significantly increase ADAR1 expression, RNAs from EBOV following several passages exhibit a high degree of editing, predominantly A-to-G substitutions, which has been confirmed to be associated with ADAR-mediated editing[194]. This observation suggests that the low level of mutations typically seen in the EBOV genome might accumulate

through serial passaging due to A-to-I editing. Similarly, the Marburg virus, which has an A-rich genome similar to that of EBOV, exhibits susceptibility in its polymerase (Pol) encoding sequence to ADAR editing[194]. This raises concerns about whether ADAR-mediated editing of Pol coding sequences could introduce mutations in critical amino acids, potentially altering polymerase function and impacting viral replication and pathogenicity.

1.8.4. Editing in linker sequence between MV's ORFs of P gene

Measles virus (MV), or Morbillivirus, is an enveloped, negative-sense, single-stranded RNA virus belonging to the family Paramyxoviridae and is classified under Group V in the Baltimore classification[196, 197]. ADAR enzymes are known to edit specific sites in the MV genome, leading to mutations that can alter viral protein structures and functions. The P gene of MV contains two open reading frames (ORFs) corresponding to two proteins: protein P and protein C[198]. Surprisingly, a third protein product is recently identified with extra length, which is caused by a guanosine insertion linking both ORFs into one ORF[198]. The origin of this guanosine insertion is suspected to be either a transcription error or the result of ADAR-mediated editing, drawing parallels to similar editing activities observed in the HDAg gene of Hepatitis D virus (HDV).

Hyper-mutated genomes have been observed during measles virus (MV) infections, particularly in cases of subacute sclerosing panencephalitis (SSPE), where ADAR-mediated editing of the MV genome was first described[199-201]. Detailed analyses of MV genomic sequences isolated from brain autopsies of SSPE and measles inclusion body encephalitis (MIBE) patients revealed that these genomes are extraordinarily hyper-mutated, predominantly showing A-to-G substitutions, equivalent to A-to-I editing events[202]. It has been established that ADAR1 can extensively edit MV genomes, with these hyper-mutations accumulating over successive passages[203, 204]. This hyper-mutation process likely contributes to the increased diversity of MV genomes and may accelerate the evolution of the virus.

1.9. RNA editing on virus genomes results in increased sequence diversity

In addition to the creation of new isoforms of viral proteins, ADAR editing can also introduce point mutations into viral genes or entire genomes, and those point mutations could be accumulated. Such hyper-editing of viral genes and genomes can increase the diversity of viral proteins and alter expression profiles, potentially impacting virus infections in different ways.

1.9.1. Hyper-editing in various genes of HIV

Human immunodeficiency viruses 1 (HIV-1) is an enveloped, positive-sense, single-stranded RNA virus belonging to the family Retroviridae, classified under Group VI in the Baltimore classification[205, 206]. HIV-1 is well known for adopting host post-transcriptional modification mechanisms such as alternative splicing, capping, and polyadenylation (poly(A) synthesis) of pre-mRNA to favour its proliferation in many aspects[207]. Research has shown that the expression of ADAR1, both in its editing-capable wild-type form and editing-incapable form with artificially mutated deaminase, facilitates the intracellular assembly of HIV-1 virions, though the editing-capable form of ADAR1 provides greater facilitation[208]. This observation suggests that ADAR1 may possess additional functions beyond RNA editing that could benefit HIV-1 infection. Additionally, ADAR1 expression also enhances the expression of the p24 antigen of HIV-1, leading to the generation of more infectious virions[208, 209]. ADAR1 has been reported to edit HIV-1 RNAs at several sites, including the 5' UTR, and the coding sequences of Env, Rev, Gag, and Tat[208, 209]. These editing-derived mutations not only facilitate the expression of these genes but also enhance their functionality, leading to increased viral replication and gene expression[208, 209]. Transcripts modified by ADAR1 are expressed more efficiently than their wild-type counterparts. Conversely, the knockdown of ADAR1 results in reduced HIV-1 production and proliferation[209]. Moreover, ADAR1 knockdown cells show a decrease in level of unspliced RNA transcripts, caused by errored alternative splicing[209]. For instance, the expression levels of Gag protein precursor p55 and Gag protein p24 are significantly lower with ADAR1 knockdown[209]. These findings imply the critical role of ADAR-mediated RNA editing as a regulator in the HIV-1 lifecycle,

influencing various stages from gene expression and protein synthesis to viral assembly and release.

In addition, HIV-1 non-coding RNA sequences such as the Rev responsive element (RRE), trans-activation responsive element (TAR), and the dimerisation domain (DIS) all feature robust double-stranded structures that are predicted to serve as ideal substrates for ADAR1[207, 209-211]. These non-coding sequences play crucial roles in the replication process of the virus, and their modification through RNA editing could lead to significant changes in their normal functions. The HIV-1 genome encodes a positive-regulator protein named Tat, which can facilitate HIV-1 replication[212-215]. The activation of Tat is dependent on a cis-acting element known as the transactivation response (TAR) site[216, 217]. This TAR site can form a specific stem-loop structure, which is vital for the activation of Tat[216-218]. Notably, the TAR site is found to be heavily edited enriched with inosines, suggesting A-to-I editing may participate in TAR site activation besides up-regulating Tat protein expression[209, 219].

1.9.2. Hyper-editing in HDV anti-genomes

In addition to specific editing of the HDAg stop codon by ADAR enzymes in Hepatitis D virus (HDV), both ADAR1 and ADAR2 also facilitate nonspecific RNA editing across the HDV antigenome, which has been shown to inhibit HDV replication[178, 220]. Although the mRNA of HDV is also edited, but significantly more abundant editing events are found in the HDV anti-genome[159]. This discrepancy may be attributed to the mRNA's instability and its relatively short lifespan, or potentially due to its limited ability to form secondary structures, which are less prevalent in shorter RNA sequences[158, 159, 221]. Furthermore, different strains of HDV exhibit variations in nonspecific ADAR-mediated editing, likely due to sequence variability that affects RNA structure. These differences in genotype are not only functional but also genetic[222]. HDV genotype I is the most widespread and extensively studied; however, genotype III is associated with more severe disease outcomes and is genetically most distinct from other genotypes[223, 224]. Notably, the antigenomic RNA structure of HDV genotype III is more susceptible to ADAR editing compared to genotype I. This increased susceptibility is due to genotype III's tendency to develop more structurally

favourable hairpin-like and unbranched rod-like formations that are more effective at recruiting ADAR enzymes[225].

While ADAR-mediated editing of viral transcripts or genomes has significant impacts during virus infections, evidence suggests that ADAR also influences viral infections in an editing-independent manner. For instance, investigations into HDV antigenome editing have revealed that even ADAR1 and ADAR2 variants incapable of editing can slightly inhibit HDV replication[178]. This observation suggests the involvement of additional molecular mechanisms in this inhibition beyond RNA editing. Furthermore, the expression of both editing-capable and editing-incapable ADAR1 variants has been shown to increase the levels of gp120, p24, and Nef viral proteins. This finding implies that ADAR1 may facilitate the synthesis of these viral proteins through a mechanism independent of its RNA editing activity[208].

1.9.3. Hyper-editing in Influenza A virus genome and transcript

Influenza A virus is an enveloped, negative-sense, single-stranded RNA virus of the family Orthomyxoviridae, classified under Group V in the Baltimore classification[226, 227]. Influenza A viruses are differentiated into various subtypes based on the combinations of their surface glycoproteins, hemagglutinin (HA) and neuraminidase (NA)[226, 228-233]. The increase in A-to-I editing level in influenza A virus (H1N1 and H3N2) infected cells are caused by up-regulated ADAR1 but not ADAR2[228]. ADAR1 has the capability to edit influenza A virus genomes, and these hyper-edited viral genomes can alter the base composition and structural configuration of the viral RNA[203, 234-237]. The hyper-editing of viral genomes or transcripts can lead to the inhibition of the normal function and replication of the influenza A virus genome[203]. It has been noted that the frequency of hyper-edited mutants in influenza A virus is higher compared to other viruses, such as Measles virus, suggesting that different viruses, or even different strains of the same virus, may exhibit varying sensitivities to ADAR1 editing[203]. Furthermore, many commercial vaccines, including those for influenza, are produced using chicken embryo fibroblasts, which also encode ADAR1-like proteins[238, 239]. This production process can result in influenza virus vaccines being hyper-edited by chicken ADAR1-like proteins, potentially leading to vaccine attenuation[203]. Additionally,

the highly pathogenic avian influenza virus strain H5N1, when infecting chickens, is enriched in ADAR A-to-I editing, which may contribute to the evolution of H5N1 by increasing the diversity of its quasispecies[240].

During influenza A virus infection, the non-structural protein 1 (NS1) of the virus is known to interact with ADAR1 through a protein-protein interaction, forming a molecular complex[241, 242]. This NS1-ADAR1 complex has been reported to antagonise the cellular interferon (IFN) response, a key defence mechanism against virus infections[243]. This interaction potentially provides an advantage to the influenza virus by helping it evade IFN-inducible antiviral activities. It worth doubting whether this NS1-ADAR1 interaction correlates with ADAR1 dimerisation property, allowing ADAR1 to form dimer with viral protein like NS1.

1.9.4. Hyper-editing in various genes of LCMV

Lymphocytic Choriomeningitis virus (LCMV) is an enveloped, multipartite, negative-sense, linear single-stranded RNA virus from the family Arenaviridae, classified under Group V in the Baltimore classification[244]. Notably, the coding sequences for the glycoprotein (GP), nucleoprotein (NP), RNA-dependent RNA polymerase (RdRp), and Z protein in LCMV are found to be hyper-edited by various ADAR molecules. However, the specific roles of different ADAR molecules in these editing events remain unclear. Hyper-mutation within the GP gene of LCMV can lead to a loss of function in the expressed GP protein[245]. This dysfunction is likely caused by the accumulation of altered amino acids, which can result in misfolding of the GP protein. Similar editing outcomes, characterised by accumulated mutations leading to protein dysfunction, have also been observed in other proteins such as NP, RdRp, and Z proteins[245]. Although the broader implications of these hyper-mutations on the coding sequences have not been extensively studied, it is understood that the majority of these mutations are likely to be eliminated by natural selection.

LCMV features two segmented RNA genomes, both of which naturally form intergenic loops and stems, providing a secondary structure that is particularly susceptible to ADAR-mediated editing[244-246]. This structured genome is favourably targeted by ADAR enzymes, resulting in extensive hyper-mutation through A-to-G substitutions within the LCMV genomic

RNA[245]. Moreover, LCMV replicates exclusively in the cytoplasm, and given this localisation, it is predominantly the cytoplasmic ADAR1L, rather than the nuclear ADAR1S or ADAR2, that is responsible for the hyper-editing of the LCMV genome[245]. Mutations induced by ADAR1L in the viral RNA can lead to the loss of function in viral proteins and a consequent reduction in viral infectivity[245]. However, this same hyper-mutation process can also fuel the creation of a more infectious virus by increasing the rate of evolution within the LCMV genomes. This dual impact highlights the complex role of RNA editing in viral pathogenesis, where the balance between detrimental and beneficial mutations can significantly influence the viral life cycle and its adaptation to host defences.

1.9.5. Hyper-editing in RSV's GP gene

Respiratory syncytial virus (RSV) is an enveloped, monopartite, negative-sense, linear single-stranded RNA virus classified within the family Pneumoviridae, under Group V in the Baltimore classification[247, 248]. The G glycoprotein of RSV is crucial for virus binding to cell surface receptors, facilitating the entry of RSV virions into host cells[249]. Notably, the G glycoprotein exhibits the highest degree of antigenic and genetic diversity among virus isolates[250-258], largely due to hyper-mutation in the glycoprotein sequences generated by ADAR editing[259]. This diversity in the G glycoprotein allows some viral variants to evade binding by specific anti-G antibodies[259], highlighting the potential for altered infection outcomes driven by mutations in this protein. Generally, the A-to-G mutations induced by ADAR are concentrated in the conserved regions of the hyper-mutated virus, which are the primary targets of RSV-specific antibodies[259]. The hyper-edited regions are characterised by an A-rich composition and features of double-stranded RNA structures such as loops, bulges, and short stems[259]. Interestingly, the flanking segments of the conserved regions are more susceptible to editing than the central part, which central part remains relatively unchanged and truly conserved[259]. The hyper-mutations on G glycoprotein leading to its functional changes can generate more profound RSV infections, which eventually leads to more increased severity of RSV infections[260, 261]. It is suspected whether the mutations enriched in the conserved region is the consequence of natural selection, which mutations in non-conserved region does not show significant impact much on the viral survival and

replication. Thus, understanding these mutation patterns may also provide insights into the evaluation of RSV evolution rate.

1.9.6. Hyper-editing in NoV's VP1 gene

Norovirus (NoV) is a non-enveloped positive-sense single-stranded RNA virus of the family Caliciviridae, under Group IV in the Baltimore classification[262]. An interesting aspect of NoV biology is the high level of RNA editing observed in the viral transcripts encoding VP1, a major structural protein of the virus. These transcripts exhibit notable U-to-C substitutions, which is later found derived from A-to-I editing in virus genomes[235]. While the high diversity and mutation rate typical of RNA viruses are primarily due to the low fidelity of RNA-dependent RNA polymerases, regions of particularly high mutation within viral genomes are often the result of targeted RNA editing by enzymes such as ADAR[263, 264]. Furthermore, the Norovirus genome is known for its exceptional genetic diversity[265-267]. The analyses of stool samples from NoV infected patients have revealed low-frequency genome sequences with a high level of A-to-G or U-to-C base substitutions, contributing to the overall diversity of the NoV genome[235]. This genetic variability is crucial for the viral ability to evade host immune responses and adapt to changing environmental pressures.

1.9.7. Hyper-editing in hTLV genome

Human T-lymphotropic virus (hTLV) is an enveloped, positive-sense, single-stranded RNA virus from the family Retroviridae, classified under Group VI in the Baltimore classification[268-270]. Human T-lymphotropic virus type 2 (hTLV-2) genome is found highly edited in vitro, but probably remains a rare phenomenon in vivo[271]. Additionally, in peripheral blood mononuclear cells (PBMCs) from hTLV-2 infected patients, there is no significant increase in A-to-I editing levels within the virus genome. However, highly edited transcripts have been identified in cells infected with hTLV-2, indicating that the viral transcripts in these cases are more susceptible to ADAR-mediated editing[271]. Similar phenomenon is found with similarly originated virus Simian T-lymphotropic virus type 3

(sTLV-3)[271-273]. Despite these observations, the functional outcomes of ADAR editing on T-lymphotropic viruses remain unclear.

1.9.8. Hyper-editing in various genes of MuV

Mumps virus (MuV) is an enveloped, monopartite, negative-sense, linear single-stranded RNA virus categorized under Group V of the Baltimore classification, within the family Paramyxoviridae[274, 275]. Notably, the P gene, V gene, and I gene of MuV are found to be hyper-mutated in viral mRNA transcripts, a phenomenon that is attributed to ADAR editing[276]. These hypermutations significantly impact the coding and functional capacities of these genes, reflecting the influence of RNA editing mechanisms on viral gene expression.

1.10. Editing on other RNA molecules and specific regions causing various subsequential impacts

Interestingly, beyond the editing of viral protein-coding genes by ADAR, other RNA molecules or specific regions can also be targets of ADAR editing, significantly impacting virus infections. This section will discuss the editing of these other RNA molecules and specific regions, highlighting their potential roles in modulating viral replication, translation, and host-virus interactions.

1.10.1. Editing on Zika virus reshape codon usage biases

Zika virus (ZIKV) is an enveloped, monopartite, linear positive-sense single-stranded RNA virus from the family Flaviviridae, classified under Group IV in the Baltimore classification[277-279]. The ZIKV's genome encodes a single polyprotein that is subsequently cleaved into three structural and seven non-structural proteins[236, 278]. The polyprotein coding sequences of ZIKV exhibit a significant enrichment in A-to-G mutations, as revealed by RNA sequencing, which are likely to be the result of ADAR-mediated editing[280]. Interestingly, the codon usage within the ZIKV genome is characterised by a high frequency of 'A-ended' codons, which are preferred targets for ADAR editing[280]. Compared to 'G-ended' codons, 'A-ended' codons are considerably less common in the human codon pool[146]. Thus, ADAR editing of 'A-ended' codons may enhance the viral ability to adapt to the human codon pool, potentially facilitating more efficient viral gene expression[280]. The role of ADAR-mediated editing in inducing mutations on viral genomes is believed as a significant evolutionary force among RNA viruses, as it contributes to increased viral mutation rates[280]. This is supported by evidence suggesting that the percentage of guanosine in ZIKV coding sequences has increased over the course of Zika virus evolution[236]. This trend may indicate that ADAR editing not only facilitates the adaptation of the virus to its human host by favouring 'G-ended' codons, which are more compatible with the human codon pool, but also plays a crucial role in driving the evolutionary dynamics of ZIKV.

When ZIKA hijack host translational machinery to synthesise viral proteins, the use of rare codons during translation will cause translation pause[235, 281]. These pauses occur because

the host ribosome stalls when encountering clusters of rarely used codons, significantly slowing down the translation process[282]. Additionally, the single-stranded RNA of ZIKV can form G-quadruplex structures that may also induce stalling of RNA-dependent RNA polymerase during viral replication[283]. Following such translational pauses, the viral replication process initially slows down then the rate will be resuming at normal rates[284]. During such translational pauses, the upcoming 3'-directed RNA fragments are exposed to oxidative damage or RNA editing[236]. It has been established that the sequences 3'-directed to the rare codon stretch are more susceptible to editing by ADAR[236]. This susceptibility is also influenced by the secondary structure of these 3'-directed sequences, which enhances their ability to attract ADAR and undergo editing. It would be surprising if the 3'-directed codons are more efficiently expressed after edited by ADAR, because it could be a mechanism for viruses adapting human host transcription/translation machinery. However, further investigation is needed to confirm this hypothesis.

1.10.2. Hyper-editing in EBV's FR region

Epstein-Barr virus (EBV), also known as human gamma-herpesvirus 4, is an enveloped, double-stranded DNA virus of the family Herpesviridae, classified under Group I in the Baltimore classification[285, 286]. EBV is particularly notable for its ability to switch between latent and lytic phases, a process essential for its lifecycle and pathogenesis. During the reactivation from latency, the EBV latency origin of replication (oriP) is transcribed bidirectionally, producing transcripts both leftward (oriPtLs) and rightward (oriPtRs)[287-290]. Interestingly, these transcripts, particularly around the Family of Repeat (FR) regions, have been identified as hyper-edited by ADAR1[287]. This editing is significant because specific knockdown of oriPtLs has been shown to contribute to the activation of EBV lytic gene expression and viral DNA replication, facilitating the switch of EBV into its lytic phase[287]. This suggests that ADAR1-mediated RNA editing may play a critical role in modulating the latency-to-lytic switch in EBV. The FR region of EBV, where these oriPtLs and oriPtRs are located, is characterised by stable double-stranded hairpin secondary structures, which are likely targets for ADAR1 due to their double-stranded nature[287, 291, 292]. These features are highly conserved across different strains of EBV, indicating that the

virus may have evolved to exploit ADAR1 editing as a mechanism to enhance its adaptability and survival within the host[287]. However, the mechanism of ADAR1 editing on oriPtLs affecting the outcome of EBV infection is still not clear[287]. Besides, A-to-I editing events have also been observed in other regions of the EBV genome, including the BHRF1 gene and the miRNAs of BamHI A rightward transcripts (BART)[287, 293-295]. The functional outcomes of these editing events are still under investigation but could potentially influence EBV's pathogenicity and its interactions with the host immune system.

1.10.3. Editing in MPV's Defective interfering RNA

Metapneumovirus (MPV) is an enveloped negative-sense single-stranded RNA virus of the family Pneumoviridae, classified under Group V in the Baltimore classification[296, 297]. A special type of RNA named Defective interfering RNA (DIs), which are derivatives of the viral genome that become partially deleted and spontaneously accumulate in virus stocks during passaging[298]. This accumulation is attributed to errors made by the viral replicase complex[298]. MPV is known to accumulate an excessive quantity of DIs during passaging, and these DIs act as potent inducers of antiviral interferon (IFN) responses[299]. Interestingly, these DIs undergo extensive mutations by ADAR, with up to 70% of adenine bases being edited into inosine[299]. It has been observed that these heavily mutated DIs exhibit an attenuation in IFN inducibility[299]. This suggests a potential molecular mechanism by which MPV may evade IFN induction and subsequent host antiviral responses, when hypermutations are accumulated[299]. In the context of MPV infection, both ADAR1L and ADAR1S are expressed in infected cells, with ADAR1L being inducible by IFN[299]. Therefore, the dynamics among ADAR expression, IFN inducibility and DI editing still needs further investigating, and interactions between MPV and antiviral defence system is more profound than that. Additionally, virus infections are known to activate innate immune pathways, including the RIG-I-like receptor (RLR) pathway involving RIG-I and MAVS, which leads to the production of IFN[300-302]. However, the G protein of MPV has been reported to interact with RIG-I, resulting in the inhibition of this crucial antiviral activation pathway[303]. This interaction further illustrates the profound strategies employed by MPV to manipulate host

cellular mechanisms to its advantage, highlighting the complicated dynamics of virus-host interactions during MPV infection.

1.10.4. Editing on HCV's IRES

Hepatitis C virus (HCV) is an enveloped, positive-sense, single-stranded RNA virus belonging to the family Flaviviridae and classified under Group IV in the Baltimore classification[304, 305]. ADAR1, by increasing the A-to-I editing level of the internal ribosome entry site (IRES) of HCV, can lead to malfunctioning of IRES, which in turn inhibits HCV replication[306]. The IRES element is crucial for the initiation of translation in HCV; thus, any modification that impairs its function can significantly disrupt the viral life cycle. The specific mechanisms through which hyper-edited IRES-containing transcripts affect HCV replication remain unclear. However, it has been observed that the introduction of exogenous RNAs that compete for ADAR1 editing can counteract the inhibition of HCV replication[306].

1.11. Other results related to ADAR regulations

In fact, not all the impacts brought by ADAR proteins are based on editing-derived mutations. ADAR proteins contain other functional domains that can significantly influence virus infections through editing-independent mechanisms. This section will discuss the editing-independent impacts of ADAR proteins on virus infections, illustrating their multifaceted roles in modulating viral pathogenesis.

1.11.1. Regulations in VV

Vaccinia virus (VV) is an enveloped, monopartite, linear double-stranded DNA virus of the family Poxviridae, categorised under Group I in the Baltimore classification[307]. VV has developed complex mechanisms to evade host immune responses, particularly through the action of its E3L protein. The E3L protein of VV is crucial for inhibiting the antiviral effects of type I IFN, and he IFN-countered E3L is found to inhibit A-to-I editing functions of both IFN-inducible ADAR1L, and ADAR1S[308]. This inhibition allows VV to replicate effectively even in cells that are actively responding to IFN signalling[309-312]. Moreover, VV's replication efficiency is significantly reduced in the absence of the N3L protein, indicating the importance of these viral proteins in counteracting IFN-mediated antiviral defences[311]. The interaction between E3L and ADAR1 is particularly notable. It has been identified that only the dsRNA binding domain of E3L and the third dsRNA binding domain of ADAR1 are crucial for the inhibition of ADAR1's editing functions[308, 309, 311]. This specific interaction suggests that E3L may inhibit ADAR1 by forming dimers with it, resulting in editing incapable E3L/ADAR1 complexes. This hypothesis points to a direct interaction mechanism where E3L could physically block ADAR1 from accessing dsRNA substrates, thereby inhibiting its editing activity. In addition to its effects on ADAR1, E3L also inhibits PKR (Protein Kinase R), a kinase involved in another antiviral signalling pathway, probably through an indirect mechanism[313-315].

1.11.2. Regulations in Adenovirus

Adenovirus, a non-enveloped, monopartite, linear double-stranded DNA virus from the family Adenoviridae, is categorised under Group I in the Baltimore classification[316, 317]. A key element of its viral machinery is adenovirus-associated RNA (VA RNA), a small non-coding RNA synthesised by adenoviruses during infection. Found in infected cells as part of replicative intermediates, VA RNA plays a critical role in the virus life cycle[318]. This long dsRNA is found to be edited by ADAR1, which can be utilised as a negative regulator of ADAR1[319]. The editing of VA RNA by ADAR1 is significant as it suggests that VA RNA may act as competitive substrates that consume the host's ADAR editing capacity, thereby protecting other viral RNAs from being edited. This mechanism serves as a strategic viral evasion technique to prevent the modification of viral RNA that could potentially inhibit viral function or enhance immune recognition. Moreover, the interferon-inducible ADAR1L can edit directly on adenovirus-associated RNAs as well as hepatitis C virus (HCV) RNAs[306]. The presence of adenovirus-associated RNAs in cells infected concurrently with HCV has been shown to compete for ADAR1L, resulting in a decrease in the level of inosines in HCV transcripts[306]. This competition can attenuate the editing burden on HCV RNAs, subsequently aiding the recovery of HCV replication and proliferation[306]. This interaction between adenovirus and HCV highlights the complex dynamics of co-infections, where the presence of one virus can significantly impact the biology of another by modulating the host's cellular machinery.

1.11.3. Regulations in Adenovirus

Polyoma virus (PV) is a non-enveloped, monopartite, circular double-stranded DNA virus from the family Polyomaviridae, classified under Group I in the Baltimore classification[320]. During the late stage of PV infection, there is a suppression of early-stage viral genes. This suppression is mediated by certain nuclear RNAs, which are derived from the antisense strand of the PV genome itself[321]. Further analysis of these regulatory RNAs has revealed that they are hyper-mutated and enriched with inosines and guanosines, suggesting extensive editing by ADAR[321]. This high level of editing by ADAR in the regulatory RNAs of PV indicates a

profound interaction between the virus and the host's RNA editing machinery. It is still unclear right now how do those hyper-editing events on regulatory RNA molecules affect virus infection but it provides another insight of indirect mechanism for ADAR to participate and facilitate viral life cycle and replication[321].

1.12. Discussion

Chapter 1 of the thesis provides a detailed summary of the intracellular molecular mechanisms of RNA editing mediated by ADARs and ADATs as well as the roles of ADAR-mediated editing in virus infections. In general, the interest of ADAR editing in virus infections are mainly based on (1) the functionality of generating A-to-I mutations (A-to-G synonymous mutations) in virus related RNA molecules, which alters the outcomes of virus infections; (2) ADAR1L is IFN-inducible and potentially participating the host anti-virus mechanism. Based on the previous research, it is no surprising that IFN-inducible ADAR1L is the major contributor of RNA editing in virus infections although some findings involved other ADARs like ADAR1S and ADAR2. Despite extensive studies on ADARs, there remains a notable gap in understanding the specific roles of ADATs in the context of viral infections. This chapter raises questions about the potential functions and implications of both ADAR and ADAT editing in viral infections, highlighting critical gaps in current research.

1.12.1. Summarising roles of RNA editing and ADAR regulations in virus infections contributes to understanding the knowledge gap in this field

Firstly, I explore the phenomenon of hyper-mutations induced by ADAR editing across various viral infections. These mutations are identified in both viral transcripts and genomes, resulting in various consequential effects on viral biology. Such hyper-mutations can significantly influence viral pathogenicity, replication strategies, and interactions with the host immune system, thereby affecting their virology. A significant focus within this research area is on finding critical mutations that can directly and obviously alter viral virology. For instance, the editing of stop codons in viral genes by ADAR can extend or modify protein sequences, potentially changing the viral behaviour and interaction with the host. Similarly, non-synonymous mutations introduced by ADAR can lead to altered viral protein functions, affecting the viral pathogenicity, replication efficiency, and immune evasion strategies. One inspiring example of the unusual implications of ADAR editing is observed in the Zika virus (ZIKV), where A-to-I editing converts A-ended codons into I-ended codons, which functionally mimic G-ended codons. This alteration can lead to an improved adaptation of the

viral genome to the codon usage preferences of the human host, potentially enhancing the viral translational efficiency and overall fitness in human cells. This observation underscores the need for further research into how ADAR editing influences codon usage biases in viral genomes. The case of ADAR-mediated editing in ZIKV invites a broader examination of the roles of RNA editing in reshaping virus codon usage biases, potentially revealing new dimensions of virus-host interactions.

Secondly, the discussion on the lack of research concerning ADAT editing in virus infections reveals a significant gap in the field. ADAT enzymes, known for their role in modifying tRNA molecules, could substantially affect the tRNA pool of the human host. This alteration may impact the translational efficiency of viral proteins, particularly in relation to the specific codon usage biases of viruses. Such changes in translation dynamics are critical for understanding how viruses optimise replication within their host's cellular environment. Furthermore, an interesting finding from my previous research project indicated that the promoters of the ADAT1 and ADAT2 genes were significantly activated in host cells infected by various viruses[3] (Appendix 5). This observation, while preliminary and not extensively studied, suggests that viral infections may influence ADAT expression through mechanisms that are not yet understood, whether they are mediated directly by the virus or as a response by the host. The activation of ADAT promoters during viral infection implies at a potentially critical role for ADAT-mediated tRNA modifications in context of virus infection. These findings align with the proposed knowledge gaps, especially concerning how viruses adapt their codon usage to optimise replication. By investigating deeper into virus codon usage biases and their correlation with host tRNA supplies, future research could uncover significant insights into the molecular dynamics of virus-host interactions.

Finally, the induction of IFN-inducible ADAR1L during various viral infections highlights its critical role in the antiviral defence mechanisms mediated by interferon (IFN) signalling. Given the broad-spectrum antiviral response triggered by IFN signalling, it is expected that ADAR1L would be up-regulated as part of this defence strategy. Unlike other ADAR enzymes such as ADAR1S and ADAR2, which exhibit higher specificity, ADAR1L is characterised by its relatively low specificity and semi-random editing patterns. This feature allows ADAR1L to counteract the diversity of exogenous RNA molecules generated during viral infections, potentially modifying a wide range of RNA substrates. Moreover, given the semi-random

nature of ADAR1L editing, it raises an interesting possibility: ADAR1L may not only target viral RNAs but could also edit host transcripts, potentially leading to altered outcomes in virus infection. IFN, which are cytokines released by host cells in response to viral infections, stimulate the expression of IFN-inducible genes like ADAR1L not only in infected cells but also in neighbouring uninfected cells[322, 323]. This widespread induction could lead to extensive editing activity across the cellular transcriptome. I hypothesise that the over-expression of ADAR1L, induced by IFN signalling, might edit transcripts of host genes that are crucial for viral entry. For example, editing of the mRNA encoding cellular receptors like Angiotensin-converting enzyme 2 (ACE2), which is critical for SARS-CoV-2 entry, could potentially modify receptor functionality and influence the overall viral infection dynamics.

1.12.2. Discussing research interest in RNA editing and its impacts on virology and proposing thesis hypothesis

The discussions thus far have highlighted three significant knowledge gaps in the current understanding of RNA editing mechanisms within viral contexts. (1) Current knowledge on how ADAR editing influences viral codon usage biases remains limited; (2) Similarly, the contributions of ADAT editing during viral infections are poorly characterised; (3) There is no evidence supporting the hypothesis that overexpressed ADAR1L, a component of the host's antiviral response, may edit host genes involved in viral infection processes, potentially altering their function in virus infections. The knowledge gap (1) and (2) are both related to virus codon usage. Thus, by studying virus codon usage biases may provide clues in roles of ADAR and ADAT editing in virus infections.

To address the knowledge gaps identified in (1) and (2), I hypothesise that there are distinguishable differences in virus codon usage biases between viruses infecting humans and those infecting non-human hosts. These differences may be indicative of the specific roles played by ADAR and ADAT editing mechanisms in shaping viral adaptation to human hosts. Accordingly, I propose a detailed examination of virus codon usage biases across a diverse range of viruses and their potential correlations with ADAR and ADAT editing activities through systematic statistical methods. This statistical research will not only fill the existing knowledge gaps but also enhance our understanding of the molecular interactions between

viruses and their hosts, potentially leading to novel insights into viral pathogenesis and host defence mechanisms.

To take a step further in the study of virus codon usage biases, I hypothesised that the systematic interpretation of virus codon usage biases in different viruses of different host ranges could be systematically studied with machine learning modelling. And by analysing predictions computed by trained machine learning model, I could develop bioinformatics tools to monitor virus codon usage fitness in the specific host range, which is, in another term, to statistically describe virus codon usage fitness in the specific host.

Different from knowledge gap (1) and (2) both related to virus codon usage biases, the knowledge gap (3) demonstrated the potential ADAR1L editing in host genes which are critical to virus infections. According to the literature review, it is found that the research interest has heavy focus on the editing in viral RNA and the impacts brought by those editing. Because the randomness in substrate selection and IFN-inducibility of ADAR1L, I hypothesise that the host gene may be optimally or sub-optimally edited by over-expressed ADAR1L causing changes in regular gene functionality. For example, SARS-CoV-2 significantly relied on protein-protein interaction (PPI) between its Spike protein and host receptors ACE2 and TMPRSS2 to infect host cells. If the IFN-induced ADAR1L in target host cells edits host ACE2 or TMPRSS2 transcripts, it might affect the functionality of ACE2 and TMPRSS2, and eventually lead to impacts on SARS-CoV-2's entry. Therefore, I propose to evaluate changes of PPI between Spike, ACE2 and TMPRSS2 under ADAR1L over-expression to study potential roles of ADAR editing as host anti-virus or potential pro-viral mechanism.

In summary, there are three aims in this thesis: (1) to study virus codon usages biases and their relationships to ADAR and ADAT editing; (2) to study virus codon fitness to specific host range via machine learning methods; (3) to study changes in PPI between SARS-CoV-2 Spike protein, human ACE2 and human TMPRSS2 under ADAR1L editing. By studying these three aims, we could have a better understanding in roles of RNA editing in context of virus infections, which is potentially beneficial to virus outbreak prevention, vaccine development and therapeutic development.

1.13. Proposed hypothesis

Hypothesis: ADAR and ADAT editing affects viral gene expressions and virus entry in virus host interactions

- **Aim 1**: To investigate impacts of ADAR and ADAT editing on virus codon usage biases by statistically comparing between human and non-human viruses
- **Aim 2**: To systematically evaluate viral codon fitness in certain host through Random Forest modelling
- **Aim 3**: To establish in-vitro molecular assay to quantitatively detect protein-protein interaction between SARS-CoV-2 Spike and human receptors ACE2 and TMRPSS2, and to study impacts of ADAR editing on host receptors and virus entry

Aim 1 is addressed in chapter 2, whereas Aim 2 is addressed in chapter 3 and the under-review manuscript in Scientific Reports (Manuscript ID: 400c1a00-9297-40bf-8a4b-d7d2bbda0d64 v1.0). Finally, Aim 3 is addressed in chapter 4.

CHAPTER 2

RNA Editing and Virus Codon Usage Biases

by Shuquan (Steve) Su

Keywords:

RNA editing, Adenosine deaminase, Virus codon usage biases

Chapter 2. RNA Editing and Virus Codon Usage Biases

2.1. Introduction

2.1.1. Introduction of RNA editing and codon usage biases reveals potential relationships

As highlighted in Chapter 1, there are notable knowledge gaps in research regarding impacts of RNA editing on virus codon usages within the context of virus infections. Specifically, the effects of synonymous mutations, which are often overlooked due to their lack of direct influence on the amino acid sequences of viral proteins, may play a critical role by altering codon usage patterns. These synonymous mutations, while not altering the protein sequence, can significantly impact the codon usage biases. Codon usage bias refers to the preferential use of specific codons over others that encode the same amino acid. This bias can significantly influence the efficiency of protein translation within a host cell. Variations in codon usage among viruses and their host cells can affect viral gene expression, potentially impacting the viral ability to proliferate and adapt to the host environment. RNA editing can potentially introduce synonymous mutations that alter the codon composition of viral genomes. These alterations can shift the codon usage biases of the viral coding sequences, potentially affecting the translational dynamics of viral genes within the host. Such changes might lead to variations in the efficiency of protein synthesis, ultimately influencing viral replication and pathogenicity. To better understand the impact of RNA editing on viral infections, it is essential to consider how edited codons influence virus genome codon usage biases. By expanding our research focus to include the effects of RNA editing on codon usage biases, we can uncover critical insights into the viruses' adaptive mechanisms and potentially identify novel targets for antiviral therapies.

As discussed in Chapter 1, ADAT2 and ADAT3 are known to perform tRNA A-to-I editing at the wobble position of adenosine, which enhances the pairing flexibility of tRNAs. Specifically, research has shown that tRNAs corresponding to eight amino acids including Arginine (R), Alanine (A), Serine (S), Leucine (L), Proline (P), Isoleucine (I), Threonine (T), and Valine (V), which are targets for ADAT2-ADAT3 dimer editing activities[87, 89, 92, 138]. Such editing modifies the host's tRNA pool, potentially altering the translational efficiency of both host and viral genes. Furthermore, my prior research demonstrated that the promoters of the ADAT1 and ADAT2 are significantly induced and activated during various virus infections[3]. Despite this induction, the direct impacts of ADAT-mediated tRNA editing on viral pathogenesis remain largely unexplored. By analysing the codon usage biases in virus genomes and specific genes, we can gain insights into the possible effects of ADAT editing. This approach could reveal how shifts in tRNA pool composition influence the translation of viral proteins, potentially affecting virus replication and host-virus interaction dynamics.

In Chapter 1, as highlighted, the tRNAs subject to ADAT editing are predominantly expressed compared to other non-edited tRNA (Figure 10). ADAT-mediated editing expands the codon pairing capabilities of these tRNAs, potentially affecting translational efficiency for their native codon matches. This effect arises because the edited tRNAs begin to accommodate additional codons, effectively 'sharing' their availability with a broader set of codons, which possibly dilute their presence for the original codon. Consider the $tRNA^{Ala}_{\underline{A}GC}$, which typically pairs with the GCU codon. Once edited by ADAT to $tRNA^{Ala}_{\underline{I}GC}$, it can also pair with GCC and GCA codons. This editing implies that the GCU codon now 'shares' its tRNA with GCC and GCA. Theoretically, this sharing could reduce the translational efficiency for GCU due to increased competition for the tRNA among these codons. To systematically analyse the implications of ADAT editing on translational dynamics, it is useful to categorise codons based on their interactions with edited tRNAs. ADAT suppress codons are codons like GCU, which may experience decreased translational efficiency due to their tRNAs being shared with additional codons post-editing. The term 'suppress' reflects the potential negative impact on translation efficiency for these codons. ADAT benefit codons are codons like GCC and GCA, which gain additional tRNA pairings due to ADAT editing. The increase in available tRNAs for these codons might enhance their translational efficiency. Other codons do not interact directly with tRNAs altered by ADAT and are therefore unaffected by ADAT editing. Their

translation is not influenced by the expanded codon pairing capacities and thus remains consistent with non-edited states.

2.1.2. Introducing metrics of codon usage biases and relationships to virus codon adaption studies

Given that different tRNA species exhibit varying levels of abundance, changes in codon usage can theoretically influence the translational efficiency of coding sequences (CDS)[324-326]. This aspect is particularly critical in the context of viral genes, where translational efficiency plays a decisive role in virology and proliferation. If the codon usage of a viral gene is markedly different from that of its host cell, indicating poor adaptation or fitness to the host's translational system, the cellular antiviral mechanisms may effectively suppress viral replication[327]. This misalignment can affect the viral ability to efficiently utilise the host's translational machinery, thereby impacting its capacity to produce essential proteins for replication and assembly. Analysing virus codon usage biases directly from viral gene or genome sequences is essential for understanding how viruses proliferate and adapt to human hosts. By systematically comparing the codon usage patterns of viruses that infect humans with those that infect non-human hosts, I can identify distinct codon usage characteristics that are potentially advantageous for viral adaptation to human cells.

The study of virus codon biases is enriched by various metrics, including Relative Synonymous Codon Usage (RSCU)[328, 329], Codon Adaptation Index (CAI)[329], and tRNA Adaptation Index (tAI)[330], among others. Each of these metrics provides unique insights into the differences of codon preference and efficiency in the context of viral and host interactions. However, many of these metrics require the gene expression levels of host genes as a reference, which can introduce biases when scaling up to species-wide comparisons. RSCU is a key parameter used to measure the frequency of synonymous codons used in a gene, normalised against a uniform usage of all synonymous codons for each amino acid. This metric is calculated purely from the coding sequence data, avoiding biases associated with external computational inputs. RSCU preferences have been studied in individual viruses including SARS-CoV-2[331], Flaviviridae Virus[332], Zika virus[333], and Transmissible Gastroenteritis Virus[334]. Despite the extensive use of RSCU in analysing individual viruses,

there is a significant gap in our understanding of how these codon usage patterns collectively relate to the broader context of viral host ranges.

In Chapter 2, I propose a comparative analysis of the Relative Synonymous Codon Usage (RSCU) values of virus genomes, specifically between viruses that infect humans and those that infect non-human hosts. This investigation will employ various statistical methodologies to elucidate distinctive characteristics of codon usage biases in viruses that are adapted to human hosts.

2.2. Methods

2.2.1. Fundamental computational environment

Visual Studio Code (VScode, Microsoft) was majorly used as coding IDE (version 1.83.1) on personal computer with fundamental Python environment and Jupyter Notebook environment established under Anaconda platform (version 23.3.1). All the python packages were acquired through either Anaconda installation or the Python Package Index (PyPI). High-performance computational sources were provided by UTS Interactive High-Performance Computing facility (iHPC) with multiple available high-performance computing nodes, which could be remotely assessed with VScode SSH (Secure Shell Protocol) extension 'Remote - SSH'.

2.2.2. Acquisition of virus genome sequences with RefSeq data

The accession IDs of all the virus genome reference sequences (RefSeq) and their corresponding host ranges (under label 'Host') are acquired from the 'Viral genome browser' of National Center for Biotechnology Information (NCBI)[335]. Hosts with a limited sample count are ignored in later studies except 'Human', 'Vertebrates', 'Invertebrates', 'Land plants', and 'Bacteria', but they remained in the dataset as negative samples. The incomplete viral genome sequences (labelled as 'Incomplete' in 'RefSeq type') were discarded in the analysis. The multi-partite virus which has multiple NCBI accession IDs for multiple genome segments are summarised as the same virus. Total 10820 samples were retrieved with 488 Human samples, 1758 Vertebrates samples, 1851 Invertebrates samples, 1763 Land plants samples, and 4041 Bacteria samples (Figure 12).

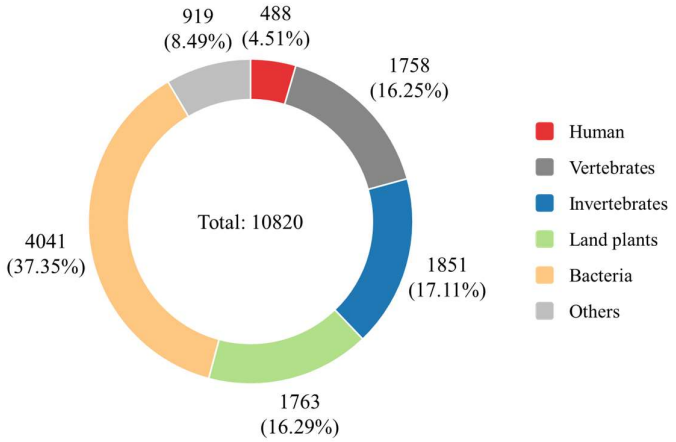


Figure 12. Counts and percentages of downloaded virus genomes with different host ranges. The overall dataset of total 10820 samples includes 488 (4.51%) Human samples, 1758 (16.25%) Vertebrate samples, 1851 (17.11%) Invertebrate samples, 1763 (16.29%) Land plant samples, 4041 (37.35%) Bacteria samples, and 919 (8.49%) Other samples. Some of samples have multiple host range labels.

2.2.3. Acquisition of coding sequences in virus genomes

The CDS of all viral genes encoded in a virus genome were downloadable from NCBI GeneBank database according to the NCBI Accession ID of virus genome provided (i.e. NC_045512 for SARS-CoV-2 RefSeq genome) through Biopython toolkit[336] (Appendix 6). Same method is applicable to other virus genomes (non-RefSeq) or just a single gene if the correct accession ID is provided. By basically extracting all the features with feature name of 'CDS' in the downloaded GeneBank file of virus genome sequences through Biopython, I am allowed to download all the CDS in a virus genome under 'CDS' feature, and the sub-feature 'mRNA_sequence' is the actual coding sequence (Appendix 6). The downloaded CDS will be discarded if the length is not dividable by 3 (considered not-complete genome).

2.2.4. Calculation of Relative Synonymous Codon Usages

The Relative Synonymous Codon Usages (RSCU) of the virus genome, as readouts of codon usage biases, are calculated based on codon counts and amino acid counts of coding sequences according to their definition proposed in previous publication (see equation below) [328, 329].

$$RSCU_{ij} = \frac{Codon\%_{ij}}{AminoAcid\%_i} \times n_i$$

The variable j represents the index of a specific codon, while i denotes the index of a specific amino acid. The RSCU is calculated by dividing the percentage of a given codon ($Codon\%_{ij}$) encoding amino acid i by the overall percentage of amino acid i ($AminoAcid\%_i$), and then multiplying the result by n_j, the number of codons that encode amino acid i. For example, n_{Ala} is equal to 4, as four codons encode alanine (GCU, GCC, GCA, GUG).

All the coding sequences, or CDS, in a virus genomes (either mono-partite or multi-partite) were converted into counts of each codon and counts of each amino acid. The counts of the same codons from all CDS in a virus genome were summed to represent codon counts for the whole genome, which same method was applied to the counts of amino acids. RSCU of each codon were calculated based on each codon count and respective amino acid count. The codons for 1-box amino acids, UGG (Try) and AUG (Met) are discarded for later use due to unchanged values (= 1). The stop codons UAA, UAG and UGA are also discarded because they were not relevant to translation efficiency. Thus, the RSCU dataset, D_{RSCU} (or D_R) consists of total 59 codon features. As the start and stop codons are discarded, our analysis was purely focused on the codon usage biases of the coding sequences related to translation efficiency.

2.2.5. Calculation of Codon% and AminoAcid%

Besides RSCUs, other feature datasets were used to achieve better machine learning prediction. Datasets 'Codon%' ($D_{Codon\%}$) and 'AminoAcid%' ($D_{AminoAcid\%}$) were simply calculated with percentages of different codons or amino acids in total codon count of amino acid count of

virus genome. Stop codons were ignored in both $D_{Codon\%}$ and $D_{AminoAcid\%}$ (thus 61 features in $D_{Codon\%}$ and 20 features in $D_{AminoAcid\%}$, stop codons were not included).

2.2.6. Acquisition of data related to Human tRNA supplies

Two types of tRNA supplies of human are utilised to study virus codon usage biases in this thesis, which are predicted tRNA gene counts and mature tRNA expression level. The predicted tRNA gene counts are data-mined from the GtRNAdb database[138]. Briefly, predicted tRNA gene counts are the predicted counts of available tRNA genes for the specific codons. Those data is predicted with previous published bioinformatics tool tRNAscan-SE[337] and the human reference genome GRCh38/hg38. The data of mature tRNA expression level is from raw data of publication regarding mature tRNA sequencing from HEK293 cells[139]. The mature tRNA expression levels are the sum of RNAseq read counts of all available mature tRNA for specific codon.

2.2.7. Statistical tests comparing different groups of data

The independent T-test was performed for different purposes through the python package Scipy. Basically, T-test is used to compare RSCU data or other data, such as Codon% and AminoAcid% of each codon or each amino acid between from different groups of virus genomes, such as human virus genomes and not-human virus genomes.

The Spearman correlation coefficient (SCC) was performed for different purposes also through the python package Scipy. Basically, the SCC was used to compare in a ranking manner between RSCU data or other data, such as Codon% and AminoAcid% and other translation related metrics, such as tRNA abundances.

2.2.8. Acquisition of codon properties for correlation analysis

To investigate more clues in codon usage distinction between human and non-human viruses, the specific amino acid or codon properties is the investigating subject. Briefly, the interested amino acid properties included in this analysis are 'Essential', 'Essential (General)', 'Polar (General)', 'Codon box', 'Codon box > 2', 'A present', 'U present', 'G present', 'C present', 'Total AU/GC-rich', 'ADAT relate', 'Sum of predicted tRNA gene counts', and 'Sum of mature tRNA expression levels'. The interested codon properties included in this analysis are 'Essential', 'Essential (General)', 'Polar (General)', 'Codon box', 'Codon box > 2', 'A present', 'U present', 'G present', 'C present', 'A/U present', 'G/C present', 'A count', 'U count', 'G count', 'C count', 'A/U count', 'G/C count', 'First nucleotide as A/U', 'First nucleotide as G/C', 'Second nucleotide as A/U', 'Second nucleotide as G/C', 'Third nucleotide as A/U', 'Third nucleotide as G/C', 'Total AU/GC-rich', 'ADAT relate', 'ADAT suppress/benefit', 'Predicted tRNA gene counts', and 'Mature tRNA expression levels'. All the details of both amino acids and codon properties are listed in Appendix 7.

2.2.9. Bootstrapping resampling for estimating data distribution

Bootstrapping resampling method is used to estimate the distribution of the data including AminoAcid%, Codon% and RSCU. Due to the limitation of the data size, sometimes the statistical results computed from comparison between two groups of data are unreliable. Thus, the distribution of the median of bootstrapping resampled data are generated. The resampled data has the same sample size as input data, and the resampled iteration is set as 10000. The confident interval is set as 5% to 95%. If the confident intervals of two median distributions have no overlapping, the two original data groups are considered as significant different.

2.2.10. Uniform Manifold Approximation and Projection

Dimensional reduction analysis was performed on the normalised and compressed data using the Uniform Manifold Approximation and Projection (UMAP) algorithm[338] for both RSCU

data and transposed RSCU data. This method visualises data points in high dimensional space with better preservation of global structure compared to other algorithms like t-SNE.

2.2.11. Bioinformatics program Multi-Codon Analyser to study multi-codon usage biases of virus genomes

'MultiCodonAnalyser' is python-based data mining bioinformatics tool to calculate RSMCU-n of different codon stretch length (CSL) *n* from coding sequences in parallel computation manner. Besides, it can identify codons (or codon stretches) with significantly different codon biases based on statistical tests including Welch T-test, Mann-Whitney U-Test, 2-sample Kolmogorov–Smirnov Test et al.

The codon count pool is first generated by simply counting codon combinations based on *n* in CSL, where codon count pools of different coding sequences with same 'Index' label will be summed together. The codon combinations and amino acids combinations containing stop codons will be ignored in the counting but remained in coding sequence length for calculations of codon combinations percentages (Codon%) and amino acid combinations percentages (AminoAcid%). The RSMCU-n were calculated with those two percentages and number of synonymous codons for certain amino acid.

Two RSMCU-n data matrix with same *n* value generated from different groups of coding sequences could be compared to find codons (or codon stretches) that were significantly different according to statistical tests. In this paper, the significantly different codons (or codon stretches) were screened out based on Mann-Whitney U-Test. Besides RSMCU-n, the non-zeroes percentages (NZP) of certain codons (or codon stretches), which are percentages of coding sequences from a group of coding sequences having those codons (or codon stretches), were also used to find significantly different codons (or codon stretches).

2.3. Results

2.3.1. Establishing RSCU computational pipeline to analyse codon usage biases of virus genomes

Before analysing virus codon usage biases, it is essential to first obtain the coding sequences (CDS) of viral genes from human viruses. These sequences form the basis for calculating various genomic features such as Codon%, AminoAcid%, and RSCU. However, comparing human viruses with non-human viruses on a broad scale presents challenges, primarily due to the uneven availability of genome sequences across different virus species, in another term, data imbalance. For example, the COVID-19 pandemic has resulted in the production of an extensive number of SARS-CoV-2 genome sequences due to the advancements in high-throughput sequencing technologies. Utilising all available virus genome sequences could disproportionately represent human viruses, particularly SARS-CoV-2, potentially skewing the statistical analysis due to sample size imbalances. To avoid this issue and enhance the generalisability of the comparisons between human and non-human viruses, I propose utilising Reference Sequences (RefSeq) from the NCBI database. This database provides ideally one reference genome sequence for each recorded virus species, offering a more balanced and standardised dataset for analysis. The virus genome sequence, and all the CDS encoded (all the virus genes) in the genome are acquired according to the NCBI accession ID of the virus (Appendix 6). For multi-partite virus, or virus with multiple genome segments, same approach is applied that all the CDS from each genome segment are considered as from the same virus sample.

Figure 13 illustrates the computational pipeline utilised to derive codon usage-related features such as Codon%, Amino Acid%, and RSCU from the data of virus genome sequences and their CDS. For each virus CDS within a genome, every codon is counted. This step involves tallying each specific triplet of nucleotides that codes for an amino acid across the entire coding sequence. The counts for each codon from all CDS within the same virus genome are aggregated. This summation provides a comprehensive view of how often each codon is used across the entire genome, rather than within individual genes or segments. The feature data of

Codon%, AminoAcid% and RSCU are calculated from summed codon counts of virus genome to represent the codon usage characteristics of the virus genome. With this computational pipeline, I generated different data matrix for Codon%, AminoAcid% and RSCU data according to the overall list of NCBI accession ID recorded for all the RefSeq virus genomes. These matrices facilitate subsequent analyses and comparisons across different viruses.

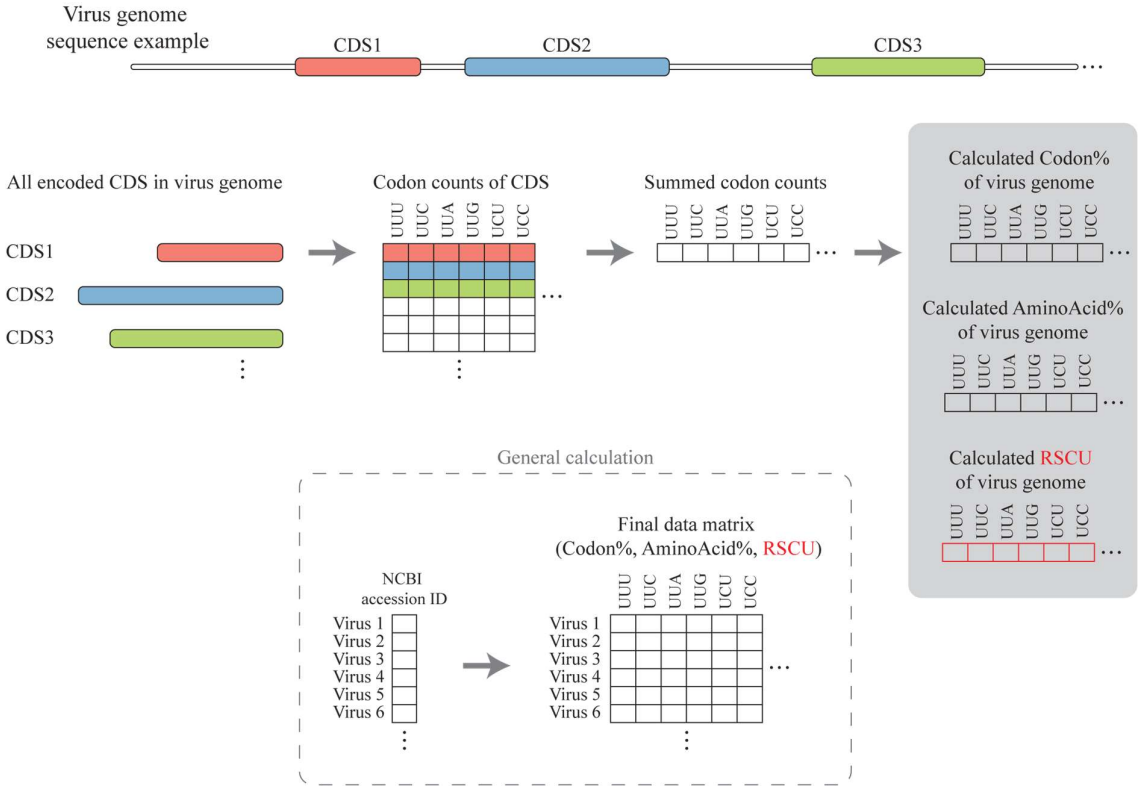


Figure 13. Computational pipeline of RSCU from all virus genomes. All coding sequences of the genes from the genome sequences are extracted according to gene annotations. The coding sequences are converted to codon counts, and aggregated together to get the codon counts of the genomes. Based on the codon counts of the genomes, data of AminoAcid%, Codon% and RSCU of the genomes are computed.

2.3.2. Analysis of SARS-CoV-2 codon usage biases

To have a better understand of the computed RSCU matrix, I first look into the ranked RSCU matrix of SARS-CoV-2 as an example to study the codon usage biases of SARS-CoV-2.

According to Figure 14.A illustrates SARS-CoV-2 RSCU matrix (one sample from the general RSCU matrix), AGA (Arg) has the highest RSCU followed by GGU (Gly), GCU (Ala), and GUU (Val), whereas GGG (Gly) has the lowest RSCU followed by UCG (Ser), CCG (Pro), and CGG (Arg). However, interpreting RSCU values across different codons requires consideration of the underlying calculation method, which might introduce comparative limitations due to the variable scales associated with different amino acids. RSCU values are inherently influenced by the number of synonymous codons available for a given amino acid. For instance, Arginine, which is encoded by six different codons, has an RSCU scale ranging from 0 to 6, where the no-bias point, indicating no preference for any synonymous codon, is set at 1. Conversely, Histidine, encoded by only two codons, has an RSCU scale ranging from 0 to 2, maintaining the same no-bias point at 1. The distribution of RSCU values is complex, as it is derived from dividing two data points assumed to be independently normally distributed. This calculation complicates characterising the distribution of RSCU values across the genome, as the ratio of two normally distributed variables does not typically result in a normal distribution, complicating statistical interpretation[339].

With the calculated RSCU metrics, I was able to correlate the codon usage biases of SARS-CoV-2 with various general properties. Using the RSCU data for SARS-CoV-2 as a reference, I grouped the percentages of either positively-biased or negatively-biased codons based on different codon properties, including nucleotide position, AU/GC richness, and their relationship with ADAT editing (Figure 14.B). In the first nucleotide (1^{st} nt) classification, codons with uracil (U) at the first nucleotide position exhibit a slight positive bias, whereas those with guanine (G) and cytosine (C) display a slight negative bias. This trend is absent in the second nucleotide position, where only codons with G demonstrate a slight negative bias. Notably, in the third position, all U-ending codons are positively biased, contrasting sharply with C-ending codons, which are uniformly negatively biased. Moreover, notable trend suggests a strong preference within the SARS-CoV-2 coding sequences for A- and U-ended codons over C- and G-ended ones. Furthermore, AU-rich codons are more favoured compared to GC-rich codons, indicating a selection for AU-rich codons in the viral genome. For estimating the potential ADAT-editing affects, the ADAT-suppress codons, which their tRNA

are substrates of ADAT editing, tend to be positively biased. On the contrast, the ADAT-benefit codons, which can be decoded by ADAT-edited tRNAs, show a negative bias.

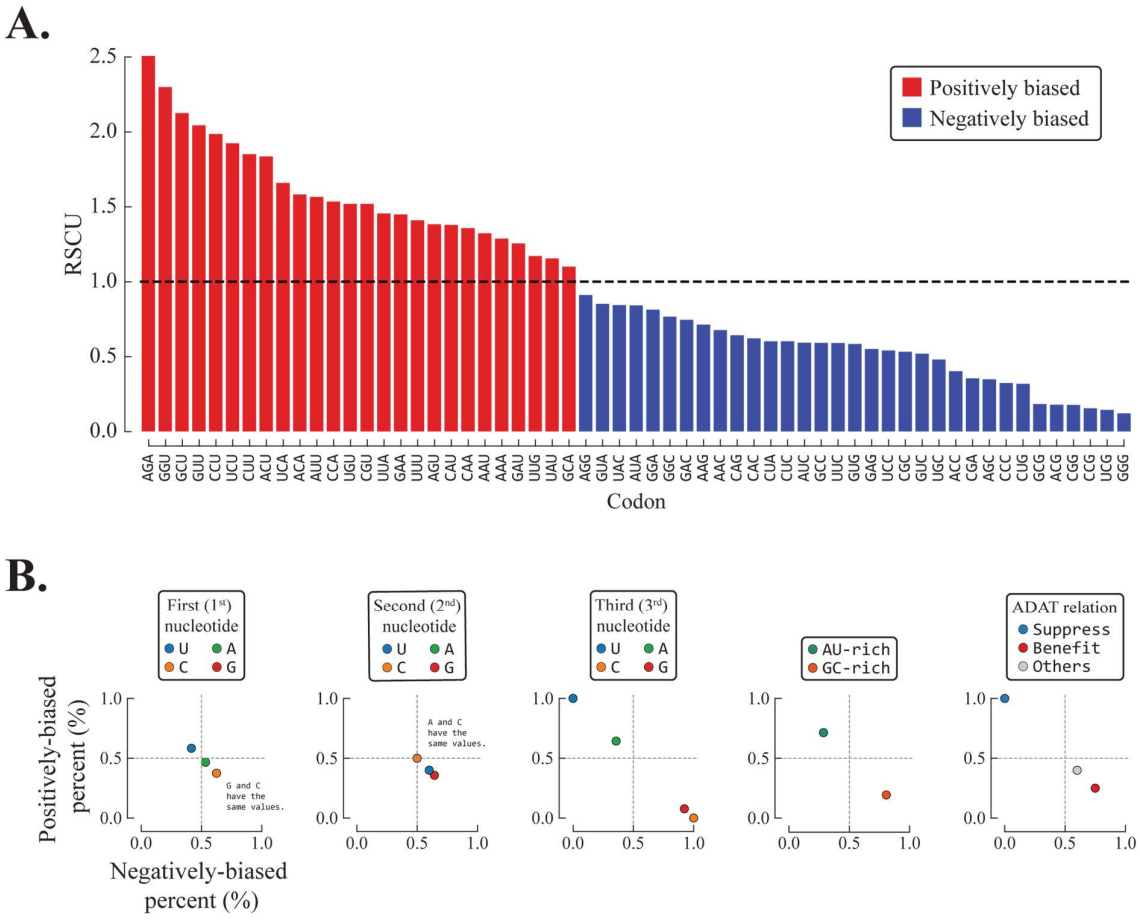


Figure 14. Illustration and analysis of RSCU matrix of SARS-CoV-2 virus genome. (A) The codon labels in red are the positively-biased codons whereas the blue ones are negatively-biased ones; (B) The percentages of positively-biased and negatively-biased codons and their codon properties. The codon properties include First nucleotide (1st nt), Second nucleotide (2nd nt), Third nucleotide (3rd nt), AU/GC-richness, and ADAT relations.

2.3.3. Statistical analyses reveal differences in codon usage biases between human and non-human virus genomes

To distinguish the differences in codon usage between human and non-human viruses, data of AminoAcid%, Codon%, and RSCU are analysed using independent T-tests with a significance threshold set at a p-value of 0.05 (Figure 15). The T-test comparisons of AminoAcid% between

human and non-human viruses reveal that Cysteine (Cys/C) emerges as the most predominant amino acid in human viruses, followed by Proline (Pro/P) and Threonine (Thr/T). Conversely, amino acids such as Aspartic acid (Asp/D) and Glutamic acid (Glu/E) are significantly less prevalent in human viruses. When examining ADAT-related amino acids, no clear preferences are observed across all amino acids involved. However, ADAT-related amino acids such as Proline (Pro/P), Threonine (Thr/T), Leucine (Leu/L), Serine (Ser/S), and Arginine (Arg/R) are significantly more abundant in human viruses, whereas Alanine (Ala/A), Valine (Val/V), and Isoleucine (Ile/I) are considerably less prevalent in human viruses. This analysis indicates distinct amino acid composition profiles that differentiate human viruses from non-human viruses, suggesting potential adaptations to the host's cellular machinery.

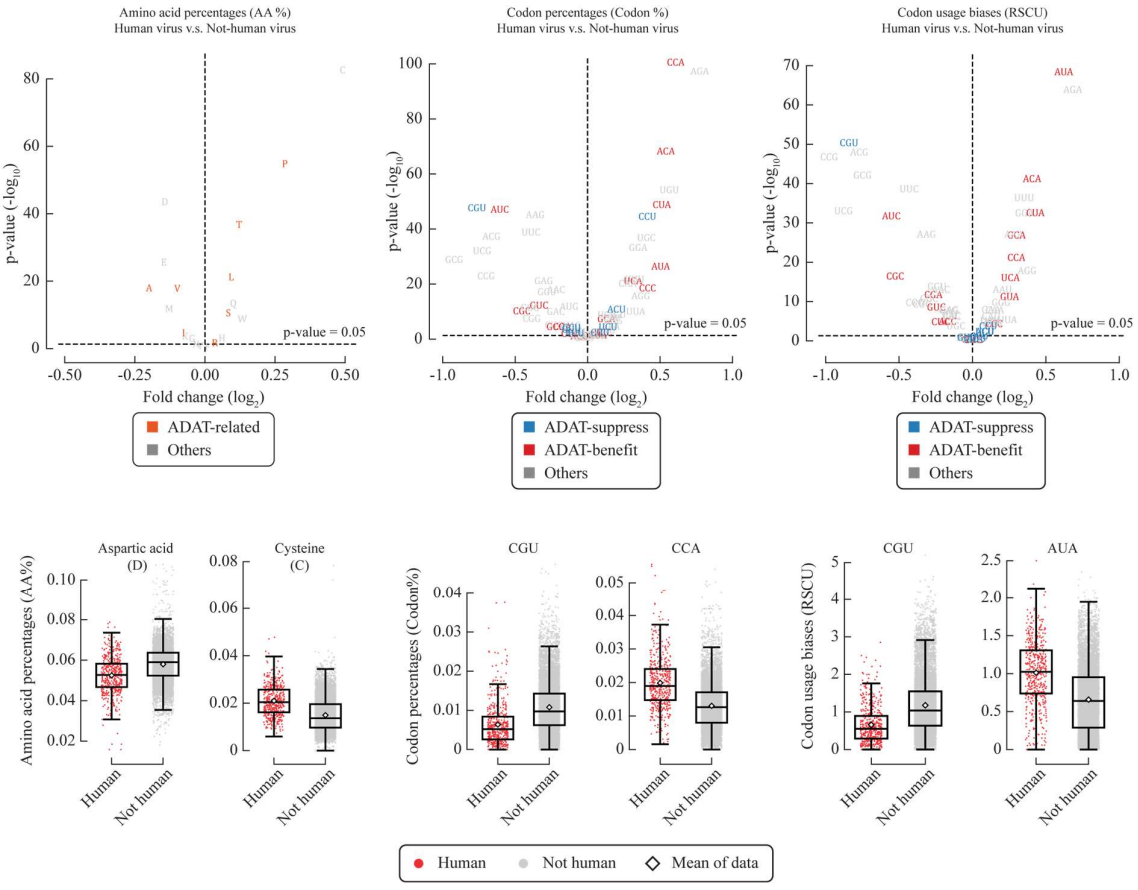


Figure 15. Independent T-test of AminoAcid%, Codon% and RSCU between human and non-human viruses. The non-human virus samples are considered as control group and baseline in the T-test analysis. The volcano plots shows T-test p-value ($-\log_{10}$) and fold change ($\log_2$), and p-values smaller than 0.05 are considered signficantly different. Besides, the top one sample of both positively and negatively changed in each analysis of AminoAcid%, Codon% and RSCU are visualised with boxplot below.

In the T-test analysis of Codon% between human and non-human viruses, codons such as CCA (Pro) and AGA (Arg) are significantly enriched in human viruses, while codons like CGU (Arg) and AUC (Ile) are notably less prevalent. However, no clear enrichment-related patterns emerge with either ADAT-suppress or ADAT-benefit codons in the Codon% data for human viruses.

In the T-test comparison of RSCU between human and non-human viruses, codons such as AUA (Ile) and AGA (Arg) are identified as significantly positively-biased in human viruses when encoding their corresponding amino acids. Conversely, codons such as CGU (Arg) and ACG (Thr) exhibit significant negative bias. Interestingly, ADAT-benefit codons tend to display lower p-values (or higher in $-\log_{10}$ transformed p-value), regardless of whether they are positively or negatively biased. On the other hand, ADAT-suppress codons appear more 'stable' without significant biases, except for CGU (Arg), which is significantly negatively-biased. This observation suggests that ADAT2/3 expression may differentially impact the translation of various amino acids in viral genes. In another term, the presence of tRNAs edited by ADAT2/3 does not necessarily impact their associated codons' translation efficiency of the viral genes from human viruses. However, codons that utilise post-edited tRNAs might experience significant impacts on translation, either enhancing or reducing efficiency.

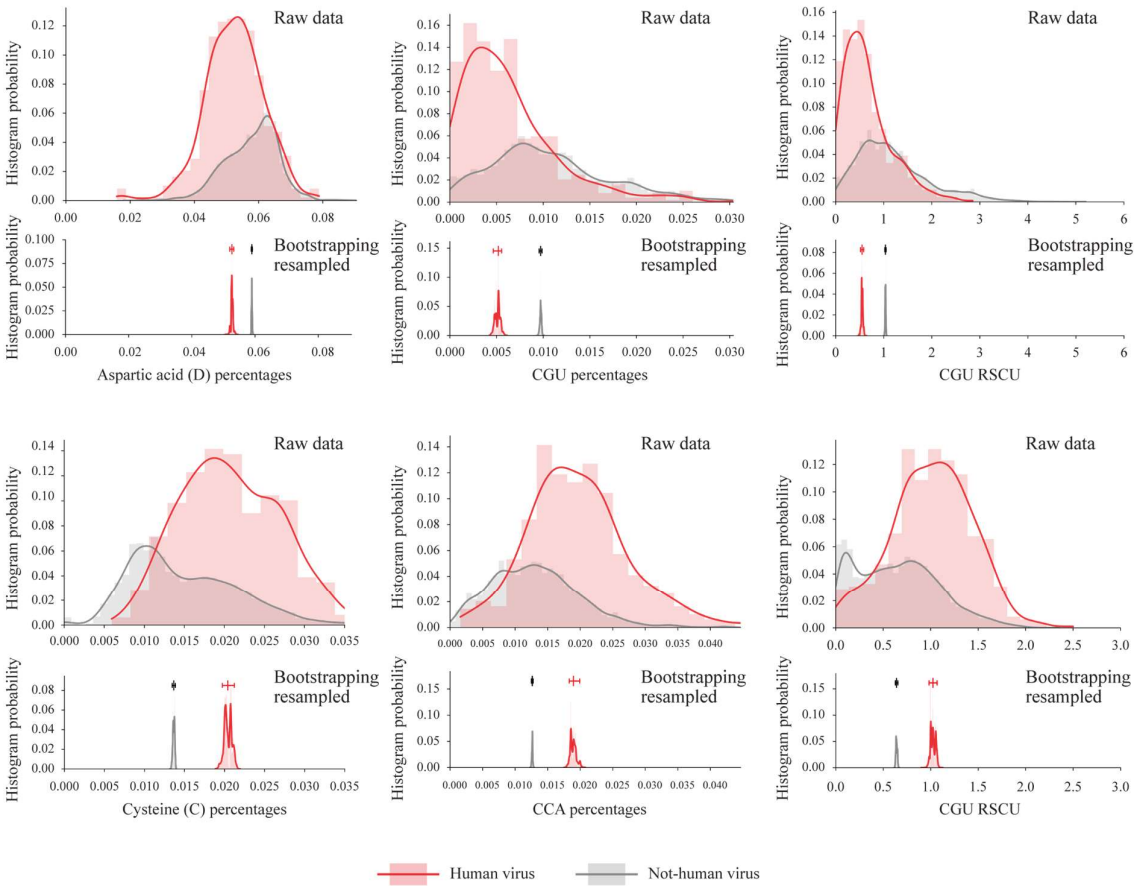


Figure 16. Bootstrapping resampling of AminoAcid%, Codon% and RSCU of human and non-human viruses to verify usage differences. Each amino acid or codon feature, which is identified significantly changed in above independent T-test analysis, are resampled to estimate their data distribution and confident interval (CI). The non-overlapping CI suggests the distribution of the two groups are significantly different.

To verify the findings from the T-test analysis, a bootstrapping method was applied to resample the data of Amino Acid%, Codon%, and RSCU for human and non-human virus groups. This approach facilitated the assessment of the distribution of resampled medians. As illustrated in Figure 16, the resampled data with 10,000 iterations show distinct differences for the top examples of codons and amino acids identified in the initial T-test results. Notably, there is no overlap between the confidence intervals of the resampled medians between human and non-human viruses. This absence of overlap further confirms the significant distinctions in Amino Acid%, Codon%, and RSCU between human and non-human viruses. Other results, similarly, verified using this method and suggested by a p-value < 0.05 in the T-test, are confirmed as

well (data not shown). This robust methodological approach ensures the reliability of the statistical differences observed, reinforcing the significance of the findings.

2.3.4. Correlation analyses reveal relationships between codon properties and T-test results regarding codon usages between human and non-human virus genomes

After identifying distinct characteristics in Amino Acid%, Codon%, and RSCU between human and non-human viruses, the investigation extends to identifying which properties of amino acids or codons contribute to these distinctions. To explore this, a ranking-based Spearman correlation method is applied to examine the relationships between the distinct characteristics found in Amino Acid%, Codon%, and RSCU and the intrinsic properties of amino acids or codons. These distinct characteristics are quantified using statistics of transformed T-test p-values ($-\log_{10}$) and transformed fold changes ($\log_2$). A higher value in the transformed T-test p-value indicates more significant changes in codons or amino acids, either positively or negatively biased, while a higher value in the transformed fold change suggests greater abundances or biases in codon or amino acid usage. For the analysis of codon properties, categorical values are numerically encoded to facilitate statistical computation. For instance, AU-rich codons are encoded as 1 and GC-rich codons as 0. This encoding allows the computation of the Spearman Correlation Coefficient (SCC) between the transformed fold change from RSCU T-test results and AU/GC richness. The outcome of this analysis could reveal whether AU-rich codons exhibit a positive bias in human viruses compared to non-human viruses. The specific properties of codons and amino acids tested in this analysis are detailed in the methods and section (Appendix 7), ensuring a comprehensive and systematic approach to understanding the underlying factors contributing to the observed codon usage biases.

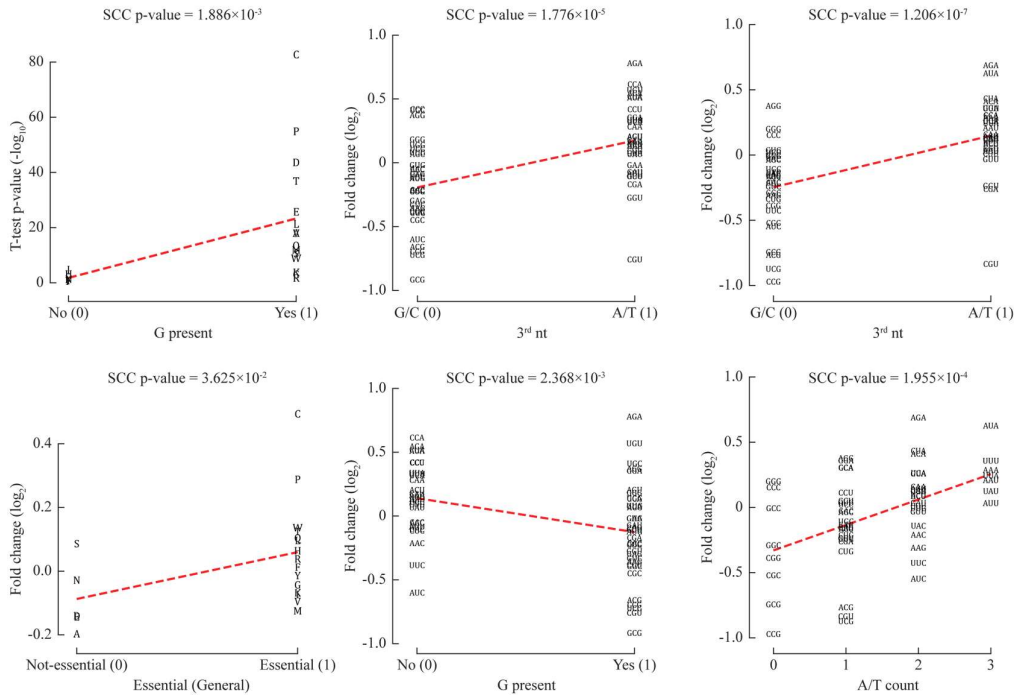


Figure 17. Top-two highest Spearman Correlation Coefficient (SCC) comparison results between T-test statistical metrics and amino acid or codon properties with best fit linear curve. The T-test metrics with AminoAcid%, Codon% and RSCU are all examined and all the amino acid or codon properties are listed in Appendix 7. Other signficant SCC results are shown in Table 2.

Figure 17 presents the top Spearman Correlation Coefficients (SCC) results for Amino Acid%, Codon%, and RSCU, and Table 2 summarises the table of all comparison results that show significant correlations (SCC p-value < 0.05). In the analysis of Amino Acid% SCC, significant positive correlations are spotted between T-test p-values ($-\log_{10}$) and the presence of G, as well as between fold changes ($\log_2$) and essential amino acids. This indicates that the variability in the abundance of G-present amino acids differs markedly between human and non-human viruses, and essential amino acids are more prevalent in human viruses. In the Codon% SCC analysis, significant positive correlations are observed between fold changes ($\log_2$) and various codon properties including AU/GC 3rd nt, A/U count, A count, A/U present, and A present, indicating that AU-rich codons, particularly those with A, are prevalent in human viruses. Conversely, significant negative correlations are detected between fold changes ($\log_2$) and properties such as the G present, G count, AU/GC-rich and C present, suggesting that GC-rich codons, especially those with G, are less frequent in human viruses. Additionally, a notable

negative correlation is found between T-test p-values ($-\log_{10}$) and the third 3rd nt (General), indicating that the variability in the abundance of GC-present codons is more pronounced between human and non-human viruses. This comprehensive correlation analysis provides insight into the distinct codon usage characteristics and their implications on the translational efficiency and viral adaptability in human versus non-human viruses.

Table 2. Summerised significantly correlation results of SCC tests in AminoAcid%, Codon% and RSCU.

Group	*T-test metrics*	*AA/Codon factors*	*Ranking order (encoding)*	*Spearman correlation coefficient (SCC)*	*SCC p-value*	*Correlation*	*BH-adjusted SCC p-value*	*BH test result*
AA%	p-value ($-\log_{10}$)	G present	Yes (1), No (0)	0.651	1.886×10^{-3}	Positive	7.923×10^{-2}	Not-passed
AA%	Fold change ($\log_2$)	Essential (General)	Essential (1), Not-essential (0)	0.471	3.625×10^{-2}	Positive	5.552×10^{-1}	Not-passed
Codon%	Fold change ($\log_2$)	3rd nt (General)	A/U (1), G/C (0)	0.520	1.776×10^{-5}	Positive	5.329×10^{-4}	Passed
Codon%	Fold change ($\log_2$)	G present	Yes (1), No (0)	-0.382	2.368×10^{-3}	Negative	3.495×10^{-2}	Passed
Codon%	Fold change ($\log_2$)	A/U count	-	0.375	2.912×10^{-3}	Positive	3.495×10^{-2}	Passed
Codon%	Fold change ($\log_2$)	G count	-	-0.338	7.738×10^{-3}	Negative	7.738×10^{-2}	Not-passed
Codon%	Fold change ($\log_2$)	AU/GC-rich	GC-rich (1), AU-rich (0)	-0.319	1.227×10^{-2}	Negative	1.052×10^{-1}	Not-passed
Codon%	p-value ($-\log_{10}$)	3rd nt (General)	A/U (1), G/C (0)	-0.300	1.889×10^{-2}	Negative	1.259×10^{-1}	Not-passed
Codon%	Fold change ($\log_2$)	A count	-	0.285	2.625×10^{-2}	Positive	1.445×10^{-1}	Not-passed
Codon%	Fold change ($\log_2$)	A/U present	Yes (1), No (0)	0.284	2.649×10^{-2}	Positive	1.445×10^{-1}	Not-passed
Codon%	Fold change ($\log_2$)	C present	Yes (1), No (0)	-0.265	3.908×10^{-2}	Negative	1.954×10^{-1}	Not-passed
Codon%	Fold change ($\log_2$)	A present	Yes (1), No (0)	0.255	4.738×10^{-2}	Positive	2.187×10^{-1}	Not-passed
RSCU	Fold change ($\log_2$)	3rd nt (General)	A/U (1), G/C (0)	0.625	1.206×10^{-7}	Positive	3.617×10^{-6}	Passed
RSCU	Fold change ($\log_2$)	A/U count	-	0.467	1.955×10^{-4}	Positive	2.933×10^{-3}	Passed
RSCU	Fold change ($\log_2$)	A count	-	0.432	6.381×10^{-4}	Positive	6.727×10^{-3}	Passed
RSCU	Fold change ($\log_2$)	C present	Yes (1), No (0)	-0.430	6.727×10^{-4}	Negative	6.727×10^{-3}	Passed
RSCU	Fold change ($\log_2$)	AU/GC-rich	GC-rich (1), AU-rich (0)	-0.411	1.238×10^{-3}	Negative	1.061×10^{-2}	Passed
RSCU	Fold change ($\log_2$)	A present	Yes (1), No (0)	0.395	1.965×10^{-3}	Positive	1.473×10^{-2}	Passed
RSCU	Fold change ($\log_2$)	C count	-	-0.375	3.385×10^{-3}	Negative	2.257×10^{-2}	Passed
RSCU	Fold change ($\log_2$)	G/C present	Yes (1), No (0)	-0.351	6.436×10^{-3}	Negative	3.862×10^{-2}	Passed
RSCU	Fold change ($\log_2$)	G present	Yes (1), No (0)	-0.321	1.325×10^{-2}	Negative	7.229×10^{-2}	Not-passed
RSCU	p-value ($-\log_{10}$)	U count	-	-0.308	1.767×10^{-2}	Negative	8.833×10^{-2}	Not-passed
RSCU	Fold change ($\log_2$)	G count	-	-0.272	3.684×10^{-2}	Negative	1.501×10^{-1}	Not-passed
RSCU	p-value ($-\log_{10}$)	U present	Yes (1), No (0)	-0.272	3.738×10^{-2}	Negative	1.501×10^{-1}	Not-passed
RSCU	p-value ($-\log_{10}$)	A count	-	0.271	3.754×10^{-2}	Positive	1.501×10^{-1}	Not-passed
RSCU	p-value ($-\log_{10}$)	ADAT benefit	Benefit (2), Others (1), Suppress (0)	0.266	4.154×10^{-2}	Positive	1.514×10^{-1}	Not-passed
RSCU	Fold change ($\log_2$)	A/U present	Yes (1), No (0)	0.265	4.289×10^{-2}	Positive	1.514×10^{-1}	Not-passed

In the RSCU SCC analysis, significant positive correlations are observed between fold change (log_2) and various codon properties such as the general 3rd nt (General), A/U count, A present, A/U present, that AU-rich codons exhibit a positive bias in usage. Additionally, positive correlations are found between T-test p-values ($-log_{10}$) and properties like A count and ADAT benefit, suggesting that the usage biases of codons with increasing A count and ADAT benefit exhibit greater variability between human and non-human viruses. Conversely, significant negative correlations are observed between fold change (log_2) and properties such as C present, AU/GC-rich, C count, G/C present, G present and G count, indicating that GC-rich codons are negatively biased in usage. Furthermore, negative correlations are also noted between T-test p-values ($-log_{10}$) and properties including U count and U present, highlighting that the usage biases of U-rich codons are more stable and exhibit less variability between human and non-human viruses. These findings underscore the differential codon usage preferences between human and non-human viruses, particularly in terms of nucleotide composition at critical codon positions (i.e. 3rd nt), which could have implications for the translational dynamics and adaptation strategies of viruses in various hosts.

In summary, the analyses of SCC correlations highlight the significant role of AU/GC content in differentiating the codon and amino acid usage patterns between viruses that infect humans and those that do not. The observations point to a marked preference for AU-rich codons, particularly those with A or U at the 3rd nucleotide position, in human-infecting viruses. This preference suggests that such viruses may have adapted to optimise translational efficiency within human hosts. Furthermore, the significant bias towards AU-rich codons in viruses that infect humans implies a potential susceptibility to ADAR editing, given the abundance of adenosine, which serves as a substrate for ADAR. The A/U at 3rd nucleotide in human viruses. This sensitivity extends to ADAT editing as well, given the relevance of the bias in A/U at 3rd nucleotide position in determining ADAT's editing targets.

2.3.5. Statistical analyses reveal distinct codon usage in human viruses related to ADAT editing

Following the initial findings, a more detailed examination of ADAT-related codon and amino acid usage between human and non-human viruses was conducted. The proportion of amino

acids (AminoAcid%) corresponding to the eight amino acids influenced by ADAT2/3 editing, including Threonine (Thr/T), Alanine (Ala/A), Proline (Pro/P), Serine (Ser/S), Leucine (Leu/L), Isoleucine (Ile/I), and Valine (Val/V), was aggregated. Additionally, the AminoAcid% for amino acids not affected by ADAT was similarly calculated. An independent T-test analysis between human and non-human viruses revealed that the aggregated AminoAcid% for ADAT-related amino acids is significantly elevated in human viruses compared to non-human viruses (Figure 18.A). This statistically significant difference underscores the potential influence of ADAT editing on the adaptation of human viruses. A parallel analysis was conducted for Codon% involving the aggregation of ADAT-related codons. This was contrasted with the aggregation of codons not related to ADAT editing and specific codons within ADAT-related amino acids that are not influenced by ADAT, such as GUG in Valine (Val/V). Similar to the comparison in AminoAcid%, the summed Codon% of ADAT-related codons is significantly higher in human viruses, whereas both the Codon% of other codons and other codons in non-ADAT-related codons are significantly lower conversely (Figure 18.B). These findings suggest that ADAT editing activities could play a critical role in shaping the evolutionary and infection dynamics of viruses in humans. The higher prevalence of ADAT-related amino acids and codons in human viruses may reflect a selection pressure or an adaptation strategy that enhances viral compatibility with human host cellular machinery, potentially influencing the overall efficiency of viral protein synthesis and function.

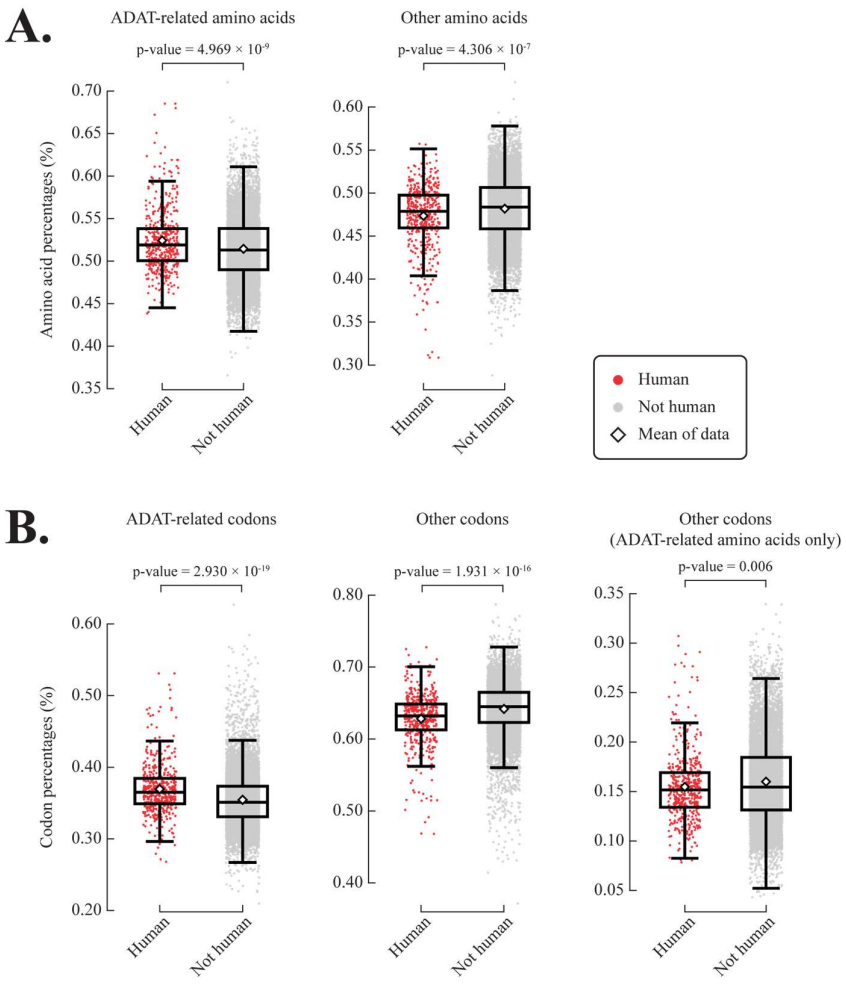


Figure 18. Aggregated AminoAcid% and Codon% of ADAT-related codons and amino acids between human and non-human viruses. (A) Boxplot of aggregated AminoAcid% of ADAT-related amino acids and comparison between human and non-human viruses. (B) Boxplot of aggregated Codon% of ADAT-related codons and comparison between human and non-human viruses.

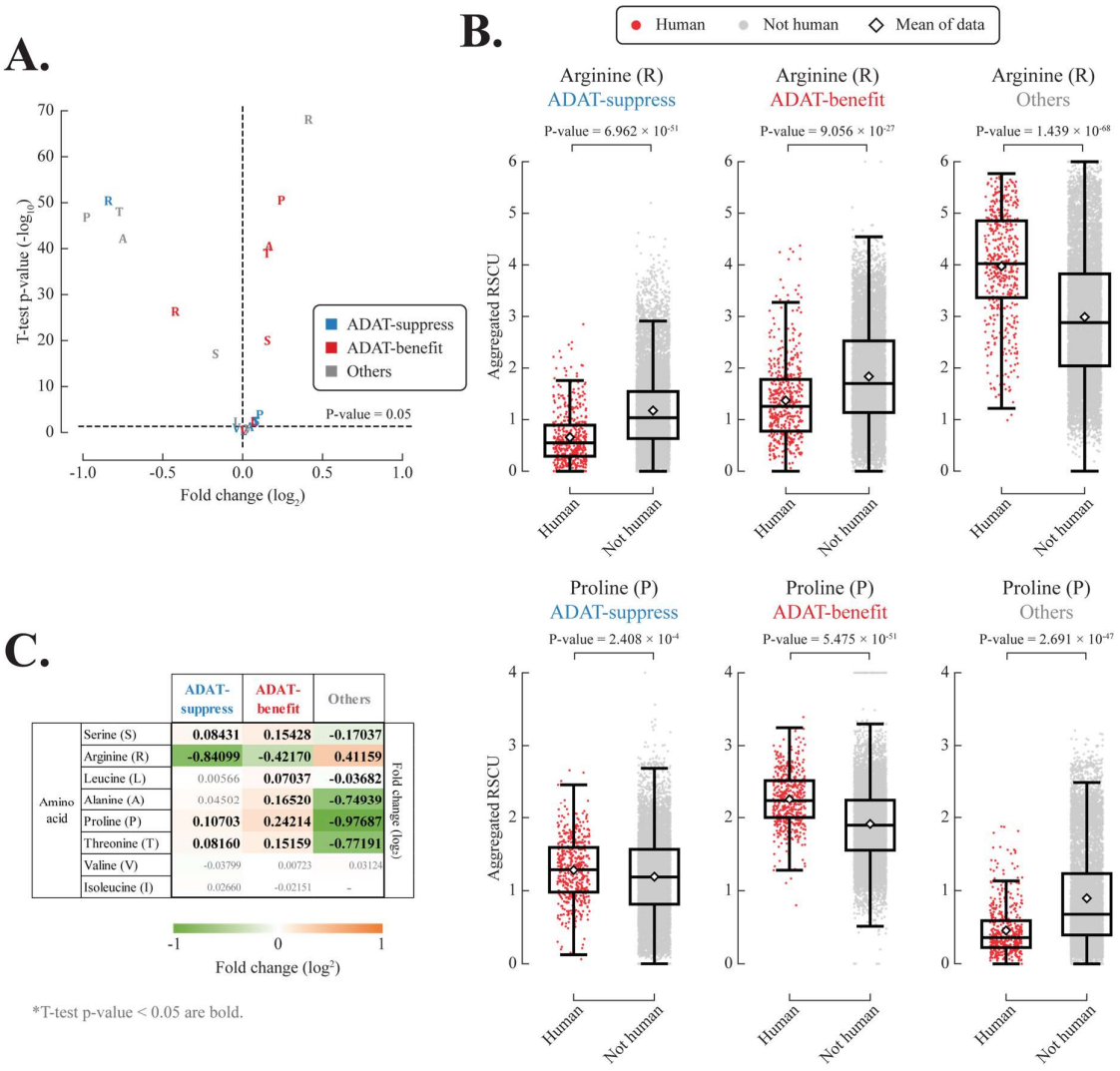


Figure 19. ADAT-related codon usage biases between human and non-human viruses. (A) Volcano plot demonstrating independent T-test analysis of aggregated RSCU in different ADAT relations of amino acids between human and non-human viruses. ADAT relations include ADAT-suppress, ADAT-benefit and others. The aggregated RSCU is the sum of different RSCU under same class of codons (i.e. ADAT-benefit) from the same amino acid. For example, the red coloured 'R' shows the T-test result comparing the aggregated RSCU of ADAT-benefit codons under Arginine between human and not-human viruses; (B) Example of aggregated RSCU with different ADAT relations within Arginine and Proline, which have distinct usage biases in ADAT related codons; (C) Table summarises fold changes (log_2) of aggregated RSCU within different ADAT-related amino acids.

To explore usage biases of ADAT-related codons, the concept of ADAT-related codon usage biases is introduced, defined by the summed RSCU of codons categorised by their relationship to ADAT2/3 editing. The metrics of ADAT-related codon usage biases could help discovery evidence of selection biases of codons when encoding the same amino acids. For instance, in the codons of Valine (Val/V), GUU is categorised as an ADAT-suppress codon, while GUC and GUA are considered ADAT-benefit codons, and GUG is categorised as other codon. Accordingly, the RSCU of GUU represents ADAT-suppress codon usage bias, the combined RSCU of GUC and GUA represents ADAT-benefit codon usage bias, and the RSCU of GUG is considered as other codon usage bias. To clarify the presentation, these categories are denoted as Sup-RSCU, Ben-RSCU, and Oth-RSCU, respectively, in subsequent analyses. Figure 19 illustrates the results from independent T-tests comparing Sup-RSCU, Ben-RSCU, and Oth-RSCU between human and non-human viruses. Notably, the Ben-RSCU values for Proline (Pro/P), Alanine (Ala/A), Threonine (Thr/T), Serine (Ser/S), and Leucine (Leu/L) are significantly higher in human viruses. Conversely, the Oth-RSCU values for these amino acids, along with the Sup-RSCU and Ben-RSCU for Arginine (Arg/R), are significantly lower in human viruses. This pattern suggests differential roles and impacts of ADAT editing across various amino acids. The summary table in Figure 19 indicates that Valine and Isoleucine do not exhibit substantial differences in ADAT-related codon differentiation between human and non-human viruses, whereas other ADAT-related amino acids generally show a positive bias in ADAT-related codons, both ADAT-suppress and -benefit ones. Remarkably, Arginine displays a unique pattern, being positively biased in non-ADAT-related codons while ADAT-related codons show a negative bias in human viruses, hinting at a specialised role for Arginine in human virus infections.

2.3.6. Dimensional reduction analysis demonstrates relationships between virus host ranges and virus codon usage biases

To investigate the differential codon usage biases among viruses with varying host ranges, an analysis of RSCU compositions across diverse viruses was conducted. This involved the visualisation of these compositions using the UMAP for dimensional reduction, as illustrated in Figure 20.A. This approach enabled a comparative analysis of viruses infecting different

hosts (e.g., vertebrates vs. non-vertebrates), revealing distinct patterns in codon usage biases. The analysis highlighted a significant distinction in the codon usage patterns of bacteriophages compared to viruses infecting other hosts. When bacteriophages are excluded from the dataset, the remaining viruses, although exhibiting somewhat similar RSCU patterns, displayed subtle differences in their distributions. This observation suggests a mild and difficult-to-identified boundary in codon usage biases that could potentially differentiate viruses based on their host range labels. Such boundaries might reflect adaptations to the codon preferences of their respective hosts.

I conducted a transposition of the RSCU data matrix to explore overarching patterns of codon behaviour across virus genomes, visualising these patterns through the UMAP dimensional reduction algorithm, as presented in Figure 20.B. To be accurate, the codon labels become the subject whereas all the virus genome samples become the data features. This analysis revealed distinct clustering based on the 3rd nucleotide of codons, demonstrating two predominant groups characterised by A/U-ended and G/C-ended codons, respectively. Notably, within these groups, further sub-clustering was evident: A-ended and U-ended codons formed distinct subgroups, as did G-ended and C-ended codons. Remarkably, this grouping patterns identified exceptional behaviours in two specific codons: UUG (Leu) and AGG (Arg), which did not cluster as might be expected with their G-ended counterparts. Instead, UUG aligned more closely with U-ended codons, and AGG clustered with A-ended codons. This unusual grouping suggests a potentially unique role for these codons in viral genomes, emphasising the importance of the codon wobble position in viruses.

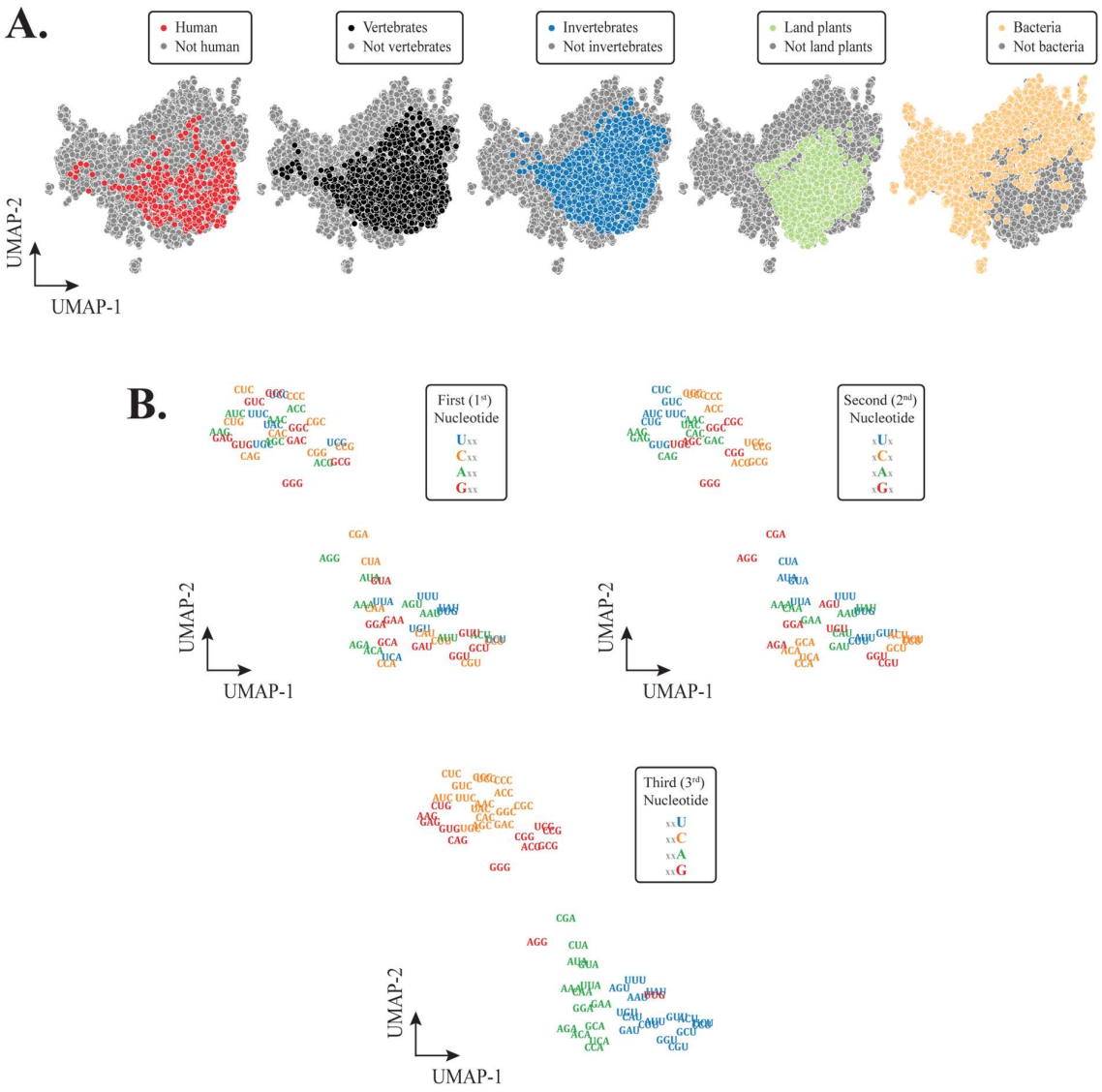


Figure 20. Dimensional reduction analysis regarding RSCU characteristics of virus with different host ranges. (A) UMAP Dimensional reduction for the RSCU of virus genomes with different host ranges. The RSCU data was first normalised with Z-score normalisation then lossless compressed with Principal component analysis (PCA). (B) UMAP Dimensional reduction for the virus genome RSCUs on different codons' wobble nucleotides. The RSCU data was first normalised with Z-score normalisation then lossless compressed with PCA.

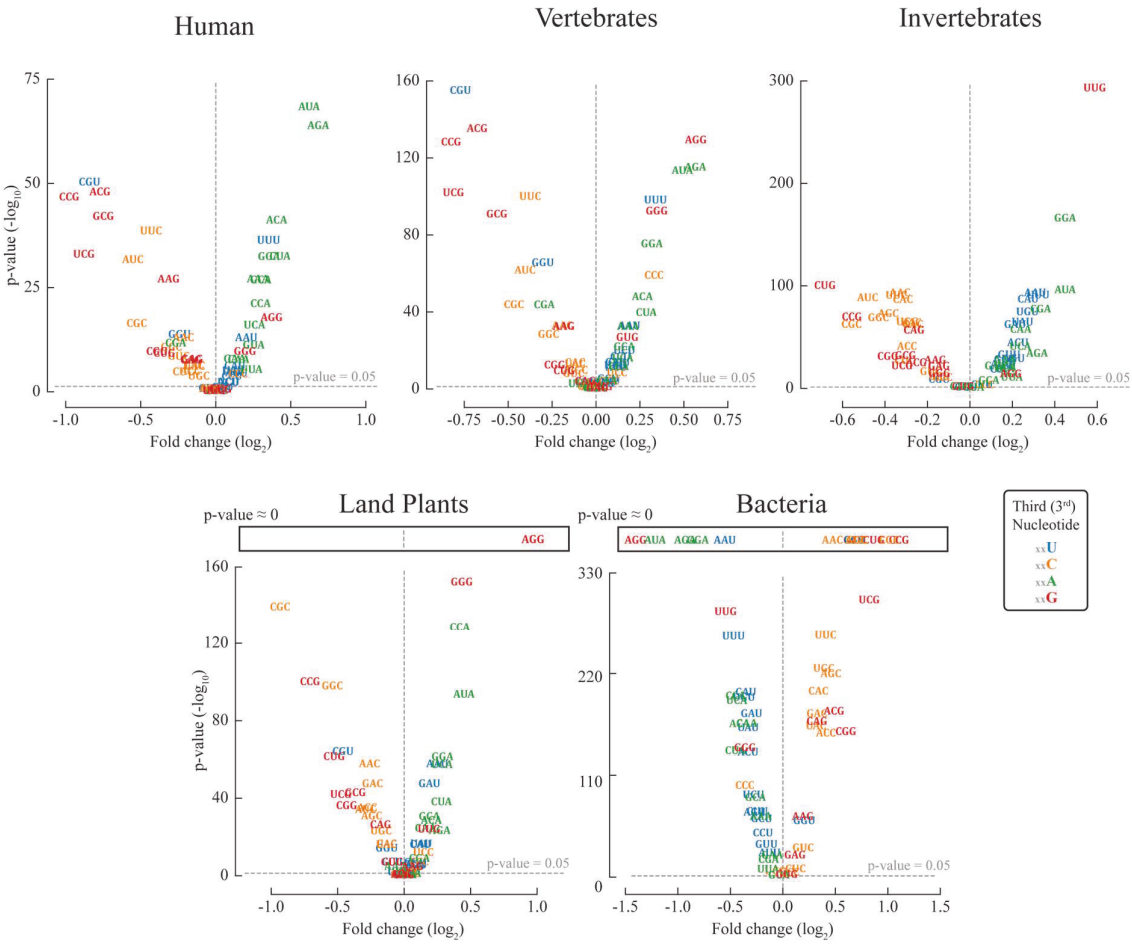


Figure 21. Volcano plots demonstrating independent T-test on the virus genome RSCU regarding comparisons between viruses of different host ranges. Virus host labels include Human, Vertebrates, Invertebrates, Land plants, and Bacteria. The codons are coloured according to third nucleotides (3rd nt).

To identify the specific codon biases within various hosts, I employed independent T-tests on the RSCU data, aiming to identify codons that exhibit significant variability in relation to the specific codon fitness of each host, as illustrated in Figure 21. In viruses that infect humans, it was observed that the RSCUs of A/U-ended codons are significantly higher than those of G/C-ended codons, supporting earlier findings that A/U-ended codons are favoured in human viruses. However, notable exceptions include the G/C-ended codons AGG (Arg), GGG (Gly), and CCC (Pro), which shows a marked abundance relative to other G/C-ended codons. Conversely, A/U-ended codons such as CGU (Arg), GGU (Gly), and CGA (Arg) show a

diminished preference. This pattern suggests a specialised role for codons encoding Arginine and Glycine in codon usage selection within human viruses, possibly due to their unique biological functions. Similar trends were observed in viruses infecting other hosts, including vertebrates, invertebrates, and land plants, where A/U-ended codons generally prevail over G/C-ended codons, although with occasional exceptions (Figure 21). In alignment with findings from the previous UMAP analysis, bacteriophages exhibit a distinct RSCU composition, uniquely favouring G/C-ended codons over A/U-ended ones, further underscoring the diverse evolutionary paths and host adaptation strategies among different viral families.

2.3.7. Using MultiCodonAnalyser to study multi-codon usage biases of various coding sequence groups

The promising outcomes of this preliminary study have motivated further exploration into the relations between virus codon usage biases and their host ranges. RSCU is calculated by quantifying the frequency of singular codons and their corresponding amino acids within a gene's coding sequence. Extending this concept, by considering stretches of codons or amino acids of length *n*, (termed n-length codon or amino acid stretches), it becomes feasible to analyse the usage biases in multi-codon stretches. For instance, one could investigate which codon combinations, such as AGA|CUA or CGA|CUG, are preferred in sequences encoding an Arginine|Leucine (R|L) stretch in a specific coding sequence. To facilitate this analysis, I propose a metric termed Relative Synonymous Multi-Codon Usage (RSMCU-n), where 'n' denotes the number of amino acids in the codon stretch (corresponding to the amino acid stretch length). Here, *n* also represents the Codon Stretch Length (CSL). For example, RSMCU-1 reflects the usage biases of a single codon (same to RSCU), while RSMCU-2 examines the biases associated with a pair of codons (2 codons). Employing the RSMCU-n metric enables the comparison of different codon stretch usage biases across diverse viral genome groups.

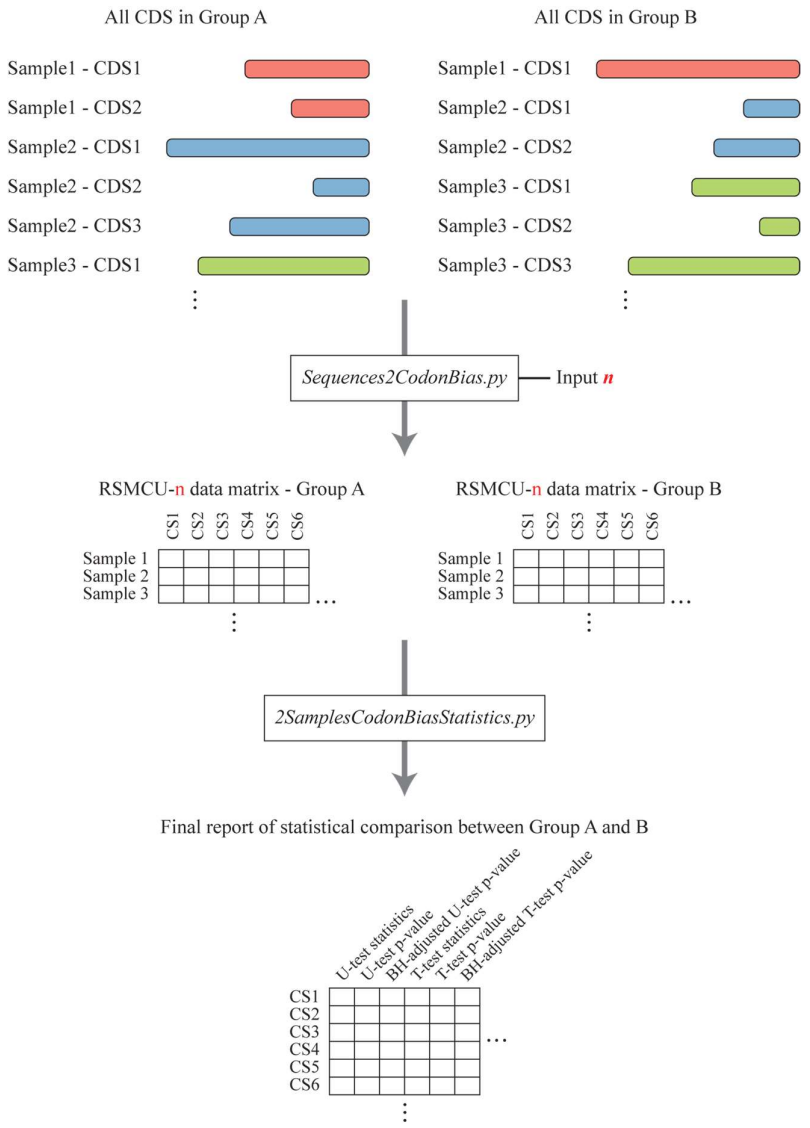


Figure 22. Computational pipeline of MultiCodonAnalyser program. Two major steps are included in this pipeline. First step is converting coding sequences into RSMCU-n matrix with given n value, which is accomplished by script *Sequences2CodonBias.py*. Second step is statistically comparing two RSMCU-n matrix to identify significantly varied codon stretches, which is accomplished by *2SamplesCodonBiasStatistics.py*.

To enhance the efficiency of computing RSMCU-n, I developed a Python-based software tool called 'MultiCodonAnalyser' (MCA), which utilises parallel processing for enhanced computational efficiency. MCA is designed to identify differences in codon usage patterns between two groups of coding sequences by incorporating several statistical tests, including the Mann-Whitney U-Test, Welch's T-test, and the Benjamini-Hochberg procedure for p-value adjustment, among others. The MCA pipeline comprises three major components. Script 'NCBI_CDS_Download.py' facilitates the downloading of coding sequences from the NCBI database using Accession IDs. Script 'Sequences2CodonBias.py' converts strings of coding sequences into an RSMCU-n matrix, where *n* is the user-defined length of the codon stretch. Script '2SamplesCodonBiasStatistics.py' compares two RSMCU-n data matrices (where both matrices are equal in *n* using the statistical tests (Figure 22). By employing MCA, it is possible to identify significant variations in codon stretches based on their multi-codon biases (RSMCU-n) and through various statistical analysis. Additional functionalities are built into MCA to accommodate different CDS data sources. However, these features are still under development and are not discussed further in this document (data not shown).

As the Codon Stretch Length (CSL) increases beyond three (CSL > 3), the RSMCU-n metric tends to exhibit a higher proportion of zero values. These zero values indicate the absence of certain codon stretches in the coding sequences under study, which can complicate statistical analysis and interpretation. To address this issue and enable more meaningful comparisons, the RSMCU-n calculations are refined to include only non-zero values. This approach ensures that the statistical analyses focus on existing codon stretches, enhancing the robustness and relevance of the findings. Further, to complement this analysis, an additional metric, the Non-Zero Percentage (NZP) of codon stretches, is employed. NZP serves as a measure to assess the appearances of various codon stretches within the dataset. This dual approach, analysing both the RSMCU-n values of non-zero codon stretches and NZP-n of zero values, provides a comprehensive understanding of the distribution and significance of codon stretch usage patterns, particularly in cases where longer codon stretches are rare or absent.

2.3.8. Statistical analysis with RSMCU-n reveal unique multi-codon usage biases of human viruses against other vertebrate viruses

The UMAP analysis previously presented in Figure 20.A highlighted that RSCU values do not significantly differ across various virus host ranges, with the exception of bacteriophages. Given the close genetic and evolutionary relationships within vertebrates, including humans, it is not unexpected that distinguishing codon usage biases solely based on single codon biases (such as RSCU or RSMCU-1) between human viruses and other vertebrate viruses proves challenging. To address this limitation and potentially uncover more distinctive features at the boundary between human and other vertebrate viruses, I extended the analysis to multi-codon biases using the MultiCodonAnalyser (MCA) pipeline. This approach focuses on examining RSMCU-n across virus genomes, employing the U-test on non-zero values of RSMCU-n to identify significant codon or codon stretch usage biases. This analysis was specifically conducted for codon stretch lengths (CSL) ranging from 1 to 4, as illustrated in Figure 23. It is noteworthy that the analysis confirmed an increase in the percentage of zero values in the RSMCU-n matrix as the codon stretch length increased, a trend that was documented in Appendix 8.

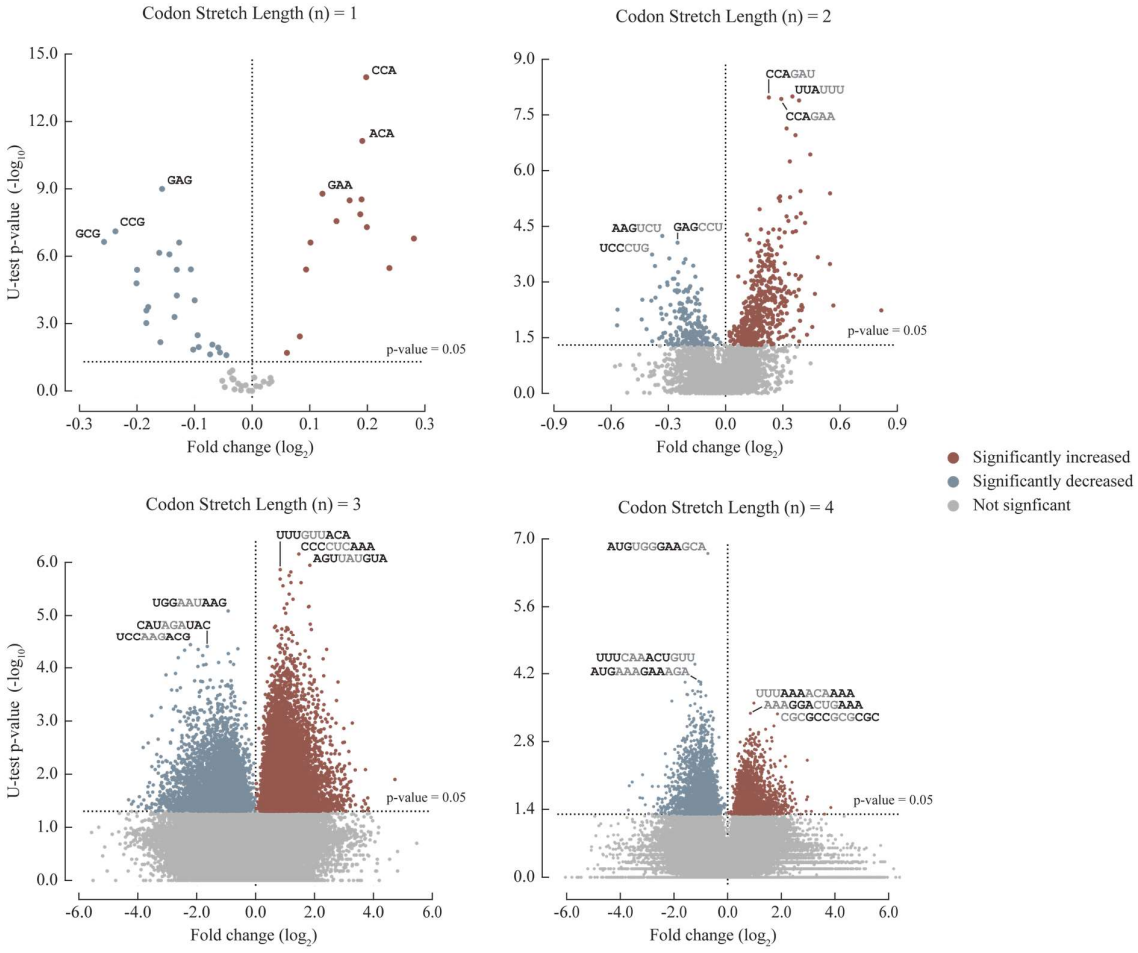


Figure 23. Volcano plot of Mann-Whitney U-test results comparing RSMCU-n ($n \leq 4$) from virus genomes between human and other vertebrate viruses. The significantly increased or decreased codon stretches (p-value < 0.05) are highlighted and the top three codon stretches either positively-biased or negatively-biased are listed.

The extensive analysis conducted using the RSMCU-n metric reveals significant differences in codon stretch usage biases between human and other vertebrate viruses across varying codon stretch lengths (CSL). Here's a summary of the findings. In the RSMCU-1 analysis, out of 61 codon stretches analysed, 14 codons (~22.95%) including CCA (Pro), ACA (Thr), GAA (Glu), and others were found to be relatively enriched in human virus genomes. Conversely, 25 codons (~40.98%) such as GAG (Glu), CCG (Pro), GCG (Ala), and others were relatively enriched in vertebrate virus genomes. In the RSMCU-2 analysis, a total of 3721 codon stretches were analysed. In human virus genomes, 497 codon stretches (~13.35%) including UUA|UUU

(Leu|Phe), CCA|GAU (Pro|Asp), CCA|GAA (Pro|Glu), and others were relatively enriched. In vertebrate virus genomes, 188 codon stretches (~5.05%) including AAG|UCU (Lys|Ser), GAG|CCU (Glu|Pro), UCC|CUG (Ser|Leu), and others showed enrichment. In the RSMCU-3 analysis, among the 226,981 codon stretches analysed, 16,536 codon stretches (~7.29%) such as CCC|CUC|AAA (Pro|Leu|Lys), AGU|UAU|GUA (Ser|Tyr|Val), UUU|GUU|ACA (Phe|Val|Thr), and others were relatively enriched in human virus genomes. On the other hand, 4661 codon stretches (~2.05%) including UGG|AAU|AAG (Trp|Asn|Lys), UCC|AAG|ACG (Ser|Lys|Thr), CAU|AGA|UAC (His|Arg|Tyr), and others were more prevalent in vertebrate virus genomes. In the RSMCU-4 analysis, this level of analysis covered a total of 13,845,841 codon stretches. Only 2583 codon stretches (~0.02%) including UUU|AAA|ACA|AAA (Phe|Lys|Thr|Lys), AAA|GGA|CUG|AAA (Lys|Gly|Leu|Lys), CGC|GCC|GCG|CGC (Arg|Ala|Ala|Arg), and others were relatively enriched in human virus genomes. In contrast, 4569 codon stretches (~0.03%) such as AUG|UGG|GAA|GCA (Met|Trp|Glu|Ala), UUU|CAA|ACU|GUU (Phe|Gln|Thr|Val), AUG|AAA|GAA|AGA (Met|Lys|Glu|Arg), and others were enriched in vertebrate virus genomes.

For a better validation in the findings from the RSMCU-n analyses, I applied the Benjamini-Hochberg (BH) adjustment method to control the false discovery rate among the p-values obtained from the U-tests. Upon applying the BH adjustment, the results demonstrated that all p-values for RSMCU-n analyses with CSL greater than 2 were adjusted to values greater than 0.05. This adjustment indicates that the variations in codon usage between human and vertebrate viruses for these longer codon stretches may be possibly unreliable (Appendix 9).

2.3.9. Statistical analysis with NZP-n reveal unique multi-codon usage biases of human viruses against other vertebrate viruses

To further investigate the usage of codon stretches based on increasing zero percentages in the RSMCU-n matrix, we analysed the non-zero percentages (NZP) of this matrix to determine whether certain codon stretches are significantly over-represented or under-represented in the gene coding sequences of human virus genomes. The variations in codon NZP might, however, be attributable to changes in the NZP of amino acids. To address this, the expected codon NZP were computed by dividing the amino acids NZPs by the number of possible encoding codon

stretches, thereby excluding codon stretches influenced by amino acid variations. In my methodology, codons or codon stretches within the top 5% of absolute codon NZP values were categorised as significantly altered. Codons or codon stretches were excluded and deemed affected by amino acid imbalances if their codon NZP values and the expected codon NZP values shared the same directional sign (+/-) and also ranked within the top 5% in expected codon NZP values.

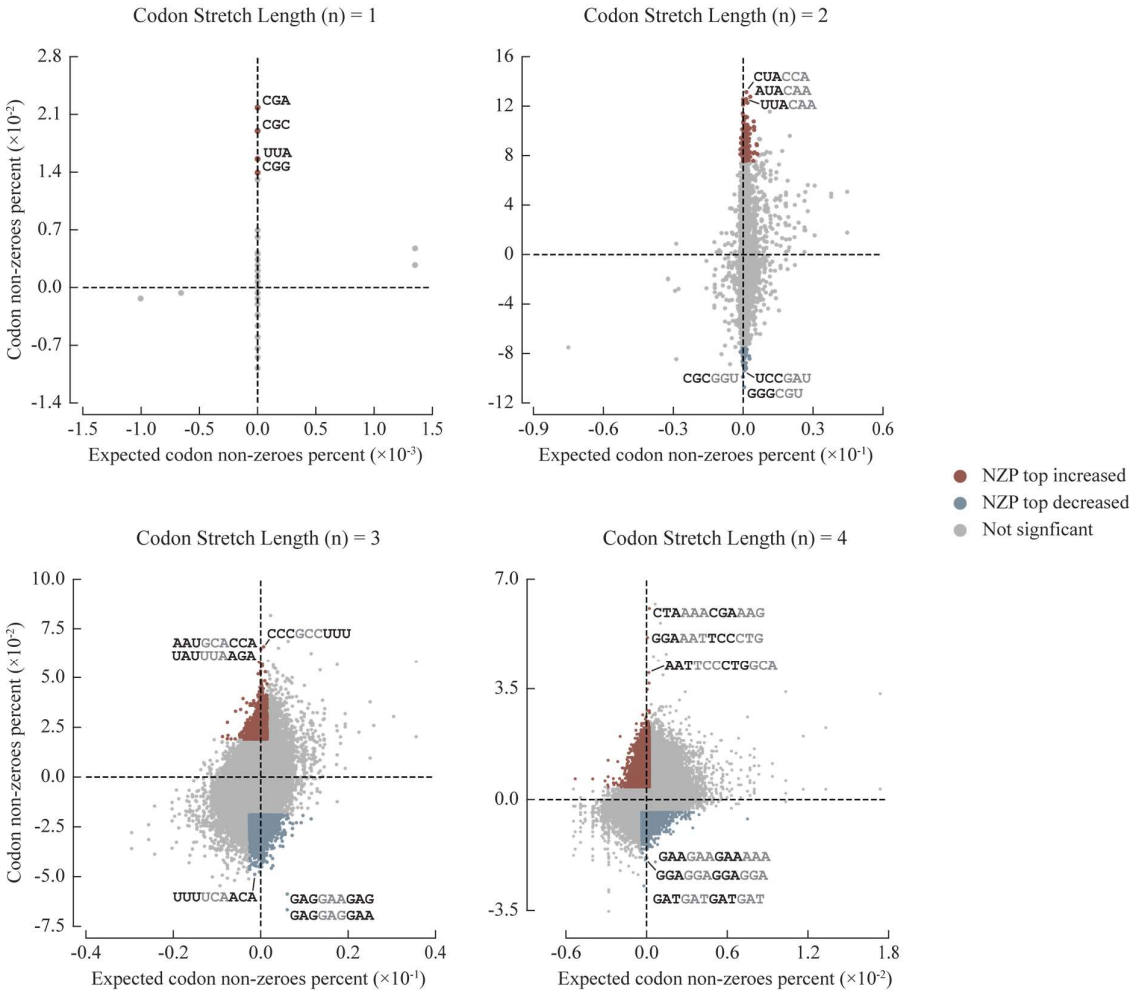


Figure 24. Critical codon stretches identified in NZP-n analysis ($n \leq 4$) from virus genomes between human and other vertebrate viruses. The critical codon stretches are those ranked within the top 5% in absolute NZP values, which shared the same directional sign (+/-) and ranked within the top 5% in expected NZP values.

In the NZP-1 analysis, four specific codons, CGA (Arg), CGC (Arg), UUA (Leu), and CGG (Arg), were identified as occurring more frequently in the gene coding sequences of human virus genomes, while none of these codons showed increased frequency in vertebrate virus genomes. The NZP-2 analysis revealed that 131 codon stretches, such as CUA|CAA (Leu|Gln), AUA|CAA (Ile|Gln), and UUA|CAA (Leu|Gln), were more prevalent in human virus genomes, whereas 43 stretches, including GGG|CGU (Gly|Arg), CGC|GGU (Arg|Gly), and UCC|GAU (Ser|Asp), were more common in vertebrate virus genomes. In the NZP-3 analysis, 2,840 codon stretches, including CCC|GCC|UUU (Pro|Ala|Phe), AAU|GCA|CCA (Asn|Ala|Pro), and UAU|UUA|AGA (Tyr|Leu|Arg), are more prevalent in human viruses, and 6,366 stretches, such as GAG|GAG|GAA (Glu|Glu|Glu), GAG|GAA|GAG (Glu|Glu|Glu), and UUU|UCA|ACA (Phe|Ser|Thr), being more common in vertebrate viruses. Further extending the analysis to NZP-4, there were 185,234 codon stretches, such as CUA|AAA|CGA|AAG (Leu|Lys|Arg|Lys), GGA|AAU|UCC|CUG (Gly|Asn|Ser|Leu), and AAU|UCC|CUG|GCA (Asn|Ser|Leu|Ala), identified more frequent in human viruses, in contrast to 65,687 stretches like GAU|GAU|GAU|GAU (Asp|Asp|Asp|Asp), GAA|GAA|GAA|AAA (Glu|Glu|Glu|Lys), and GGA|GGA|GGA|GGA (Gly|Gly|Gly|Gly) found more frequent in vertebrate viruses. Notably, codon stretches comprising multiple identical amino acids such as GAG|GAG|GAA (Glu|Glu|Glu), GAG|GAA|GAG (Glu|Glu|Glu), GAU|GAU|GAU|GAU (Asp|Asp|Asp|Asp), GGA|GGA|GGA|GGA (Gly|Gly|Gly|Gly) are standing out, raising the question of whether the biases in codon usage for continuous identical amino acids stretches play a significant role in determining virus host ranges.

2.4. Discussion

2.4.1. Summary of research finding in chapter 2

In this chapter, I describe the development of a bioinformatics pipeline designed specifically to analyse codon usage biases within virus genome sequences. This robust pipeline can be applied universally to investigate distinct characteristics in codon usages biases across various groups of coding sequences from any source. Its development involved a comprehensive approach to decode the complexity of codon distribution patterns, facilitating a deeper understanding of genomic adaptations and evolutionary dynamics among viruses. This tool not only enhances our capability to study viral genomes but also offers a methodological framework that can be adapted to a wide range of genetic research scenarios.

In my research, I employed independent T-tests to identify significant differences in amino acid and codon usage between human and non-human viruses, analysing data on Amino Acid%, Codon%, and RSCU. Through these analyses, I aimed to uncover specific properties of amino acids and codons that are associated with these usage distinct characteristics, achieved using Spearman correlation methods. Furthermore, I extended similar analytical methodologies to distinguish key codons and amino acids in viruses having distinct host range labels, based on their codon usage biases. The findings suggest the presence of a potential boundary in codon usage biases, which could contribute to virus host range identification. This insight could be critical in understanding host specificity and cross-species transmission in viral pathogens.

My research has demonstrated that viral genomes exhibiting different host ranges manifest distinct codon usage biases. Specifically, AU-rich codons are generally more prevalent in human viruses. Furthermore, codons ending in A/U at the 3rd nucleotide position typically exhibit characteristics that differ markedly from those ending in G/C, with the notable exceptions of UUG (Leu) and AGG (Arg). These findings indicate the critical role of the wobble position in viral codons. Additionally, the significant enrichment of A/U-ended codons, particularly those ending in A, suggests that viral RNA transcripts infecting humans may be more susceptible to A-to-I editing in general and at the wobble position. This post-transcriptional modification could potentially affect virus adaptability and immune evasion.

However, the exact implications of this susceptibility require further detailed investigation to fully understand its impact on virus-host interactions and pathogenicity.

In addition to identifying codon usage biases, my study further investigates into the relationship between these biases in human viruses and ADAT editing, focusing specifically on codons related to ADAT-mediated adenosine deamination. I discovered that several codons corresponding to the ADAT-related amino acids, namely Serine, Leucine, Alanine, Proline, and Threonine, are preferentially utilised in human viruses, no matter the codons are suppressed or benefited from ADAT editing. This preference suggests that an overexpression of ADAT during viral infections in human cells could potentially enhance the translation of these amino acids, given that codons suppressed by ADAT editing originally correspond to tRNAs with higher abundance. Contrastingly, the usage pattern of ADAT-related codons for Arginine markedly diverges from the other ADAT-related amino acids, with these codons being less preferred in human viruses. This observation indicates that the translation efficiency of Arginine might not be significantly impacted by ADAT overexpression. However, there are more factors could contribute to the impacts on the virus gene expression efficiency under ADAT editing, including tRNA abundance. tRNA expression level, ADAT editing efficiency, ADAT expression level et al.

The dimensional reduction analysis of the RSCU matrix across viruses with varying host ranges revealed notably distinct patterns for bacteriophages, compared to other virus groups that showed more homogeneous patterns. This observation suggests that identifying boundary of codon usage biases between viruses of different host ranges can be challenging, particularly beyond bacteriophages. This leads to an interesting possibility whether these codon usage patterns be distinguishable using machine learning models. By incorporating additional virus features into these models, it may be possible to refine our predictions and better understand the contribution of virus host ranges and codon fitness.

Last but not least, I developed a Python-based tool, the MultiCodonAnalyser (MCA), designed to investigate the usage biases metrics RSMCU-n and NZP-n for multi-codon stretches with a codon stretch length (CSL) of *n*. Utilising MCA, I was able to identify significant usage biases across codon stretches of CSL ranging from 1 to 4 between human and other vertebrate virus genomes. Despite these achievements, the specific characteristics and properties of the codon stretches that exhibit usage biases remain largely unexplored and need further investigations.

2.4.2. Discussion of research findings and potential improvement in the future investigations

In my analysis to identify significant differences in the usage of codons and amino acids between human and non-human viruses, I primarily employed the independent T-test on metrics such as Amino Acid%, Codon%, and RSCU. A key concern with using the T-test on RSCU data arises from the nature of RSCU calculation, which involves division of two variables presumed to be normally distributed. The T-test fundamentally assumes that the data adheres to a normal distribution and that variances between groups are equal, conditions that RSCU data might not meet due to its calculated nature (unknown distribution). While I utilised the Welch T-test as an alternative, which does not assume equal variances, this adaptation still does not completely fix the issue since the normality of the RSCU distribution remains uncertain. Despite these methodological concerns, the significance of the findings is supported by the large sample size of the study (10,820 samples), which tends to relieve some of the limitations associated with smaller datasets.

To enhance the robustness of our conclusions, I propose using additional statistical methods that do not assume normal distribution or equal variances. The Mann-Whitney U-test, or Wilcoxon rank-sum test, is particularly suitable as it is based on the ranking of data points rather than their distribution parameters. Employing this non-parametric method would provide a more justified evaluation of the differences in amino acid and codon usage between human and non-human viruses, potentially confirming the initial findings derived from T-test analyses. This approach would thereby strengthen the evidential basis of the study, ensuring that any identified differences in codon or amino acid usage are both statistically significant and methodologically sound.

Another limitation of the current analysis arises from the methodology employed to compute the codon usage metrics for virus genomes. Specifically, the analysis involves aggregating codon counts across all coding sequences (CDS) within a virus genome. This approach, while simplifying the analysis, may not accurately reflect the true codon usage bias of virus genomes. The primary issue with this method is its potential bias toward longer coding sequences. Since different CDS vary significantly in length, those with greater lengths and, consequently, more codons, disproportionately influence the overall codon usage metrics such as RSCU. This

aggregation could skew the results, overemphasising the codon usage patterns of longer sequences while underrepresenting those of shorter ones. To address this shortcoming, a more refined approach could involve weighting the codon counts by the length of each CDS. This adjustment would ensure that each sequence contributes proportionally to the overall codon usage metrics based on its length, thereby reducing the bias towards longer sequences. Alternatively, analysing codon usage within individual CDS before aggregating these data could provide a more detailed and balanced view of codon preferences across the entire genome. Such methodological enhancements would likely yield a more accurate and representative analysis of codon usage biases in virus genomes, enhancing the reliability of the findings.

In addressing the potential biases arising from uneven sample sizes among different virus groups, I opted to use the RefSeq database, which is a curated collection of reference sequences. This approach was intended to mitigate the skew in data caused by overrepresentation of certain viruses (i.e. too much data of SARS-CoV-2). Despite this strategy, differences in sample sizes across various virus groups persist, which could still influence the analysis results. To further counteract these imbalances, I implemented resampling techniques, specifically bootstrapping. This method allows for the estimation of sample distribution properties through repeated random sampling with replacement, thereby providing a more robust statistical analysis that is less dependent on the original sample size distributions.

CHAPTER 3

Predicting Virus Host Ranges with Virus Codon Usage Biases

by Shuquan (Steve) Su

Keywords:

Virus codon usage biases, Virus host range, Machine learning

Chapter 3. Predicting Virus Host Ranges with Virus Codon Usage Biases

3.1. Introduction

Chapter 2 successfully illustrates that features like RSCU exhibit significant distinctions between viruses with different host ranges. This finding leads to an intriguing question whether these virus codon usage biases possess predictive capabilities regarding the host ranges of viruses, or the codon fitness in viruses adapted to specific hosts.

3.1.1. Introducing codon usage biases of virus genomes and virus codon fitness in certain host

During the COVID-19 pandemic, massive virus genome sequencing data were generated from environmental sampling to identify critical mutations and to monitor the evolution of SARS-CoV-2 during the pandemic with development of sequencing technology and incredible efforts of scientists[340], especially by small-size sequencing equipment such as Nanopore MinION sequencer allowing scientists to sequence virus genome on-site directly after sample harvest[341, 342]. An important question after sample collection is often that whether the virus infect human, or what is the virus host range for the on-site scientists to identify the virus' potential threat.

Host range can be defined as a group of host species where the pathogen can proliferate. It is one of the most important concepts helping understand pathogen epidemiology and evolution. A pathogen's host range is difficult to characterise due to the lack of quantitative measurements to define a host range, which further leads to ineffective predictions for early precaution alerts. Host range shifting is a chain of changes in host ranges (i.e. non-human to human) of a virus. Currently, there is no effective methodology to predict characteristics of host shifting between

different virus strains (i.e., critical mutation of significant host range change); instead, the evolution paths are mainly studied through constructing phylogenetic trees using the viruses' genomes after a host range shifting happened[343, 344]. Historically, many harmful virus outbreaks are attributed to their unknown host range shifting, especially towards human, such as the MERS-CoV epidemic (2012, over 800 deaths) from bats or dromedary camels[345, 346], H1N1 influenza virus pandemic (2009, over 280,000 deaths) from swine[347, 348], SARS-CoV epidemic (2003, over 700 deaths) from horseshoe bats or palm civets[349-351]. Thus, lacking quantitative measurements of virus host ranges may be the stumbling block of virus evolution study. Although a lot of research have been using sequencing method with environmental sampling of viruses to study potential threat, there are still limited computational methods to predict movements in host range shifting[340].

There are various potential determining factors in virus host range such as codon fitness to hosts[352], mechanism entering hosts[353, 354], immune evasion mechanisms[355] et al. Because translations of viral genes rely dramatically on host translational machinery, and codon fitness is a correlation between virus codon usage biases and host tRNA pool, thus incompatibilities in codon fitness will eventually lead to inefficiency in virus proteins translation and failure in virus proliferation[356, 357]. Thus, virus codon fitness (VCF) is one of the most vital determining factors to virus host range, which has huge potential in virus host range prediction. Host tRNA pool is dramatically affected by the host genotypes while it is still difficult to represent it at the species level consisting of different host genotypes with significant variety. Although I may set a reference genome with certain individual, the reference genotype may be insusceptible to certain human virus but not to the majority population. Thus, I propose to study virus host codon fitness with virus codon usage bias directly from viral genomes and virus host range label for generalisation.

Virus codon usage bias, as a major metric for host translational adaptions, is the key property of coding sequences to decide intracellular translation efficiency[324-326], and the intracellular translational efficiency of viral proteins directly determines the efficiency of virus replications[358-360]. I hypothesise that the virus codon biases would have relation to the host translational mechanism (i.e., tRNA pool), and generally reflect the adaptation level if studied by machine learning, which therefore could be used to predict viral host codon fitness. There are many metrics to study virus codon biases such as Relative Synonymous Codon Usage

(RSCU)[328, 329], Codon Adaptation Index (CAI)[329], tRNA adaptation index (tAI)[330], et al. However, most of them required gene expression level of host genes as reference, which may lead to extra biases in species-scale representation and in later prediction. Relative synonymous codon usage, or RSCU, is a statistical propensity parameter representing essential biases of the codon usages in a coding sequence[328, 329], which is purely computed from coding sequences without computational loss. RSCU preferences have been studied in individual viruses including SARS-CoV-2[331], Flaviviridae Virus[332], Zika virus[333], and Transmissible Gastroenteritis Virus[334]. However, most of these studies are only focused on the statistical analysis of the RSCU contents of the individual virus in an aim to find RSCU correlations between the individual viruses and their host labels[332, 361-364]. The integrative RSCU contents and preferences about the collection of all the viruses have never been systematically examined, and there is no study aiming to bridge the gap between the virus codon biases and the viral host codon fitness. Although the micro-environments of virus-host interactions are extremely complicated and they should be clearly distinct between different species of viruses[365, 366], it is possible through competent machine learning algorithms[367] to discover previously unknown rules underlining the association between codon usage biases and the viral codon fitness in hosts.

3.1.2. Introducing Random Forest model and potential use in studying virus codon usage biases

In this study, I propose to use tree-based machine learning algorithms such as random forest (RF) to establish accurate models predictive to the probability of virus host codon fitness with RSCU of virus genomes and other virus genome composition properties as input data. This classification technique, as empowered by entropy or information gain dichotomy, is specially used by this study due to their advantages in dealing with non-linear features such as RSCU, which the RF model is a committee of different Decision Tree models making the prediction by voting. Additional important features of the input data include coding sequences (CDS) length profiles, and virus taxonomy classifications. The tree-based algorithms are a branch of supervised machine learning technique, where each tree is a dichotomy hierarchy structure of true/false decision-making rules for deciding the output classification according to the input

feature values. Here, I propose using the predicted probability from trained RF model as representative readout score for virus codon fitness (VCF) in certain host range (i.e. human). In this study, the human virus codon fitness score, or HVCF score, predicted from the trained RF model for the human host is further explored for virus genomes sequence data from different sources, and for monitoring of SARS-CoV-2 human VCF shifting during COIVD-19 pandemic. Moreover, I attempted to predict codon-based evolution path of SARS-CoV-2 from other Betacoronavirus through examining changes in HVCF when applied mutations. I have found that the virus codon biases, and machine learning models can serve as measurements in defining the boundaries of virus host codon fitness and can make predictions for virus host codon fitness shifting and virus evolutionary path.

3.2. Methods

3.2.1. Fundamental computational environment

The fundamental computational environment is identical to Chapter 2 (See section 2.2.1.).

3.2.2. Acquisition of additional virus genome sequences

Complete virus genome sequence data (non-RefSeq) of MERS-CoV, Zaire Ebolavirus, West Nile virus, Zika virus, Orthohantavirus, Influenza A virus, Henipavirus, Lyssavirus Rabies, and SARS-CoV-2 were acquired for other use (i.e. test trained model). The accession IDs of all those virus genomes used by this study are acquired from the NCBI Virus database[368], which other information such as 'Host', 'Pangolin' et al were also downloaded there (incomplete genomes were discarded). The accession IDs was used to download coding sequences through Biopython toolkit. Total 639 genome sequences of MERS-CoV (256 human-sourced, 383 non-human-sourced), 563 genome sequences of Zaire Ebolavirus (435 human-sourced, 128 non-human-sourced), 1823 genome sequences of West Nile virus (137 human-sourced, 1686 non-human-sourced), 240 genome sequences of Zika virus (208 human-sourced, 32 non-human-sourced), 826 genome sequences of Orthohantavirus (142 human-sourced, 684 non-human-sourced), 614 genome sequences of Influenza A virus (131 human-sourced, 483 non-human-sourced), 55 genome sequences of Henipavirus (35 human-sourced, 20 non-human-sourced), 1862 genome sequences of Lyssavirus Rabies (30 human-sourced, 1832 non-human-sourced), and 755151 genome sequences of SARS-CoV-2 (all human-sourced) were retrieved. The WHO Name information related to SARS-CoV-2 was acquired from 'cov-lineages.org' database[369].

3.2.3. Calculation of other data for applying additional features in training machine learning models

Besides RSCUs, other feature datasets were used to achieve better machine learning prediction. Datasets 'Codon%' ($D_{Codon\%}$) and 'AminoAcid%' ($D_{AminoAcid\%}$) were simply calculated with percentages of different codons or amino acids in total codon count of amino acid count of virus genome. Stop codons were ignored in both $D_{Codon\%}$ and $D_{AminoAcid\%}$ (thus 61 features in $D_{Codon\%}$ and 20 features in $D_{AminoAcid\%}$, stop codons were not included). Dataset 'ATGC%' ($D_{ATGC\%}$) was simply calculated with frequency of each nucleotide (A%, U%, G%, C%, AU%, GC%), which AU% and GC% were calculated by summing A%/U% and G%/C%. Dataset 'Start-Stop Codon%' ($D_{StartStopCodon\%}$) is the calculated with frequency of the start codon (AUG%) in all the CDS of virus genomes, and the frequency of each stop codons (UAA%, UAG%, UGA%) in all the CDS of virus genomes. Dataset 'CDS Length' ($D_{CDS\ Length}$) consists of features genome length, Concatenated CDS length, CDS count, and the mean and standard deviation of CDS length, which Concatenated CDS length is the sum of all CDS length. Dataset 'HumanCorr' ($D_{HumanCorr}$) is correlation coefficients calculated between virus RSCU and Human reference RSCU, which is Human reference RSCU acquired from the CoCoPUTs database[370]. Dataset 'HumanCorr(AA)' ($D_{HumanCorrAA}$) dataset is correlation coefficients calculated among between virus RSCU and Human standard RSCU from each Amino Acids. Met (M) and Trp (W) as well as Stop codons were ignored. Dataset 'Partite' ($D_{Partite}$) consist of the virus classifications of either mono-partite or multi-partite. Dataset 'Taxonomy' ($D_{Taxonomy}$) consist of tertiarily encoded data based on taxonomy information acquired from the NCBI Taxonomy database with Realm (or Clade), Kingdom, Phylum, Class, Order. The positive samples were labelled as '1' and negative samples were labelled as '-1', which unknown samples were labelled as '0'. The raw dataset of Taxonomy without encoding was noted as $D_{Taxonomy\_Raw}$, which may be used in some study.

3.2.4. Data normalisation

The raw RSCU data matrix, and the transposed RSCU data matrix were both normalised through the z-score normalisation method through machine learning package Scikit-learn[371].

3.2.5. Principal Component Analysis

The normalised (z-score) matrixes were transformed by Principal Component Analysis (PCA), where the cut-off threshold was set as 1.0 (no loss in explained variance) through Scikit-learn[371]. This method can compress the data without loss of the variances by removing the redundant features.

3.2.6. Uniform Manifold Approximation and Projection

Dimensional reduction analysis was performed on the normalised and compressed data using the Uniform Manifold Approximation and Projection (UMAP) algorithm[338] for both RSCU data and transposed RSCU data. This method visualises data points in high dimensional space with better preservation of global structure compared to other algorithms like t-SNE.

3.2.7. Synthetic Minority Oversampling Technique

Because the sample sizes were imbalanced with different host range labels. Different resampling method was used on train dataset generated by train-test split for increasing model predicting accuracy and training better classifier model.

The train datasets were resampled by the Synthetic Minority Oversampling Technique (SMOTE)[372] to overcome sample imbalance. To be brief, SMOTE finds the k-nearest neighbors (k=5) of this data point in the feature space, typically using Euclidean distance. Then, it generates synthetic samples along the line connecting the selected data point and its randomly chosen neighbours.

3.2.8. Random forest classifiers

The Random Forest (RF) models were trained with SMOTE-resampled train datasets with Scikit-learn[371], when Balance accuracy, F1 score, Recall scores and ROC-AUC scores were

used as the standard of the model performance due to sample imbalances. The open-source OPTUNA framework was used for hyper-parameters tunning with specific trials for different scenarios (20 or 50 trials)[373]. The predicted probabilities from the models' predictions were considered as the readout of infection probability to a specific host, which is acquired from embedded function of Scikit-learn. The feature importance metrics are also acquired from embedded function of Scikit-learn.

3.2.9. Other machine learning classifiers

Besides RF models, other classifier models were also tested at the beginning including Decision Tree classifiers (DT), Support Vector Classifiers (SVC), Gaussian Process Classifiers (Gau), and Neural Network (NN). They were all trained through Scikit-learn with default hyperparameters to check if prediction performances are better.

3.2.10. Leave-One-Out machine learning technique

To further confirm reliability of using RSCU and other features in predicting human virus codon fitness scores (HVCF) of viruses, the Leave-One-Out (LOO) method was carried out, which all other samples were used to train a RF model in predicting one test sample. The same process was carried out separately to all the samples, and only 5 trials were used in OPTUNA hyper-parameters tunning for less computational cost. Model performances such as Balanced accuracy and Recall score were generated with summary of total 10820 predictions (correct/wrong predictions for 10820 samples or models). The models with wrong predictions were later re-trained with 50 trials setting in OPTUNA hyper-parameters tunning to see whether it will have correct predictions (Supplemental figure 3).

3.2.11. Simulation of SARS-CoV-2-directed codon fitness shifting path

The predicted probability from the D_{RTC}-trained Recall-optmised RF model trained with all 10820 samples was considered as the readout of human virus codon fitness scores (HVCF)

because it has best prediction performance. To predict an evolution path between two viruses, the start-point virus and the end-point virus were determined between SARS-CoV-2 and a target Betacoronavirus for either the Forward or Backward mutation simulation. The Forward mutation is from the target Betacoronavirus to SARS-CoV-2, while the Backward mutation is from SARS-CoV-2 to the target Betacoronavirus. At each mutation step in the evolutionary process simulation, every possible mutation was applied to the codon count compositions of the virus, including all possible substitution (i.e. AAA→AAG), addition (i.e. +AAA), or deletion (i.e. -AAA) of codons. The D_{RSCU} and $D_{CDS\ Length}$ of the updated virus codon compositions were re-calculated except for the $D_{Taxonomy}$ (remained as Coronaviridae). The new HVCF was then predicted with the updated D_{RTC}. Among new HVCFs derived from all the possible mutations, the simulated mutation generating the lowest/highest HVCF (depending on simulation direction, Forward or Backward) was selected. When multiple mutations have the same lowest/highest HVCF, the additional analysis of correlation coefficient was calculated between updated D_{RSCU} and D_{RSCU} of the end-point virus (SARS-CoV-2 in the Forward path or the target Betacoronavirus in the Backward path). The simulated mutation generating the best correlation coefficient was selected.

Similar to the gradient descent, this process was step-by-step repeated until reaching the HVCF of the end-point virus. In some cases, the simulation may reach a stagnation because of possibility of mutually contradictory mutations (i.e., AAA→AAG then AAG→AAA). To avoid such meaningless loop stagnation, mutually contradictory mutations were forbidden in the simulation process. For instance, if an ongoing evolutionary path has AAA→AAG, then mutation AAG→AAA must be excluded in the subsequent simulation.

3.3. Results

3.3.1. Prediction of virus host labels through machine learning with RSCU matrix and other virus genome characteristics

To assess the effectiveness of various machine learning algorithms in predicting virus host ranges using RSCU data, I conducted an exploratory analysis employing several classifier models including Decision Tree Classifiers (DT), Random Forest (RF), Support Vector Classifiers (SVC), Gaussian Process Classifiers (Gau), and Neural Network Classifiers (NN). The performance of these models was evaluated using a range of metrics including Accuracy, Balanced accuracy, F1 score, ROC-AUC score, Average precision, and PR-AUC score (Figure 25). The Random Forest (RF) model demonstrated superior performance across most metrics. This superior performance likely stems from the RF model's ability to handle the non-linear relationships inherent in RSCU data effectively, as well as its robustness to overfitting compared to simpler models like Decision Trees. Based on these findings, I decided to primarily utilise the RF model for further training and analysis.

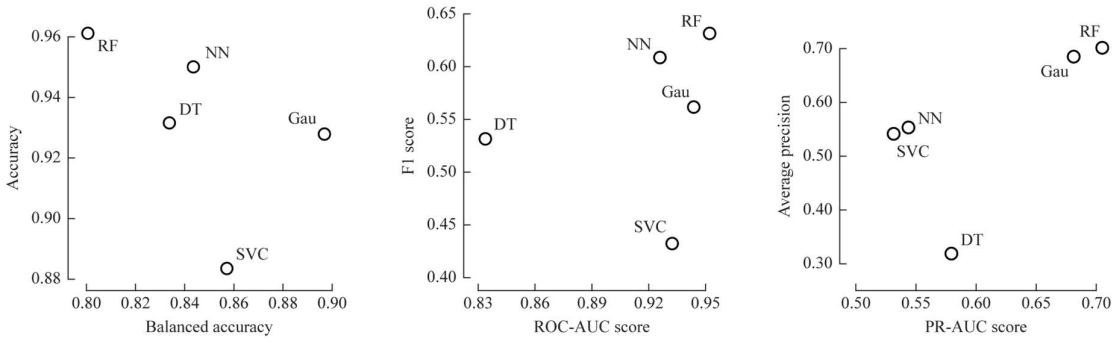


Figure 25. Performances of different trained classifier algorithms. Performance metrics such as Accuracy, Balanced accuracy, F1 score, ROC-AUC score, Average precision, and PR-AUC score are used to evaluate model performances.

I then applied RF machine learning algorithms to use the RSCU datasets of virus genomes in predicting whether a viral genome has or does not have strong potential to infect a certain host due to affinised VCF. The classification labels were binary, distinguishing between potential hosts and non-hosts (e.g., human vs. non-human). To address the challenge of sample imbalance and enhance the prediction accuracy on the test data, I utilised the SMOTE to balance the classes within the training dataset. Other resampling methods were tested but SMOTE has the best performance (data not shown). These datasets, specifically composed of RSCU features, are referred to as D_R (or D_{RSCU}). The RF models trained on these datasets demonstrated robust accuracy in predicting VCF across various host ranges, including humans, vertebrates, invertebrates, land plants, and bacteria. The models were evaluated across different train-test-split ratios, with results indicating that prediction accuracy improves with larger sizes of training data (Figure 26). Notably, even with extremely low train data ratio of 0.05, the accuracies are all better than blind guessing (0.5 in accuracy), validating the reliability of using RSCU data for predicting VCF in different host range.

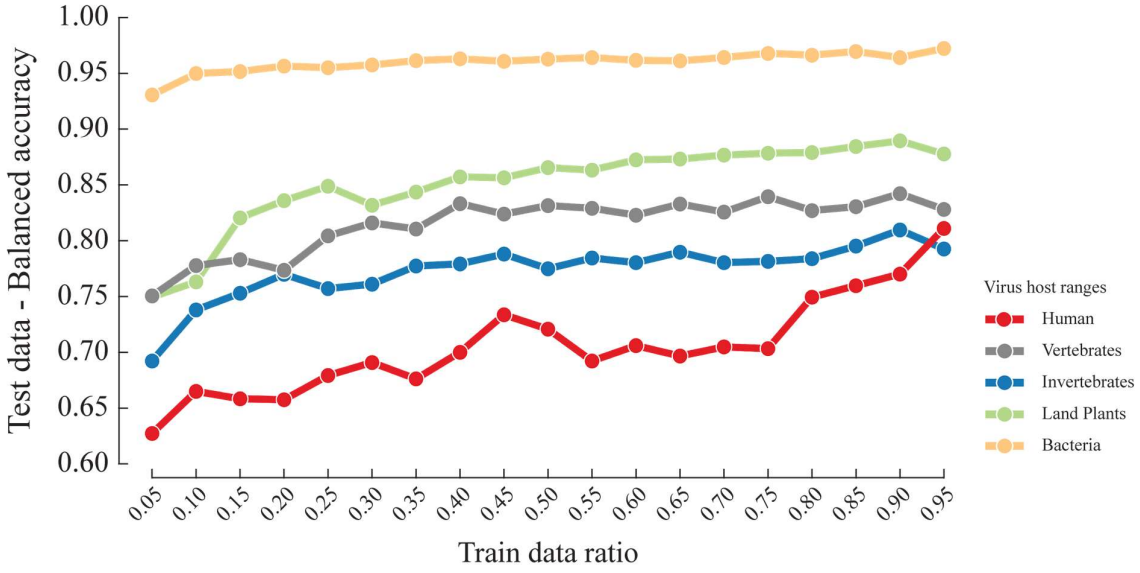


Figure 26. Performances of trained random forest models to predict different hosts. The balanced accuracy of models trained with D_{RSCU} with different train-test-split ratios, which are better than blind guessing (0.5 accuracy) even with extremely low train data ratio of 0.05.

To enhance the accuracy of the machine learning models used for predicting viral host specificity, I incorporated additional features beyond the basic RSCU data (Figure 27). These features include datasets on taxonomy and CDS length of viruses, aiming to provide a more comprehensive representation of viral characteristics. The expanded datasets used in training are listed below including $D_{Codon\%}$, $D_{AminoAcid\%}$, $D_{Taxonomy}$, $D_{Taxonomy\_Raw}$, D_{Human_Corr}, $D_{Human\_Corr\_AA}$, $D_{CDSLength}$ et al (Figure 27.A). Initial trials with various combinations of these datasets revealed that including the taxonomy dataset ($D_{Taxonomy}$) alongside the RSCU data (denoted as D_{RT} or $D_{RSCU\text{-}Taxonomy}$) significantly improved the balanced accuracy of the models (Figure 27.B). Encouraged by these results, I further explored the addition of the CDS Length dataset ($D_{CDSLength}$) to this combination. The incorporation of CDS length data alongside RSCU and taxonomy datasets (denoted as D_{RTC} or $D_{RSCU\text{-}Taxonomy\text{-}CDS\ Length}$) further enhanced the predictive performance of the models (Figure 27.C). This approach underscores the importance of a multi-dimensional dataset that captures various aspects of viral genomes, improving the models' ability to accurately reflect and predict the complex nature of virus-host interactions based on genomic signatures.

To enhance the accuracy of the machine learning models used for predicting viral host specificity, I incorporated additional features beyond the basic RSCU data (Figure 27). These features include datasets on taxonomy and CDS length of viruses, aiming to provide a more comprehensive representation of viral characteristics. The expanded datasets used in training are listed below including $D_{Codon\%}$, $D_{AminoAcid\%}$, $D_{Taxonomy}$, $D_{Taxonomy\_Raw}$, D_{Human_Corr}, $D_{Human\_Corr\_AA}$, $D_{CDSLength}$ et al (Figure 27.A). Initial trials with various combinations of these datasets revealed that including the taxonomy dataset ($D_{Taxonomy}$) alongside the RSCU data (denoted as D_{RT} or $D_{RSCU\text{-}Taxonomy}$) significantly improved the balanced accuracy of the models (Figure 27.B). Encouraged by these results, I further explored the addition of the CDS Length dataset ($D_{CDSLength}$) to this combination. The incorporation of CDS length data alongside RSCU and taxonomy datasets (denoted as D_{RTC} or $D_{RSCU\text{-}Taxonomy\text{-}CDS\ Length}$) further enhanced the predictive performance of the models (Figure 27.C). This approach underscores the importance of a multi-dimensional dataset that captures various aspects of viral genomes, improving the models' ability to accurately reflect and predict the complex nature of virus-host interactions based on genomic signatures.

A.

Dataset	Description	Total sample size	Total feature size	Data composition
D_{RSCU} / D_R	Codon usage biases	10820	59	Unknown distribution
$D_{Codon\%}$	Codon percentages	10820	61	Assumed normal distribution
$D_{AminoAcid\%}$	Amino acid percentages	10820	20	Assumed normal distribution
$D_{Taxonomy}$	Taxonomy information (encoded)	10820	344	1, 0, -1
$D_{Taxonomy\_Raw}$	Taxonomy information	10820	7	Strings
D_{Human_Corr}	RSCU correlation with human	10820	3	Unknown distribution
$D_{Human\_Corr\_AA}$	RSCU correlation per AA with human	10820	36	Unknown distribution
$D_{CDSLength}$	CDS Length information	10820	5	Unknown distribution
D_{RT}	Concated D_{RSCU} and $D_{Taxonomy}$	10820	403	-
D_{RTC}	Concated D_{RSCU}, $D_{Taxonomy}$, $D_{CDSLength}$	10820	408	-

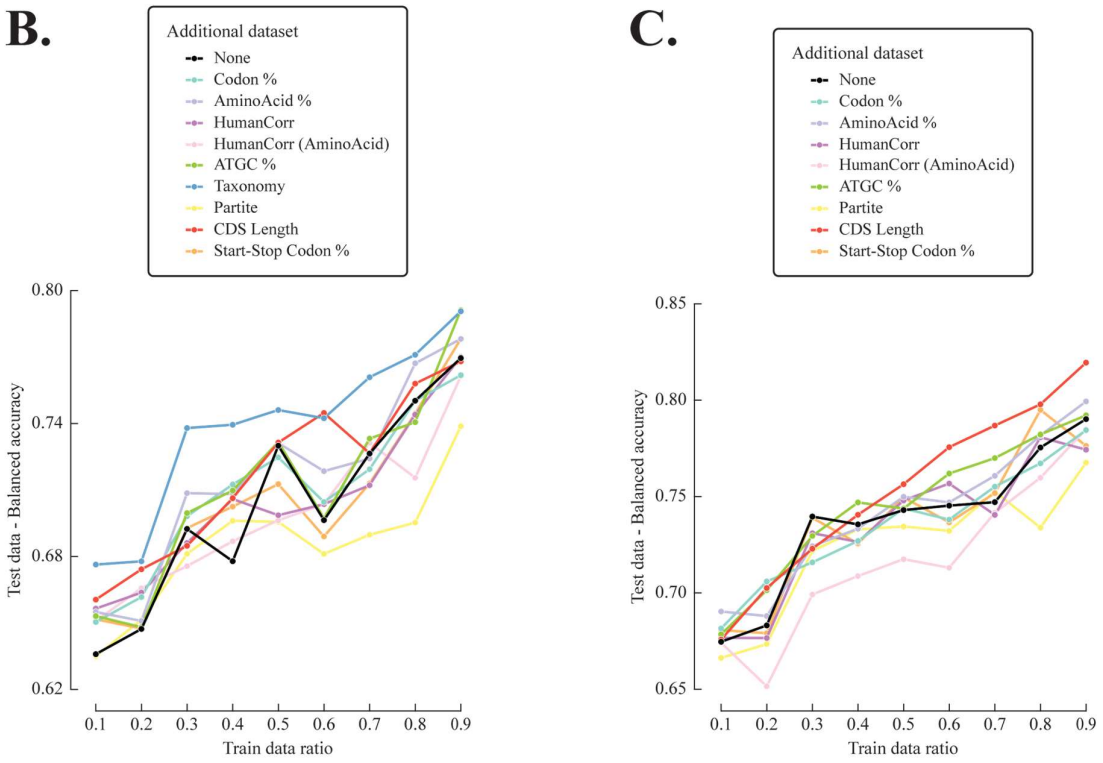


Figure 27. Additional feature dataset selections. (A) Information of different additional datasets to the input features. (B) Additional features for D_R or D_{RSCU}. Adding Taxonomy dataset has best performances compared to others. (C). Additional features for D_{RT} or $D_{RSCU\text{-}Taxonomy}$. Adding CDS Length dataset has best performances compared to others.

The models, trained with a substantial proportion of the data (train data ratio = 0.9), exhibit remarkable performance metrics, including Balanced accuracy, F1 score, and ROC-AUC score, which are detailed in Figure 28. The high accuracy of these models further proves the

hypothesis that viruses adapted to different hosts exhibit distinct codon usage biases. Consequently, I propose employing the predicted probabilities generated by these trained RF models as quantitative indicators of the Viral Codon Fitness (VCF) relative to corresponding hosts. This methodological approach allows for a nuanced assessment of how well a viral codon usage aligns with the translational machinery of its potential hosts, thereby providing insights into the likelihood of successful infection and replication within those hosts.

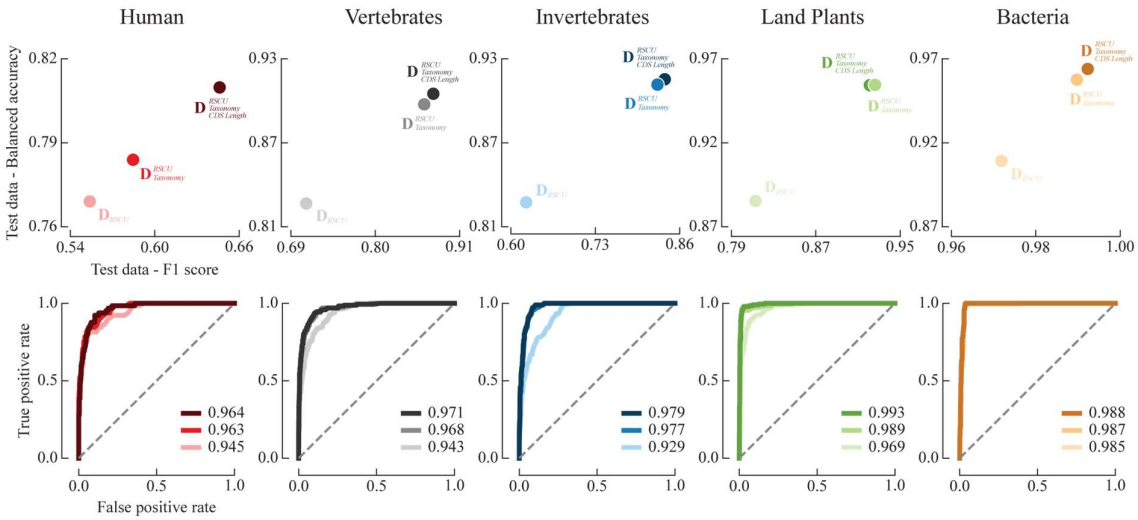


Figure 28. Performances of trained random forest models to predict different hosts based on different datasets. The model performances (Balanced accuracy and F1 score) and ROC curve of models trained with different datasets: D_R (D_{RSCU}), D_{RT} ($D_{RSCU\text{-}Taxonomy}$), D_{RTC} ($D_{RSCU\text{-}Taxonomy\text{-}CDS\ Length}$). The ROC-AUC scores are shown.

3.3.2. Leave-One-Out method verifies the model's reliability

To further verify the feasibility of creating practical tool in predicting human virus codon fitness score (HVCF score), the Leave-One-Out (LOO) train-test-split method was employed. This technique involves training the model on all samples except one and then testing on the excluded sample. This cycle repeats such that each sample is used once as the test dataset. The performance metrics including balanced accuracy, F1 score, recall score, among others, are

aggregated from the true or false predictions generated across all models. The optimisation of both balanced accuracy and recall score during hyperparameter tuning is crucial. Emphasis on enhancing the recall score is particularly significant as it may yield a model with greater sensitivity, reducing the likelihood of false-negative predictions about human viruses. The comprehensive performance metrics, including balanced accuracy and recall score, are displayed in Figure 29.A. Here, the D_{RTC} dataset consistently outperforms others, suggesting superior predictive capability to human host. Interestingly, no significant differences are found between models optimised for recall versus balanced accuracy, although the recall-optimised LOO training demonstrates slightly better results. Moreover, further refinements in model performance could potentially be achieved with more extensive hyperparameter tuning iterations (Figure 29.B). Subsequent analysis extends to assessing prediction performance across various key virus families especially those historically caused pandemic (Figure 30.A). Key virus families such as Coronaviridae, Filoviridae, Flaviviridae, Orthomyxoviridae, Paramyxoviridae, Poxviridae, and Retroviridae were evaluated, with the model demonstrating exceptionally low rates of false negatives and high predictive accuracy for these groups. This accuracy extends to pathogens responsible for past pandemics, with all but SARS coronavirus Tor2, which scored an HVCF of 0.482, just below the threshold for correct classification) being accurately predicted in LOO tests (Figure 30.B). These results verify the reliability of using the D_{RTC}-trained, Recall-optimised RF model to predict the host codon fitness of new virus genome sequences or even unknown viruses.

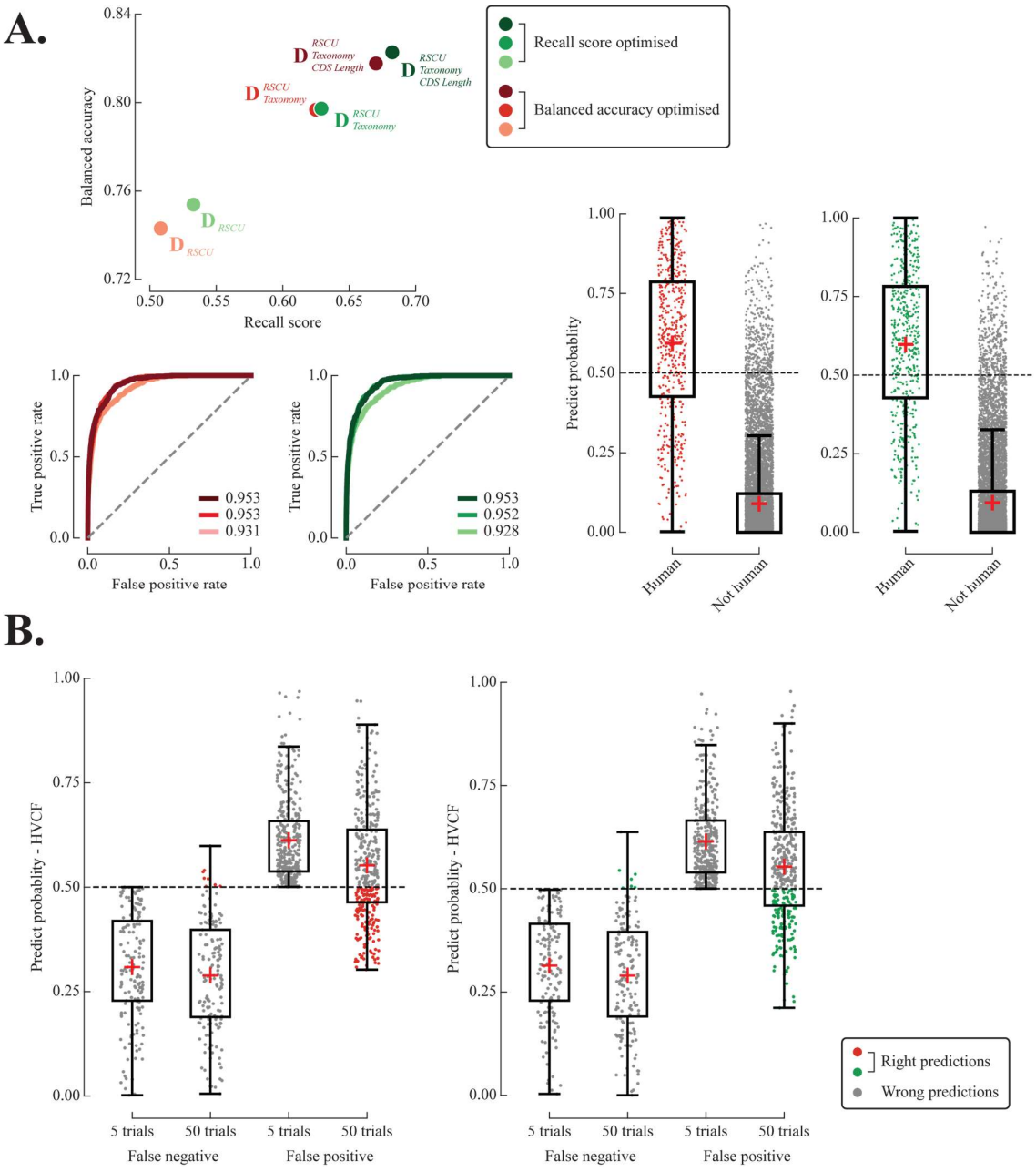
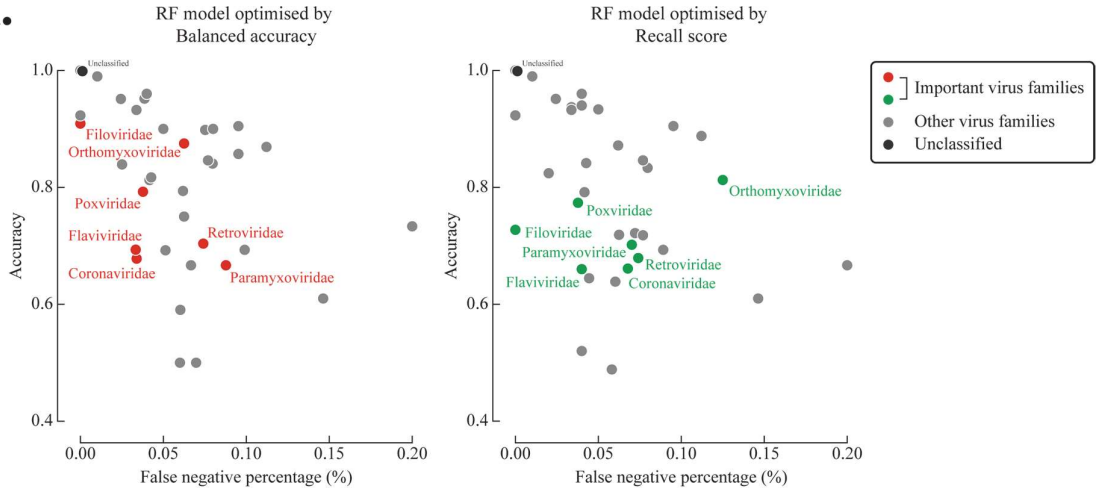


Figure 29. Leave-One-Out train-test-split method to prove possibility of generating predictive tool to VCF. The optimising score in hyper-parameters tuning is set either to balanced accuracy or recall scores. (A) Performances of all models trained by Leave-One-Out methods, including balanced accuracy and Recall score of different Datasets: D_R, D_{RT}, D_{RTC}. The ROC curve with ROC-AUC scores, and the boxplot of predict probabilities are also shown; (B) DRTC-trained LOO machine learning with 50 trials in OPTUNA optimisation on wrong predicted samples (either false positive or false negative) in 5-trials OPTUNA optimisation. 150 out of 519 previously falsely predicted samples have correct predictions in DRTC-trained balanced-accuracy-optimised modelling, while 142 out of 536 samples have correct predictions in recall-optimised modelling. The red crosses in all the strip-plot show the mean of the data.

A.



B.

Pandemic	NCBI Name	NCBI Accession	Taxonomy (Family)	HVCF score (5 trials)	HVCF score (50 trials)
COVID-19	Severe acute respiratory syndrome coronavirus 2	NC_045512	Coronaviridae	0.68137117	-
SARS	SARS coronavirus Tor2	NC_004718	Coronaviridae	0.41566970	0.48194337
MERS	Middle East respiratory syndrome-related coronavirus	NC_019843	Coronaviridae	0.80040527	-
Marburg virus	Marburg marburgvirus	NC_001608, NC_024781	Filoviridae	0.55311611	-
Ebola virus disease	Zaire ebolavirus	NC_002549	Filoviridae	0.58626284	-
Zika virus	Zika virus	NC_035889, NC_012532	Flaviviridae	0.80287746	-
Swine flu	Influenza A virus (A/California/07/2009(H1N1))	NC_026433, NC_026438, NC_026435, NC_026432, NC_026437, NC_026431, NC_026434, NC_026436	Orthomyxoviridae	0.92389210	-
Asian flu	Influenza A virus (A/Korea/426/1968(H2N2))	NC_007374, NC_007380, NC_007377, NC_007382, NC_007381, NC_007376, NC_007375, NC_007378	Orthomyxoviridae	0.92642469	-
Hong Kong flu	Influenza A virus (A/New York/392/2004(H3N2))	NC_007370, NC_007367, NC_007368, NC_007369, NC_007366, NC_007371, NC_007372, NC_007373	Orthomyxoviridae	0.69597415	-
Avian influenza	Influenza A virus (A/goose/Guangdong/1/1996(H5N1))	NC_007357, NC_007360, NC_007362, NC_007361, NC_007359, NC_007358, NC_007363, NC_007364	Orthomyxoviridae	0.86106302	-

Figure 30. Leave-One-Out machine learning results of critical virus families. (A) The prediction performances including accuracy and false negative percentages (%) towards important virus families. The important virus families contain the viruses that caused pandemic in human society in the history (i.e. Coronaviridae has SARS-CoV-2 causing COVID-19). Two RF models are tested, which are separately optimised by Balance accuracy and Recall score; (B) The detailed results in LOO machine learning analysis of viruses in previous pandemic to examine whether HVCF scores could provide clues for pandemic precaution. 5 trials are used in OPTUNA optimisation for all the data. 50 trials are used in samples that have predicted wrong labels in the previous model training.

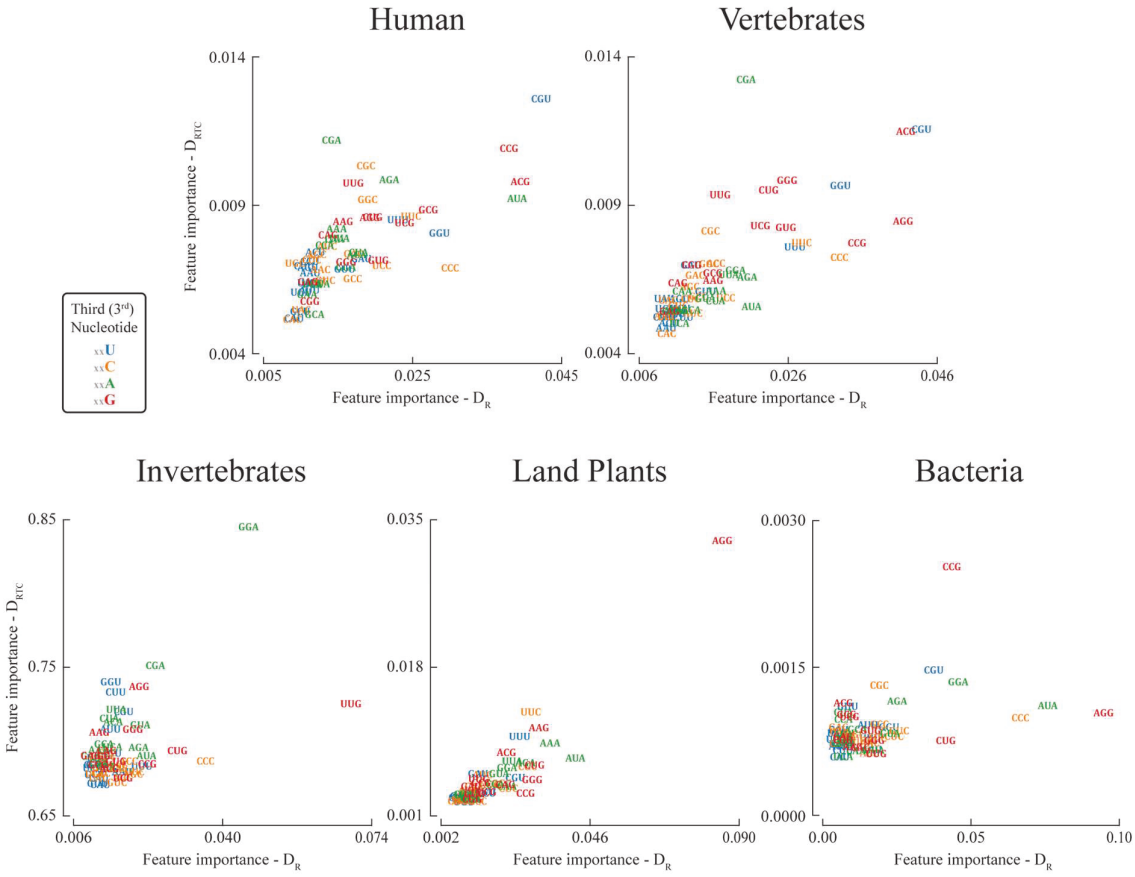


Figure 31. Feature importance of different codon RSCUs in D_{RTC}-trained Recall-optimised RF models trained with all 10820 samples (train data ratio = 1.0). Only codon features are shown from either D_R or D_{RTC} trained models, and the codon features are colour-coated based on their third nucleotide (3rd nt).

Figure 31 presents the feature importances extracted from the D_{RTC}-trained Recall-optimised RF model and the D_{RTC}-trained Recall-optimised RF model, predictive to different host labels. The analysis highlights the significant impact of specific codons usage bias on the model's predictive performance. Notably, the RSCU of codons like CGU (Arg), CGA (Arg) et al emerge as crucial features within the model. This finding further implies their potential key roles in determining the infectiousness of viruses to human hosts.

3.3.3. Comparing human codon fitness of virus genome sequences harvested from different sample sources

I employed the HVCF score, derived from the virus genomes' RSCU and additional features, as an index to assess the VCF in human hosts. This index was primarily generated using the D_{RTC}-trained Recall-optimised RF model. The HVCF score serves as a predictive measure to analyse the fitness of viral genomes in adapting to the human host, providing insights into their potential infectiousness.

The HVCF scores, derived from virus genome sequence data sourced from both human and non-human sources, exhibited no significant distinctions in the predicted labels between viruses isolated from human and non-human hosts (Figure 32). Interestingly, the model often categorised non-human-sourced viral genomes as potential human codon fitted. Nonetheless, it was noted that HVCF scores for non-human-sourced viruses, including MERS-CoV, Zaire Ebolavirus, Zika virus, Influenza A virus, and Henipavirus, were generally lower compared to those sourced from humans (Figure 32). Different patterns were also identified across various taxonomies of source species, as shown in Figure 33.A. Additionally, the distribution of human-sourced samples within the model training dataset is detailed in Figure 33.B. Notably, virus orders such as Orthomyxoviridae and Filoviridae have a higher proportion of samples labelled as originating from human hosts within the training dataset. This enriched representation correlates with the model's excellent performance in accurately distinguishing between human and non-human labels within these groups. This observation suggests that the less satisfactory performance in predicting labels for viruses from other orders may derive from an imbalance in the training dataset between human-labelled and non-human-labelled samples.

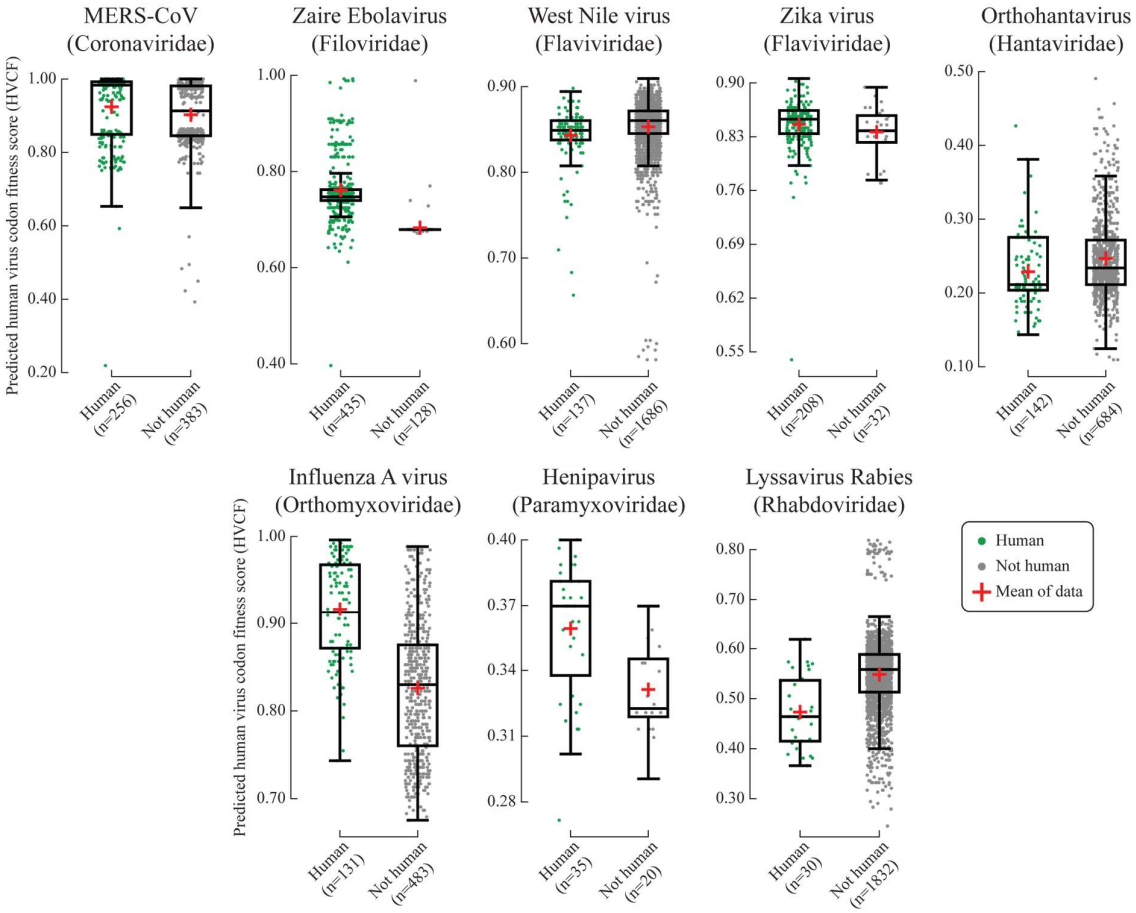


Figure 32. Using D_{RTC}-trained recall-optimised RF model to predict HVCF scores of virus genome sequence data from human or non-human sources. Eight viruses from different taxonomic orders are examined including MERS-CoV (Coronaviridae), Zaire Ebolavirus (Filoviridae), West Nile virus (Flaviviridae), Zika virus (Flaviviridae), Orthohantavirus (Hantaviridae), Influenza A virus (Orthomyxoviridae), Henipavirus (Paramyxoviridae), and Lyssavirus Rabies (Rhabdoviridae).

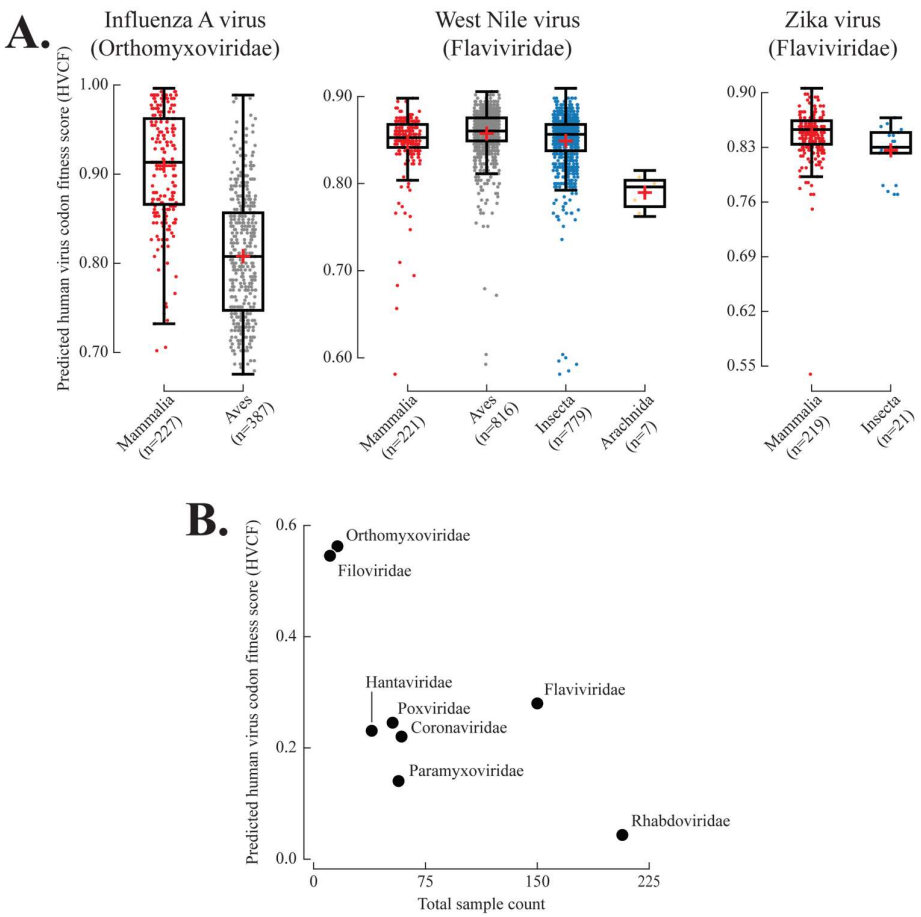


Figure 33. Using D_{RTC}-trained recall-optimised RF model to predict HVCF scores of virus genome sequence data from environmental source. (A) HVCF of virus genomes from species having different taxonomic orders; (B) Sample counts and percentages of different virus taxonomic orders in the train data.

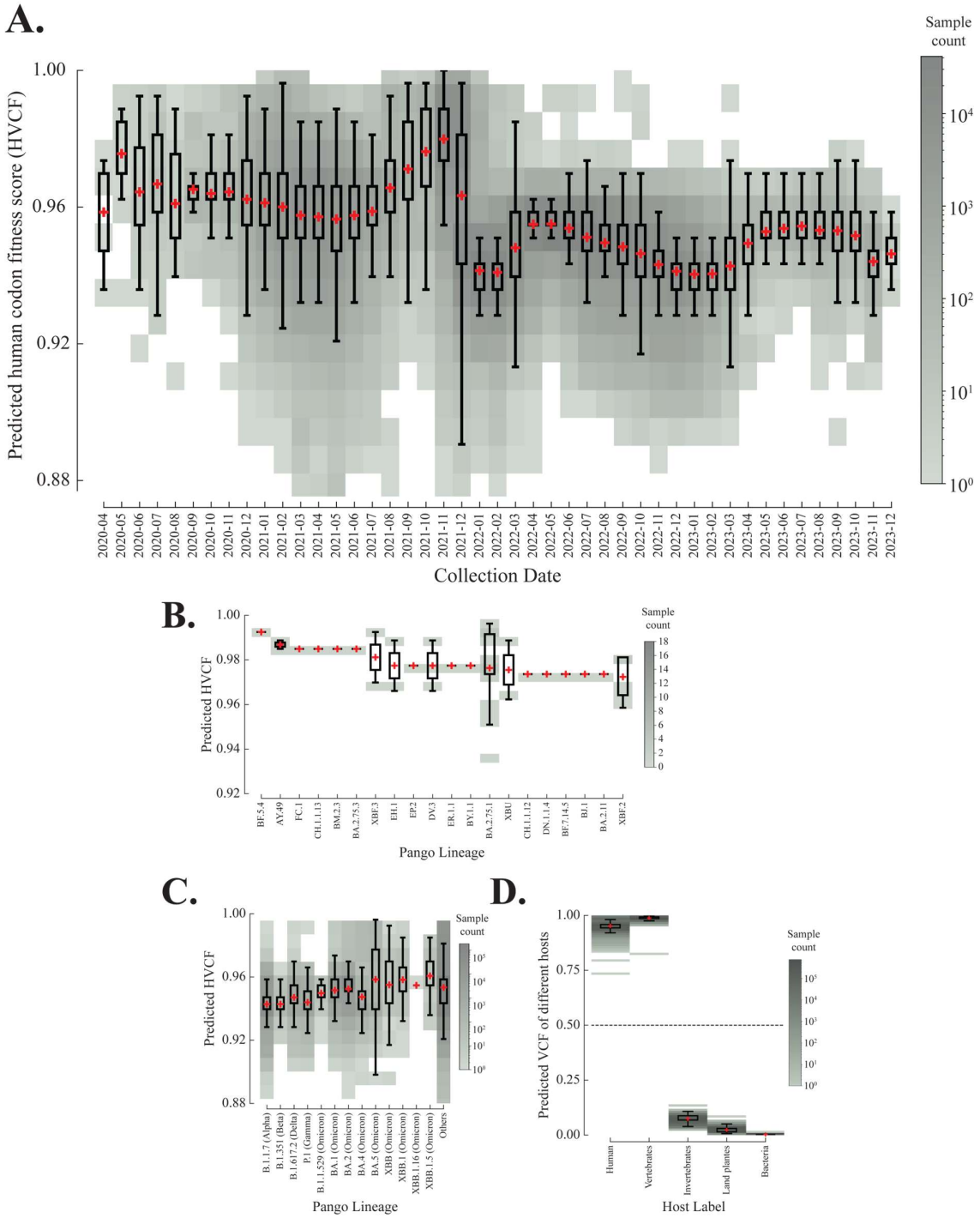


Figure 34. Using trained model to monitor HVCP shifting of SARS-CoV-2 genomes. (A) Predicted HVCF scores of SARS-CoV-2 in USA across timeline from April 2020 to December 2023. (B) Pango Lineage of SARS-CoV-2 have highest predicted HVCF scores. (C) Predicted HVCF scores of major Pango Lineage. (D) Predicted HVCF scores of USA SARS-CoV-2 genome sequence data in other host labels, including vertebrates, invertebrates, land plants, and bacteria.

I also calculated and ranked the HVCF scores of SARS-CoV-2 genomes sequenced in the USA throughout the pandemic timeline (Figure 34, Appendix 10). The initial or reference genome of SARS-CoV-2 (NC_045512[374]) registers a HVCF score of 0.992, indicating a very high probability of human infectivity. Subsequent SARS-CoV-2 genomes sequenced after the pandemic outbreak generally exhibit lower HVCF scores. The lowest observed HVCF score is 0.740 (January 2021), and the highest is 1.000 (November 2021), with the scores typically fluctuating around an average of 0.953 (Figure 34.A). The overall trend does not show a significant shift towards reduced human infectivity in the evolution of SARS-CoV-2 during the pandemic. Notably, an increase in the mean HVCF score is observed between August and November 2021. Subsequently, there is a gradual decline in the mean HVCF score, with a marked decrease in December 2021. Post-December 2021, the mean HVCF score continues to fluctuate.

To identify potentially threatening strains of viruses, I also ranked the predicted mean HVCF scores of different pango lineages, showcasing the top 20 in Figure 34.B. The lineage BF.5 recorded the highest mean HVCF at 0.992, closely followed by AY.49 at 0.987. Additionally, the codon fitness of the virus with respect to other hosts were also investigated (Figure 34.D). SARS-CoV-2 consistently exhibits significantly high VCF to humans and vertebrates, while the codon fitness in other hosts remain low (< 0.138), suggesting potential risks to other vertebrate species but not to non-vertebrate hosts. Thus, the VCF of SARS-CoV-2 has remained largely within the human and vertebrate range throughout the pandemic, indicating no substantial host-shifting trends.

3.3.4. Prediction of SARS-CoV-2 codon fitness shifting path starting from other Betacoronaviruses through HVCF gradients

To unveil the unknown genetic links between SARS-CoV-2 and human-non-infectious Betacoronavirus, I used HVCF readouts as gradient scores to simulate paths of codon-mutation-driven (including codon substitutions, codon addition, codon deletion) virus evolution for False-to-True VCF jump (i.e., human-non-infectious to human-infectious jump). In the simulation, each of the human-non-infectious Betacoronavirus is taken to perform a step-by-step mutation to screen efficient codon-mutations evolving till SARS-CoV-2's HVCF score

(see more details in the method section). At each step, it is required to generate a mutated strain such that it has a possibly highest HVCF score and the best correlation to SARS-CoV-2's RSCU matrix (Forward Mutation Path). Similarly on the other hand, SARS-CoV-2 is also taken to screen efficient mutations to evolve into a target Betacoronavirus, then the mutation path is reversed after to generate a False-to-True result (Backward Mutation Path). These paths are shown in Figure 35.A.

From the construction of these putative evolutionary paths, I can see that the Tylonycteris bat coronavirus HKU4 (NC_009019) has the highest level of efficiency to evolve into SARS-CoV-2 equivalent VCF according to the HVCF score changes per mutation compared to other Betacoronavirus reference genomes (Figure 35.B). More importantly, the compositions of both the forward and backward mutation path are also studied, I found that UUA(Leu)-to-CUC(Leu) and GAU(Asp)-to-GAC(Asp) mutations are significantly abundant in forward mutation path, while UAU(Tyr)-to-UAC(Tyr) and GCU(Ala)-to-GCC(Ala) are significantly abundant in reverse mutation path (Figure 35.C). Besides, GGU(Gly)-to-GGA(Gly) is abundant in both paths. Based on the simulation results, it is predicted that significant codon usage changing in amino acids of Leu, Asp, Tyr, Ala, Gly may be spotted in the intermediate strain of the viruses, if SARS-CoV-2 evolved from intermediate viruses related to Tylonycteris bat coronavirus HKU4. Interestingly, the third nucleotide mutations are spotted in all those mutations, especially U-to-C mutation, where most of those mutations are synonymous mutations. Similar results are observed in codon usage changes which multiple U-ended codons are significantly decreased in abundance including GCU(Ala), UAU(Tyr), GGU(Gly), CGU(Arg), CCU(Pro) in both paths, while multiple C-ended codons are significantly increased in abundance like CUC(Leu), GCC(Ala), UAC(Tyr) besides GGA(Gly). This finding provides clues in virus evolution for searching human infectious intermediate viruses from environmental sampling.

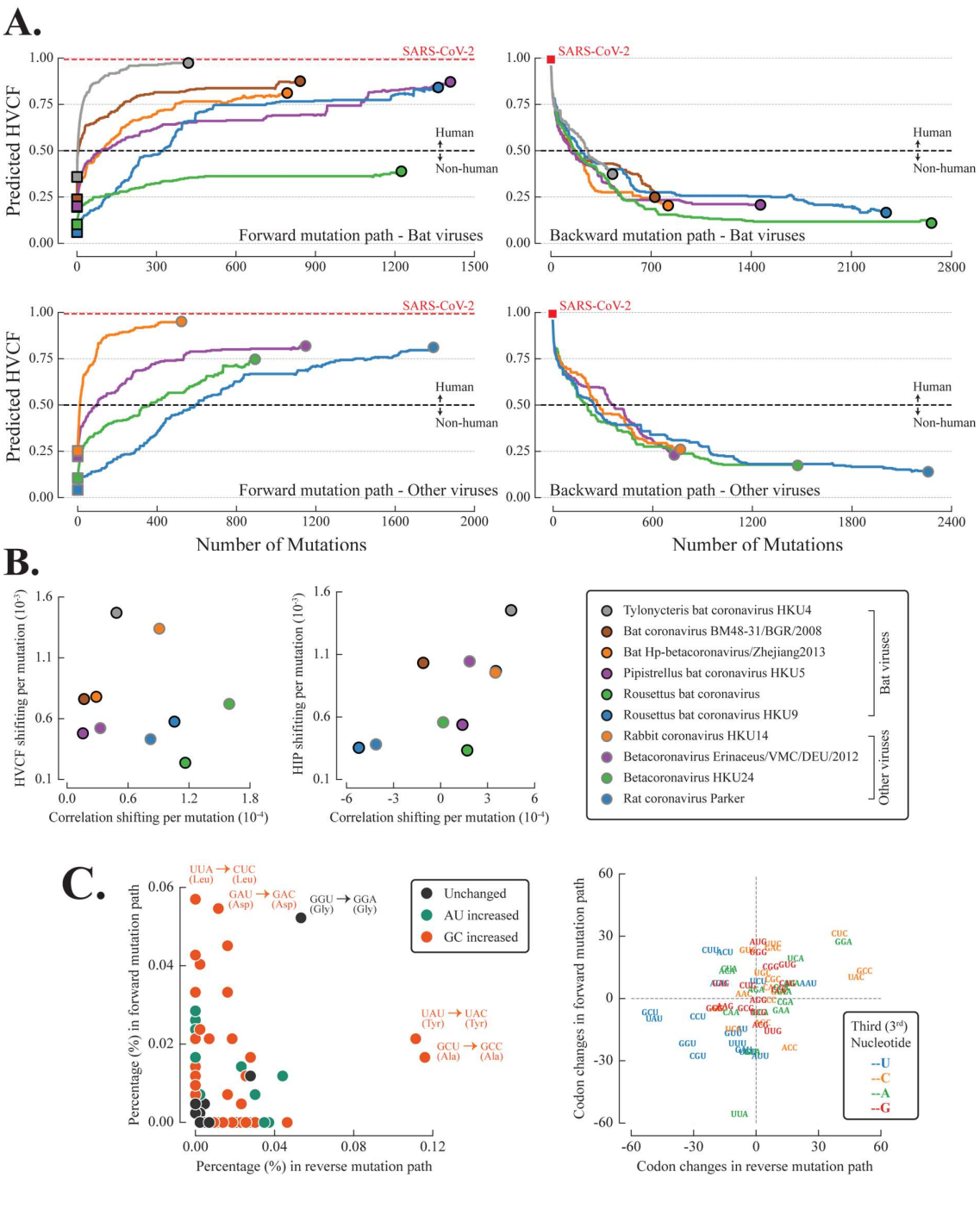


Figure 35. Prediction and analysis of SARS-CoV-2 codon fitness shifting path using codon mutations from other Betacoronavirus. (A) Predicted SARS-CoV-2 codon fitness evolution path using codon mutations from other Betacoronavirus. (B) Evolution efficiencies in both HVCF changes and correlation coefficient changes of different Betacoronavirus are shown. (C) Analysis of codon mutations in predicted codon fitness evolution path from Tylonycteris bat coronavirus HKU4 to SARS-CoV-2 with both abundant codon mutations and codon abundancy changes (another figure format in Appendix 11).

3.4. Discussion

In this chapter, I examine the codon usage biases in viruses with different host ranges, and try to train machine learning models to predict virus host ranges. It verifies that machine learning can detect distinguishable boundaries of codon usage biases from virus genomes having different host ranges, and codon usage biases have predictive power to virus host ranges and the underlying probabilities of infectiousness.

This modelling methodology has advantage in its generalisability because it purely relies on the codon usage biases of viral genomes and other general genomic characteristics regardless the diversity in virus-host interactions of different viruses such as expression regulation, protein interactions, cellular immunity, tRNA pool regulation et al. It overcomes the dilemma that the real micro-environment of virus-host interactions are complicated, and the significant diversity in different individuals from the same host type. Incomplete virus genome sequence data can also generate codon usage biases for sub-optimal prediction, expanding the range of application scenarios. This new way of predicting virus host codon fitness provides new insight into how I understand virus host ranges complementing the current major research focus on host entry of virus (i.e. Spike-membrane protein interaction)[375-377]. Moreover, data mining of using codon usage biases to represent coding sequences is significantly more computationally efficient compared to other methods such as natural language processing (NLP)[378]. The sample quantity limitation and imbalance need improvement when using only virus reference genomes, especially the imbalances in virus sample amount of different host ranges (i.e. human virus vs not-human virus). This may be possible to overcome with sample synthetic algorithms or generative deep learning networks to simulate virus genomes. Additionally, the representation of virus genome through summing codons counts within all gene CDS may not be biased towards the gene CDS of longer length. This may be improved through other embedding algorithms or through derivatives like Transformer model.

The concept of human virus codon fitness score (HVCF score) sourced from the decision tree models provides capable potential of monitoring the dynamics of virus host codon fitness shifting, which could help assess the potential host codon fitness and host ranges of emerging viruses which may cause disease outbreaks or even pandemic. However, there is still no evidence supporting that this readout of VCF is correlating to virus lethality to host or virus

infection outcomes. The accuracy of predicting different types of viruses may be different because the limitation and imbalance of training data. This results with HVCF scores of human-sourced and not-human-sourced viruses in different viruses suggest that this modelling method has potential to development accurate prediction tools to monitor virus host codon fitness shifting accordingly. In the SARS-CoV-2 pandemic analysis, the HVCF score remains at a similar level suggesting that the current attenuation in COVID-19 mortality rate is less likely leading to gradual vanish, but it remains a long persisting disease[379, 380]. Besides, this method could also identify the potential threating viruses with routine virus genome sequencing of environmental sampling (e.g., bats, mice, rats et al). The deficiency of this method is the difficulty in acquiring new samples to build models in species-specific scope (i.e. cats, dogs et al) because it is unethical and dangerous if infecting various species with various specific viruses.

More importantly, an innovative method has been proposed by this study to simulate possible evolutionary path between two viruses (original virus and target virus). Comparisons among different evolutionary paths could help identify the relations of VCF between the two viruses. SARS-CoV-2 and other Betacoronavirus are taken as example by this study, where Tylonycteris bat coronavirus HKU4 stood out closely relating to SARS-CoV-2 in terms of VCF. Further studies on the predicted evolutionary paths conclude that codon-related evolutionary signatures have significant abundancy in synonymous mutations, especially with U-to-C mutations in wobble position, of the Leu, Asp, Tyr, Ala, Gly in the intermediate viruses. Moreover, this finding of abundant synonymous mutations in the evolution simulation also demonstrates the importance of synonymous mutations in virus evolution. This method provides guidelines for searching evolutional relations between viruses and guidance for virus traceability research. The predicted probabilities generated from the RF models are discontinuous due to the nature of the algorithm leading to inefficiency and inaccuracy in predicting impacts of different codon-related mutations, which may be overcome with deep learning algorithms in the future work.

CHAPTER 4

Editing on Host RNA affects Virus Entry

by Shuquan (Steve) Su

Keywords:

RNA editing, Adenosine Deaminase, Host receptors, FRET, PPI

Chapter 4. Editing on Host RNA affects Virus Entry

4.1. Introduction

4.1.1. ADAR editing on host transcripts in virus infections and potential impacts on functionalities of host proteins

In Chapter 1, I identified two major overlooked aspects within the current scope of RNA editing research. The first knowledge gap I highlighted concerns the relations between RNA editing, conducted by ADAR and ADAT, and virus codon usage biases. To address this, I applied multiple statistical analysis to find distinct characterises in codon usage biases of human viruses and others in Chapters 2. Additionally, I developed a random forest-based machine learning tool in Chapters 3 to assess the relationships between codon usage biases and the host ranges of viruses. In Chapter 4, I will turn my attention to the second knowledge gap, focusing on RNA editing events targeting human host receptors and their subsequent effects on viral entry, a critical factor in determining the outcome of infections.

During process of virus infection, multiple host anti-viral immune mechanisms were triggered to counteract virus infections, where interferons (IFNs) signalling pathway is one of most important host anti-viral immune mechanism, where host cells will express and release IFNs as signalling molecules with gene regulatory functions[322, 355]. IFNs activate cellular immune responses through IFN pathway to up-regulate immune-related gene expressions. According to previous research, ADAR1L is classified as one of this IFN inducible gene, which abundantly expressed (up-regulation) during virus infections. Because of the randomness of A-to-I editing conducted by over-expressed ADAR1L, mutations caused by A-to-I editing were largely found in viral RNA molecules resulting in hyper-mutated viral RNA[183, 199-201, 203, 259]. In addition, the exogenous RNA molecules often contain ADAR1L-favored double-stranded-like structures[2, 120, 121]. The outcomes of those RNA editing events became the

biggest argument in this field. A part of researchers believed ADAR1L editing is anti-viral because A-to-I mutations on viral RNA cause malfunction of viral genes and/or decrease in translational efficiency. Another part of researchers believed ADAR1L editing is pro-viral because abundance of A-to-I editing derived mutations eventually lead to elevated virus evolution rate. This is the biggest debate in the research field of RNA editing in virus infections. When it comes to this debate, after reviewing considerable amount of research related to ADAR family and RNA editing, I tend to believe that ADAR1L editing fundamentally functions as anti-viral mechanisms based on the undeniable fact regarding IFN-inducibility of ADAR1L. IFN signalling pathway is one of the most vital anti-virus mechanisms of human host. In another term, I suggest that ADAR1L is originally a part of host cells immune defencing mechanism for host survival purposes regardless its side-affect as Darwinian natural selection pressure in virus perspective.

Based on this perspective, I assume there are other anti-viral related outcomes caused by ADAR1L editing that haven't been discovered yet. Compared to other genes in ADAR family, ADAR1L editing is relatively lacking specificity with higher randomness, which matches the function of hyper-mutating different types of viral RNA related to complexity of virus classifications and properties. Theoretically, all the intracellular RNA molecules are also potential editing targets with this intracellular semi-random hyper-mutating dynamic, which is not solely targeting viral RNA. If the proteins of the host, which are critical to virus infections, are mutated by RNA editing, this theoretically causes negative impacts on virus infections, in another term, anti-virus effect. Therefore, I hypothesise that RNA editing conducted by IFN-induced ADAR1L hyper-mutating dynamics could potentially affect functionality of host proteins that are critical to virus infections, and eventually lead to anti-viral effects. In this study, I focus on the ADAR1L editing on host receptor proteins, and their subsequent effects on protein-protein interactions (PPI) between host receptor and virus Spike proteins. The host entry of SARS-CoV-2 has been largely studied during the pandemic, and it is well-established model compared to other viruses. The host entry of SARS-CoV-2 involved protein-protein interactions (PPI) among SARS-CoV-2 Spike protein and mainly two human genes including Angiotensin-converting enzyme 2 (ACE2)[354, 381] and Transmembrane Serine Protease 2 (TMPRSS2)[382]. Especially ACE2-Spike binding is believed as the most important part of SARS-CoV-2 entry. Therefore, by studying RNA editing on ACE2 and TMPRSS2 may

provide clues of RNA editing on host genes in altering virus infection related functions (Figure 36).

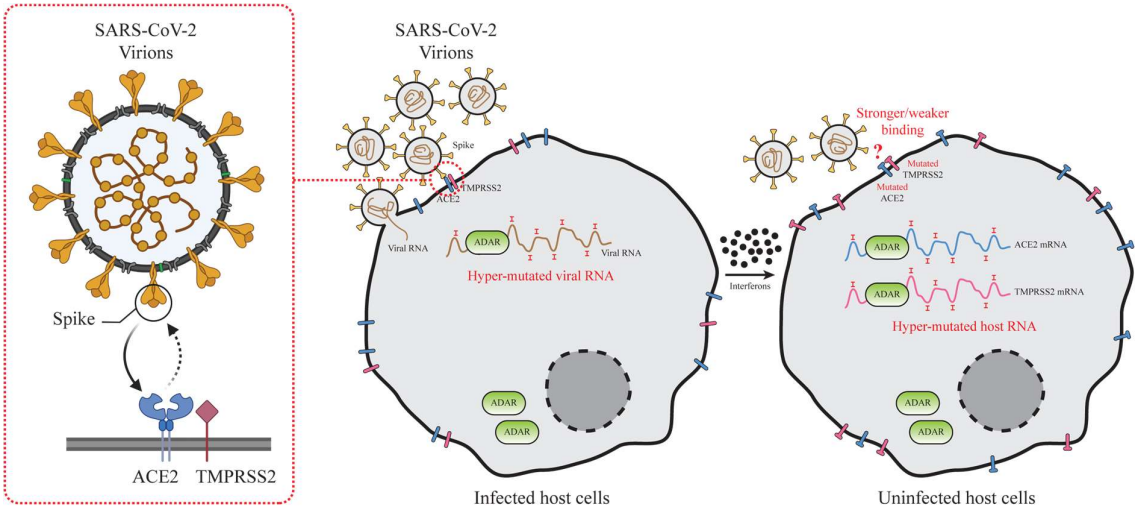


Figure 36. Schematic drawing of the hypothesis that ADAR editing derived mutations on ACE2 and TMPRSS2 transcripts would generate mutated ACE2 and TMPRSS2 proteins, eventually affects their PPI against Spike protein and SARS-CoV-2 entry.

4.1.2. Potential use of FRET-based assay to detect protein-protein interactions between SARS-CoV-2 Spike protein and human receptors ACE2 and TMPRSS2

To find evidence of RNA editing causing PPI changes, there are two major tasks: (1 finding RNA editing events and (2 detecting PPI changes caused by RNA editing derived mutations. As I mentioned in articles reviews in the Chapter 2, finding RNA editing events caused by ADAR1L hyper-mutations is relatively easy, and there are many well-established workflows for allocating RNA editing events with RNA sequencing data by comparing between control sample and ADAR1L overexpressed sample (either induced or artificially overexpressed). In this study, I use RNA sequencing techniques to find A-to-I editing events in RNA purified from ADAR1L over-expressed cell line and use recently published bioinformatic tools (i.e.

REDItools[383, 384]) to identify RNA editing events on human receptor proteins ACE2 and TMPRSS2.

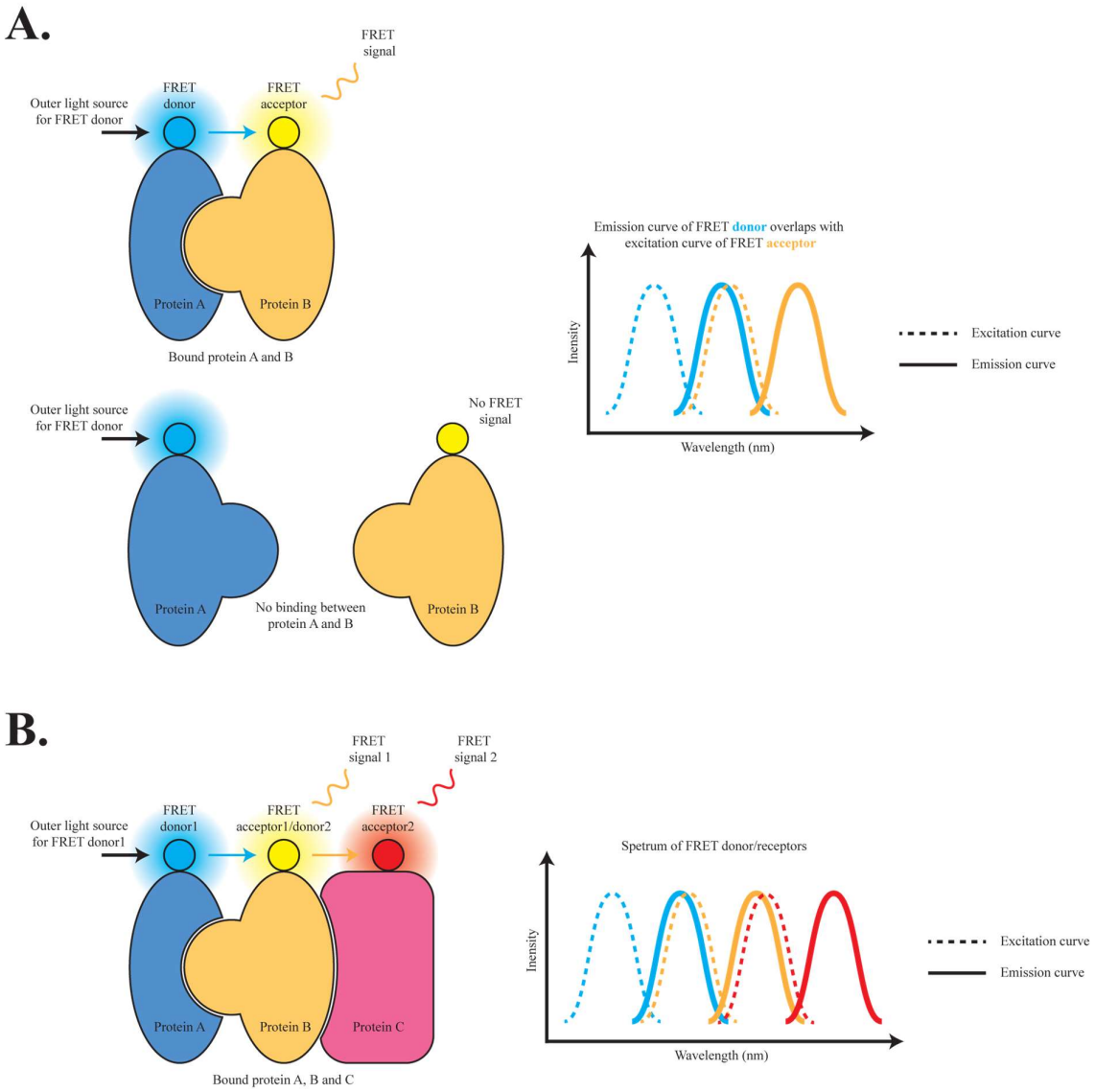


Figure 37. Schematic drawing of the principle regarding FRET-based PPI detection assay. (A) Two-way FRET assay is contained two gene-reporter fusion proteins, which the emission curve of FRET donor heavily overlaps excitation curve of FRET acceptor. When the protein A and protein B is bound, the FRET signal could be detected. The FRET signal is not detectable when protein A cannot bind protein B. (B) Three-way FRET assay is contained three gene-reporter fusion proteins, which the emission curve of FRET donor heavily overlaps excitation curve of FRET acceptor. When three proteins are bound together forming a complex, various FRET signals could be detected.

Unlike allocating RNA editing events, detecting changes in PPI between viral Spike and host receptors caused by those A-to-I mutations is the major difficulty in studying subsequential impacts of RNA editing. I could use traditional method to study PPI such as protein structure analysis with X-ray crystallography[385] or even cryogenic electron microscopy[386, 387] with purified protein complexes. However, those methods are very time-consuming when dealing with large number of different proteins complexes, which are protein complexes with various RNA-editing-derived mutations in this case. Therefore, I planned to use intra-cellularly expressed proteins with fluorescence tag fusion, and FRET reaction among fluorescence tags to study protein-protein distance between two proteins.

Fluorescence resonance energy transfer, or FRET, is a phenomenon that two or multiple fluorescence molecules, that having special spectrum overlapping, transfer energy between them, which is one of the most important and widely used methods to detect protein-protein interaction. The FRET donor, which its fluorescence emission spectrum has certain level of overlapping with the FRET receptor's excitation spectrum, transfers partial energy to activate the FRET receptor when the donor is activated by outer light source and emitting receptor-favoured fluorescence[388, 389]. Thus, when the FRET donor and FRET acceptor are close to each other, the FRET receptor will be excited by donor's emission, and start to emit detectable fluorescence, which is in another word, FRET signal (Figure 37.A). FRET assay has been used as effective tool in staying molecular biology in various research, which has great potential in studying protein-protein interactions[388]. Especially when separately linking proteins of interest with FRET acceptor and donor through protein fusion, the FRET signal could be detected if the proteins of interest conduct binding interactions bringing the FRET donor and receptor close enough for FRET. Moreover, with plasmid-based fusion protein assay, the coding sequence or equivalent amino acid sequence is easier to change with mutagenetic molecular cloning method (i.e. site-direct mutagenesis), compared to protein-based of workflow (Figure 38). In another term, the throughput of using FRET to study PPI is significantly higher than protein-based methods allowing me to screen larger quantity of mutations in short time.

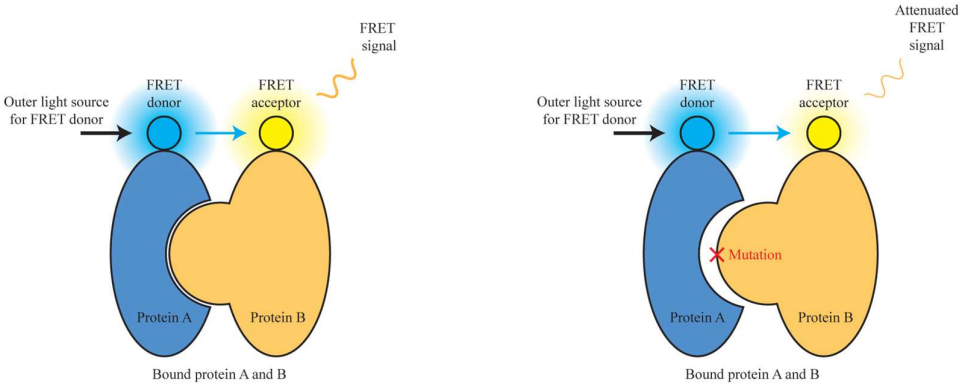


Figure 38. Schematic drawing of how mutations on proteins leading to alternations in FRET signals. The mutations on proteins causing changes in their PPI and the distance between fused fluorescence tags (increase, decrease, no change). The alternations in protein-protein distance will correspondingly change the detected FRET signals.

Based on the biology of FRET, it is not surprising that we could apply additional FRET molecules into the FRET system and detect the FRET signals from the multimer-formed complex. When we have three fluorescence reporters that have continuous resonance spectra, we could detect FRET signals from different FRET receptors once the FRET donors are activated, which it is named three-way FRET (Figure 37.B). In my colleague Dr Ni's research, he successfully used FRET assay consist of enhanced cyan fluorescent protein (eCFP), enhanced yellow fluorescent protein (eYFP), and monomeric red fluorescent protein 1 (mRFP1) to detect trimers' formation dynamics of tumor necrosis factor receptor (TNFR)[390]. Therefore, in this chapter, I propose to use FRET-based protein-fusion assay to detect PPI changes among SARS-CoV-2 Spike, human ACE2 and TMPRSS2 caused by RNA editing derived mutations, to investigate evidence of ADAR editing impacting SARS-CoV-2 infections. Here is some general information of Spike, ACE2 and TMPRSS2 from NCBI database (Table 3). By fusing Spike, ACE2 and TMPRSS2 with different FRET-related fluorescence reporters, the PPI among them could be studied based on the FRET signals detected from the fluorescence emission profiles of fused fluorescence reporters. Likewise, the changes in PPI caused by any types of mutations could also be quantitatively studied according to changes in FRET signals. Thus, this method is ideal to study impacts of RNA editing derived

mutations on protein-protein interactions among Spike, ACE2 and TMPRSS2, as the representative model to host entry of virus infections.

Table 3. Summarised general information of SARS-CoV-2 Spike gene, human ACE2 and human TMPRSS2 gene from NCBI Gene database.

Gene	Species	NCBI ID	Location	Exon count	Chromosome	NCBI Annotation location (GRCh38.p14)	Verified transcripts count	Reference transcript accession ID
S	SARS-CoV-2	43740568	-	-	-	NC_045512.2 (21563..25384)	-	-
ACE2	Homo Sapiens	59272	Xp22.2	22	X	NC_000023.11 (15518197..15607211, complement)	6	NM_021804.3
TMPRSS2	Homo Sapiens	7113	21q22.3	15	21	NC_000021.9 (41464305..41508158, complement)	3	NM_005656.4

4.2. Methods

4.2.1. Molecular biology methods

4.2.1.1. Plasmid extractions from transformed E. coli

For preparations of larger quantity of plasmid for cell transfection purpose, PureLink HiPure Plasmid Midiprep Kit (Thermofisher) was used to extract transfection-grade plasmids solution according to manufacturers' protocol. Here is a brief Midiprep protocol. The 10 mL of Equilibration Buffer EQ1 was first added into the HiPure Midi Column, which the solution was drained by gravity. 100 mL transformed *E. coli* cultures were pelleted by centrifuging at 4,000 ×g for 10 mins, which the bacterial pellet was resuspended with 4 mL of Resuspension Buffer R3 after LB supernatant was removed. 4 mL of Lysis Buffer L7 was added to the resuspended bacteria, and the bacteria was lysed for 5 mins at room temperature. The mixture was then neutralised and precipitated by adding 4 mL of Precipitation Buffer N3, which the insoluble was precipitated by centrifuging at 12,000 ×g for 10 mins. The supernatant was transferred into the HiPure Midi Column, and the solution was drained by gravity. 10 mL Wash Buffer W8 was added to the column and the solution was also drained by gravity, which was repeated twice. 5 mL Elution Buffer E4 was added to the column for eluting the plasmid DNA, and the eluted DNA was mixed with 3.5 mL of pure isopropanol. The DNA (in elute-isopropanol mixture) was pelleted by centrifuging at 12,000 ×g for 30 mins at 4 ℃, and the supernatant was discarded. The DNA pellet was then washed by adding 3 mL of 70% ethanol, which ethanol supernatant was discarded after centrifuging at 12,000 ×g for 5 mins at 4 ℃. The ethanol residue was carefully removed with pipetting before dissolving the DNA pellet in 100 µL of Nuclease-Free MilliQ water. The plasmid DNA solution was often diluted into 500 ng/µL later for subsequential experiment (i.e. lipofectamine transfection).

The sequences of extracted plasmids were verified with Sanger sequencing service provided by Macrogen (South Korea).

4.2.1.2. Long-term maintenance of transformed E. coli

The insert sequences of all the plasmids in different experiments were verified with Sanger sequencing in either unidirectional or bidirectional manners. For long-term storing, the verified plasmids were transformed into *E. coli*, and maintained as glycerol stocks, which were stored at -80 ℃ for long-term storage. The glycerol stocks were prepared by mixing 750 µL of bacterial LB broth containing transformed *E. coli* cultures with 750 µL of 50 % glycerol solution.

4.2.1.3. Protein extraction from mammalian cells

To extract protein from HEK293T cells (either transfected or non-transfected), the protein lysis buffer was used, which was made by mixing Radioimmunoprecipitation Assay lysis buffer (RIPA lysis buffer, Sigma-Aldrich) and Protease Inhibitor buffer (100×, Roche). The HEK293T cells were treated with 100 µL of protein lysis buffer for 5 mins before the lysed cells were harvested with scraper, and transferred into Eppendorf tubes, which were later incubated on ice for 10 mins. The lysed cells were precipitated by centrifuging at 14,000 rpm for 10 mins at 4 ℃, which the supernatant was transferred and stored as extracted protein solution for subsequent work (i.e. BCA assay and Western blotting). The concentration of extracted protein was quantified with Bicinchoninic acid (BCA) assay with Pierce BCA Protein Assay Kits (Thermofisher) following manufacturer's instruction.

4.2.1.4. Western blotting with extracted proteins

30 µg of extracted proteins were used for each sample. The extracted proteins were first separated by pulling down in the 10% (for p-H2AX) SDS-PAGE gel (Thermofisher) with PowerPac Basic Power Supply (Bio-Rad), which the separated proteins were later transferred onto 0.45 µM pore size (0.2 µM for p-H2AX) PVDF membranes (Merck). The PVDF blots were blocked with 5% skimmed milk (A2) prepared in TBS-T buffer before being incubated with primary antibodies at 4 ℃ overnight on a roller. Primary antibodies used in this study include mouse anti-3×FLAG (1:1000 dilution, Sigma-Aldrich) and mouse anti-β-actin (1:5000 dilution Cell Signaling). Blots containing separated proteins were then incubated with horseradish peroxidase-conjugated anti-mouse secondary antibodies (1:5000 dilution, Abcam) for 1 hour, followed by TBS-T washing for three times. The stained protein bands were

visualised with Pierce ECL Immunoblotting Substrate (Thermofisher) and subsequential chemiluminescence reaction under the Chemidoc mode in Amersham Imager 600 (Amersham). The protein size in kDa is estimated with online bioinformatic tool 'Quest Calculate'[391] (i.e. ADAR1L) for proteins with unknown sizes.

4.2.1.5. Total RNA extraction from mammalian cells

TRIzol-based method was used to extract RNA from cells transfected with different over-expression plasmids. Briefly, the cells from 6-well plates were treated with 1 mL TRIzol reagent each well and transferred into 1.5 mL Eppendorf tube after incubating 5 mins at room temperature. 0.2 mL chloroform was added into each tube and the samples were mixed well with vortex before centrifuging with 12,000 ×g at 4 ℃ for 15 mins. The colorless aqueous phase was transferred to a new Eppendorf tube and 0.5 mL of isopropanol was added to the sample, which was subsequentially incubated at 4 ℃ for 10 mins before centrifuging with 12,000 ×g at 4 ℃ for 10 mins. The supernatant was carefully discarded without interrupting the formed RNA pellet. The RNA pellet was resuspended with 75 % ethanol with vortex before centrifuging with 7,500 ×g at 4 ℃ for 5 mins. The supernatant was carefully discarded, and the RNA pellet was air-dried for 10 mins before resuspending with 50 µL of RNase-free water.

4.2.2. Cell biology methods

4.2.2.1. Mammalian cell culture

Human embryonic kidney cells HEK293T[392, 393], kindly gifted by colleague Ms. Tao Xie (UTS) and Ms. Pattarasiri Rangsrikitphoti (UTS) were all cultured in Dulbecco's Modified Eagle medium (DMEM; Gibco) supplemented with 5% ($DMEM_5$, or 10% of $DMEM_{10}$) heat-inactivated fetal bovine serum (FBS; Gibco), 100 U/mL penicillin and 100 mg/mL streptomycin (Gibco), 4.5 g/L D-Glucose and L-Glucose (Gibco), 1 mM Sodium Pyruvate (Gibco) – i.e. $DMEM_5$ or $DMEM_{10}$. Mammalian cells were cultured at 37 ℃ and 5 % CO_2 in a Tissue culture incubator (Binder). Cell cultures were grown in T25 (25 cm^2 surface area), T75 (75 cm^2 surface area) or T175 (175 cm^2 surface area) TC flasks (SPL). Cell cultures were

sub-cultured approximately every three days (once they reach ~90 %-100 % confluency). For each sub-culture, media was tipped off and discarded, and cells were rinsed with Dulbecco's Phosphate Buffered Saline (PBS; Gibco) equal amount to media then incubated in Trypsin-EDTA solution (0.25%, Gibco) at 37 ℃ for 10 mins, until they were visually dislodged from the plastic TC flask (viewed under phase contrast microscopy (Nikon Eclipse Ts2 Inverted Biological Microscope). Cells were helped to detached from the plastic by vigorous shaking of the flasks. Once dislodged, the Trypsin-EDTA was neutralised by adding $DMEM_5$ (or $DMEM_{10}$). Cells were transferred into 50 mL sterile tubes (Falcon) and centrifuged at 150 rpm for 5 mins at 4 ℃ then the supernatant was discarded, and the cell pellets were resuspended into fresh $DMEM_5$ (or $DMEM_{10}$). A proportion of the resuspended cells was placed into a new TC flask to continue the culture, or seeded into wells of a 6-well, 12-well, or 24-well TC dishes (Falcon), according to need – i.e. for a transfection experiment. All the HEK293T cells had been maintained between passage 6 (P6) and passage 35 (P35).

4.2.2.2. Lipofectamine transfection into mammalian cells

To transfect different plasmids for constitutively expressing either fluorescence reporter fusion proteins or p3×FLAG tagged proteins in HEK293T cells, Lipofectamine 3000 Transfection Kit (Thermofisher) was used according to manufacturer's instruction. The transfection mixture was optimised according to manufacturer's instruction by transfecting various plasmids into HEK293T cells and live-cell imaging with a time series (Appendix 14, Appendix 15). Before making ready-to-go lipofection buffer, lipofectamine solution was diluted by mixing 125 µL of Opti-MEM media (Gibco) and 7.5 µL of Lipofectamine 3000 solution for each well of 6-well plate, where 50 µL of Opti-MEM media and 3 µL of Lipofectamine 3000 solution for each well of 12-well plate). Besides, the plasmid DNA solution was diluted by mixing 125 µL of Opti-MEM media, 5 µL of plasmid DNA solution (500 ng/µL, in Nuclease-Free MilliQ water) and 5 µL of P3000 solution (2 µL per ug of DNA) for each well of 6-well plate, where 50 µL of Opti-MEM media, 2 µL of plasmid DNA solution and 2 µL of P3000 solution for each well of 12-well plate. The diluted lipofectamine solution and diluted DNA solution were mixed to make ready-to-go lipofection buffer, which was later added to the cell culture. The cell culture media was replaced by fresh media 6 hours after lipofection buffer was added, which was considered as the starting point for post-transfection time.

4.2.2.3. Cell imaging with transfected mammalian cells

Fluorescent protein produced within cell cultures, transfected by fluorescent protein CDS containing plasmids, were detected by EVOS FL cell imaging system (Life technologies). Images were taken using specific filters for each of the fluorescence proteins: CFP filter set (Ex 445/45 nm; Em 510/42 nm) for detecting eCFP fluorescence, YFP filter set (Ex 500/24 nm; Em 524/27 nm) for detecting eYFP fluorescence, RFP filter set (Ex 531/40 nm; Em 593/40 nm) for detecting mCherry or mRFP fluorescence, or via the transmission channel. The intensity of the light source (LED) was identical throughout all images, whereby the transmission setting was set to 45%. Similarly, intensity of the light sources for the fluorescent channels were also kept constant: eCFP 30%, eYFP 30%, and mCherry 30%. All images were saved as .tiff files and analysed with ImageJ software (version 1.54p).

Besides, transfected cells were also examined with Incucyte S3 Live-Cell Imaging System (Sartorius) for transfection optimisation purposes. The images taken from different time points were analysed with IncuCyte 2021C software (version 2023A, Sartorius), which the mean fluorescence intensity, or total integrated intensity (RCU × μm^2/image), of different samples were computed with build-in function.

4.2.2.4. Flow cytometry with transfected mammalian cells

Transfected HEK293T cells from each well were first broken up into single cells with Trypsin-EDTA solution (0.25%, Gibco) and 10 mins of 37 ℃ incubation. The Trypsin-EDTA was neutralised by adding $DMEM_5$ (or $DMEM_{10}$), and cells were transferred into 1.5 mL Eppendorf tubes, which were centrifuged at 300 ×g for 5 mins. Then the supernatant was discarded, and the cell pellets were resuspended in 4 % Paraformaldehyde/PBS (4 % PFA/PBS) solution, which were incubated at RT for 15 mins before being neutralised by adding EDTA-PBS solution. The cells were pelleted by centrifuging at 300 ×g for 5 mins before removing supernatant and resuspending with EDTA-PBS. Detached cells were broken up into single cells by gauze filtering through 100 μm mesh (Sefar), which were transferred to FACS tubes. Harvested cells were further broken into single cells with gentle vortex before analysing in flow cytometer.

The HEK293T cells were analysed on a 3-laser BD LSR Fortessa X20 flow cytometer (Becton-Dickinson) equipped with 405nm (50 mW), 488nm (100 mW), 635nm (40 mW) lasers. Briefly, approximately 10,000 cells were acquired using Diva software (version 8.0.1, Becton-Dickinson), and the data was saved as FCS 2.0 files. All the of channels in BD LSR Fortessa X20 are used when recording data (Appendix 17) without any pre-set compensation matrix. FACS files were subsequently analysed with FlowJo software (version 10.8.1) or Python scripts.

4.2.3. Bioinformatics analysis

4.2.3.1. Protein structure visualisation of pdb files

The protein structures were examined and visualised through ChimeraX software (Daily Build, version 1.8.dev202403260833) with Protein Data Bank files (pdb files) of proteins.

4.2.3.2. Protein structure prediction with amino acid sequences

The structures of fusion proteins were predicted through ColabFold[394] extension (AlphaFold in Google Colab) embedded in ChimeraX software (Daily Build, version 1.8.dev202403260833). The amino acid sequences of the fusion proteins were uploaded to the server, which all options were unselected (Default setting).

4.2.3.3. Cytometry Utilities Box Expansion

Cytometry Utilities Box Expansion, or CUBE, is a bioinformatics pipeline to efficiently compute FRET signals and FRET efficiency from data harvested by flow cytometry. It was originally developed by my colleague Dr Zhongran Ni to study conformational changes of tumor necrosis factor receptor (TNFR) trimers with FRET-based TNFR-reporter fusion protein assays[390]. Thus, CUBE is ideal to compute pure FRET signals and FRET efficiency in the FRET-based PPI detection assay. CUBE includes 5 steps with different algorithms. The first step is the single-cell auto-extraction, which automatically extracts single-cells events from the event pool through a DBSCAN-based clustering algorithm (DBSCAN: Density-based spatial

clustering of applications with noise). The second step is background noise removal through predicting negative values with positive values by collaborative-filtering-based algorithm. The third step is the cell auto-fluorescence removal, which uses a collaborative-filtering-based algorithm to predict cell auto-fluorescence. The fourth step is spectral unmixing to purify fluorescence signals of the reporters, which removes spillovers across channels through RANSAC linear regression algorithm (RANSAC: Random sample consensus). The final step is to compute pure FRET signals and corresponding FRET efficiency based on the purified fluorescence data matrix. CUBE requires specifications of primary flow cytometry channels to detect involved reporters, which is listed in flow cytometry configuration in Appendix 17.

4.2.3.4. Next-Generation RNA sequencing

The extracted and purified RNA samples are delivered to Australian Genome Research Facility (AGRF, Australia) for Next-Generation RNA sequencing services. Before sequencing run, the RNA samples received were first evaluated through AGRF quality control protocol, which the sample are in excellent condition for subsequential sequencing run (Appendix 37). The raw data was generated through the Illumina NovaSeq X Plus platform with RNAseq sequence production of a 150 bp paired-end run. All the sequence reads generated from different samples were analysed according to AGRF quality control measures, which the per base sequence quality are excellent with >94% bases above Q30 (data not shown). Additionally, the sequence reads were cleaned by screening for the presence of any Illumina adapter/overrepresented sequences and cross-species contamination before I received the raw data of sequence reads in FASTA files.

4.2.3.5. RNA sequencing data Alignment to reference genome

The raw data was aligned to human genome with the Spliced Transcripts Alignment to a Reference software (STAR, version 2.7.10b)[395], which the reference genome used is Homo sapiens Reference genome GRCh38.p14 (RefSeq version: GCF_000001405.40). The genome directory was first generated with built-in module '--runMode genomeGenerate', which the FASTA file of genome sequence and the GTF file of Annotation features were supplied after downloaded from NCBI database. The sequencing reads from both FASTQ files (due to paired-

end sequencing) was then aligned to reference genome with genome directory and other input arguments including '--outSAMtype BAM SortedByCoordinate', and '--quantMode GeneCounts'. Two major data files for subsequential analysis were generated including the BAM file containing aligned sequences was generated and the table file summarising read counts of all genes according to the annotation files. The BAM files generated from STAR alignment were later indexed with Samtools software (version 1.9). Besides, the read counts of different genes were normalised by calculation of Fragments per Kilobase of transcript per Million (FKPM). Shortly, the read counts were first divided by corresponding scaling factor (sequence depth / 1,000,000), before dividing by gene length (in kb, acquired from NCBI Gene database). Several housekeeping genes were used for evaluating expression levels across samples, including GAPDH, ACTB, SDHA and PPIA.

4.2.3.6. Identifying RNA editing events from aligned BAM files

The indexed BAM file containing aligned sequences, generated with STAR and indexed with Samtools, served as the input file of REDItool2 (version 2.0) to identify the sequence variations within specific regions of the reference genomes, which are obviously the regions where genes ACE2, TMPRSS2, ADAR1L locate[383, 384]. The gene location indexes of reference genome are 'NC_000023.11:15518197-15607211' for ACE2, 'NC_000021.9:41464305-41508158' for TMPRSS2, and 'NC_000001.11:154582057-154627997' for ADAR1L, and they are set as input argument '-g' for extracting sequence reads of ACE2, TMPRSS2, and ADAR1L. At the end, a table containing all the sequence variations was generated.

In fact, sequence variations identified from REDItool2 contain all different nucleotide variations, not only variations related to A-to-I editing. A-to-I editing events are often detected as A-to-G variations due to I-C pairing. However, RNA sequencing technique is generating sequence reads of cDNA (complementary DNA) derived from RNA molecules. Therefore, the A-to-I editing events here are detected as T-to-C variations in REDItool2 results if the genes are on the complement strand of human genome (i.e. ACE2, TMPRSS2). And the I/A Ratio for editing efficiency analysis are equivalent to C/T Ratio in REDItool2 results. But the A-to-I editing events are still detected as A-to-G variations if genes are on the reference strand, thus the I/A Ratio are equivalent to G/A Ratio. To avoid zero-division problems in subsequential

studies, the counts of aligned nucleotides with both T and C (or A and G) are added by 1 before calculating C/T Ratio (or G/A Ratio)

Because the sequence variations are identified from the gene sequences from the genome, thus some of them are located on the introns. Because the transcripts of genes are originated from the overexpression plasmids, the transcripts expressed do not contain any introns. Thus, the sequence variations, that are located in the exons and identical to plasmid sequences, are extracted for subsequential analysis.

4.3. Results

4.3.1. Constructions of different plasmids to establish intracellular FRET assay

For the construction of an intracellular FRET assay designed to detect protein-protein interactions, several overexpression plasmids were contructed. These plasmids will encode fusion proteins, each tagged with fluorescence reporters, for the genes of interest: Spike, ACE2, and TMPRSS2.

To construct an intracellular FRET-based system for examining protein-protein interactions (PPI), I initially prepared empty overexpression plasmids (pcDNA3 vector) featuring coding sequences for fluorescence reporters suitable for either 5' or 3' fusion. For this study, eCFP, eYFP, and mCherry were selected as the primary fluorescence reporters, with mRFP serving as an alternative backup to mCherry. The efficiency of the eCFP-eYFP-mRFP three-way FRET assay was previously validated in the doctoral thesis of my colleague, Dr. Zhongran Ni[390]. mCherry, chosen for its nearly identical spectral profile to mRFP and more robust molecular stability[396], has also been successfully utilized as a FRET donor in other studies[397], proving its suitability for this application.

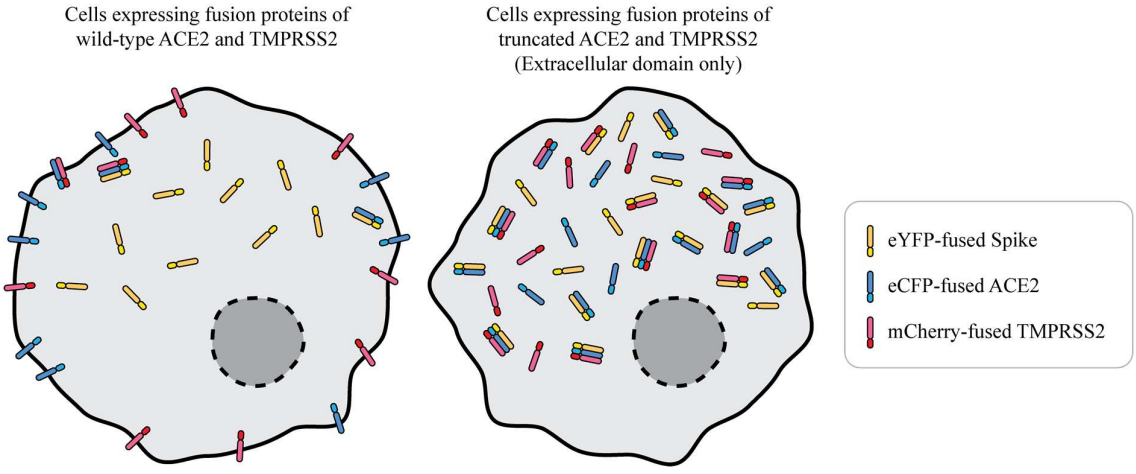


Figure 39. Demonstrating intracellular expression of fluorescence-tag fusion proteins containing either wild-type or truncated versions of SARS-CoV-2 Spike protein, human ACE2 and TMPRSS2. The Spike protein is fused with eYFP, whereas the ACE2 and TMPRSS2 are fused with eCFP and mCherry respectively.

To investigate the protein-protein interactions (PPI) underlying SARS-CoV-2 cell entry mechanisms and to assess the impact of RNA-editing-derived mutations on these interactions, the genes encoding ACE2, TMPRSS2, and Spike were integrated into plasmids containing fluorescence reporter coding sequences (CDS) based on the pcDNA3 vector. This setup enabled the creation of plasmids expressing fusion proteins of these genes with fluorescent reporters. Unlike the Spike protein, ACE2 and TMPRSS2 are membrane proteins that, once expressed, are typically translocated to the cell membrane. However, to facilitate their interaction with the cytoplasmic Spike protein and enhance the potential for detecting PPI changes, the transmembrane and extracellular domains of ACE2 and TMPRSS2 were deleted through molecular cloning. This modification resulted in truncated versions that remain cytoplasmic, thus increasing their likelihood of interacting with Spike and producing detectable FRET signals (Figure 39). Specifically, the truncated version of ACE2, designated as ACE2tr, comprises amino acids 0~740 of the full-length protein, which totals 805 amino acids. Similarly, the truncated version of TMPRSS2, referred to as TMPRSS2tr, includes amino acids 106~492 of the wild-type protein, which totals 492 amino acids. These truncated versions were cloned into both 5'- and 3'-oriented reporter plasmids using appropriate molecular cloning strategies. Unfortunately, the sub-cloning of pcDNA3-TMPRSS2tr-mCherry was unsuccessful, as no correct colonies were obtained. However, the successful constructs included pcDNA3-eYFP-Spike, pcDNA3-Spike-eYFP, pcDNA3-eCFP-ACE2tr, pcDNA3-ACE2tr-eCFP, and pcDNA3-mCherry-TMPRSS2tr, as shown in Figure 40. Detailed cloning methodologies and additional procedural information are comprehensively outlined in Appendix 18, which documents all cloning steps and methods. A summary of all constructed plasmids utilised in this chapter is provided in Appendix 34.

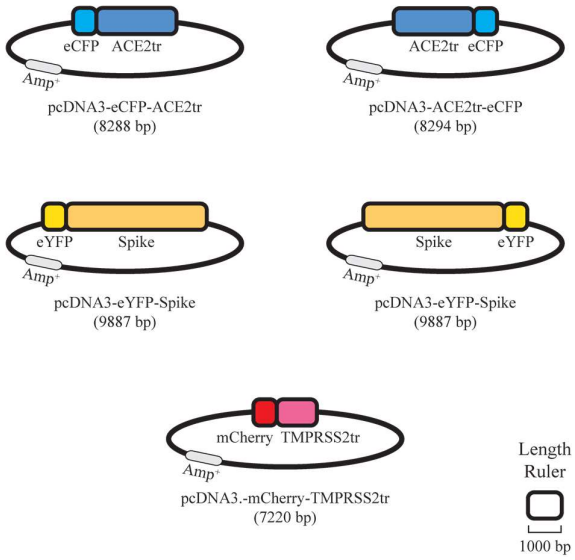


Figure 40. Structure overview of constructed plasmids expressing gene-reporter fusion proteins, consist of FRET-based PPI detection assay. Coding sequences of ACE2 and TRMPSS2 are truncated with only extracellular domain encoded, and coding sequence of Spike is wild-type. Plasmids include pcDNA3-eCFP-ACE2tr, pcDNA3-ACE2tr-eCFP, pcDNA3-eYFP-Spike, pcDNA3-Spike-eYFP, and pcDNA3-mCherry-TMPRSS2tr. Plasmid of pcDNA3-TMPRSS2tr-mCherry is failed in construction.

4.3.2. AphaFold2 predicts structures of fusion proteins and examinations of impacts on PPI among fusion proteins

To evaluate the structural integrity and functionality of the fusion proteins, their three-dimensional structures were predicted using the AlphaFold2 algorithm, which provided insights into the spatial arrangement of the fluorescent reporters relative to the proteins of interest (Figure 41). The predictive analysis indicated that for both the 5' and 3' fusions of the fluorescent reporters with ACE2tr and Spike proteins, there were no significant structural alterations that might impinge on the functional domains of either the reporter or the target proteins. Notably, the fluorescence reporter segments were positioned externally relative to the core structures of ACE2tr, TMPRSS2tr and Spike, suggesting that the fluorescent tags do not interfere with the native protein functions. Furthermore, the spatial separation of the fluorescent reporters from the proteins of interest implies that the efficiency of energy transfer, critical for FRET-based assays, remains unimpeded by direct physical interactions between the proteins

of interest. This structural configuration ensures that any FRET signal changes observed in experimental setups are likely due to alterations in PPI rather than obstructions caused by the fusion constructs themselves.

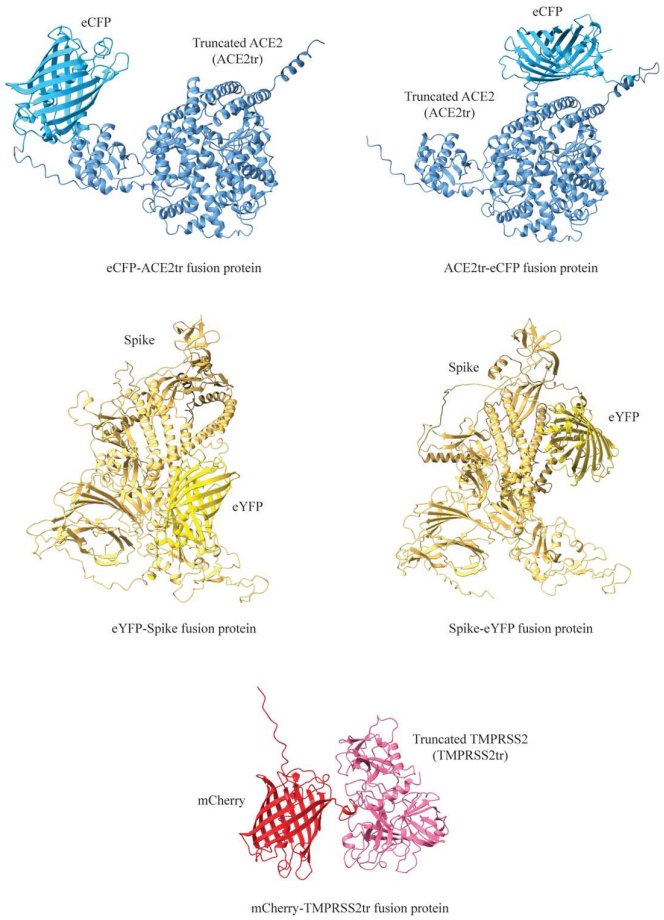


Figure 41. Predicted folding structures of fusion proteins through AlphaFold2. The amino acid sequences derived from coding sequence of gene-reporter fusion proteins are used in folding structure predictions as one complete protein.

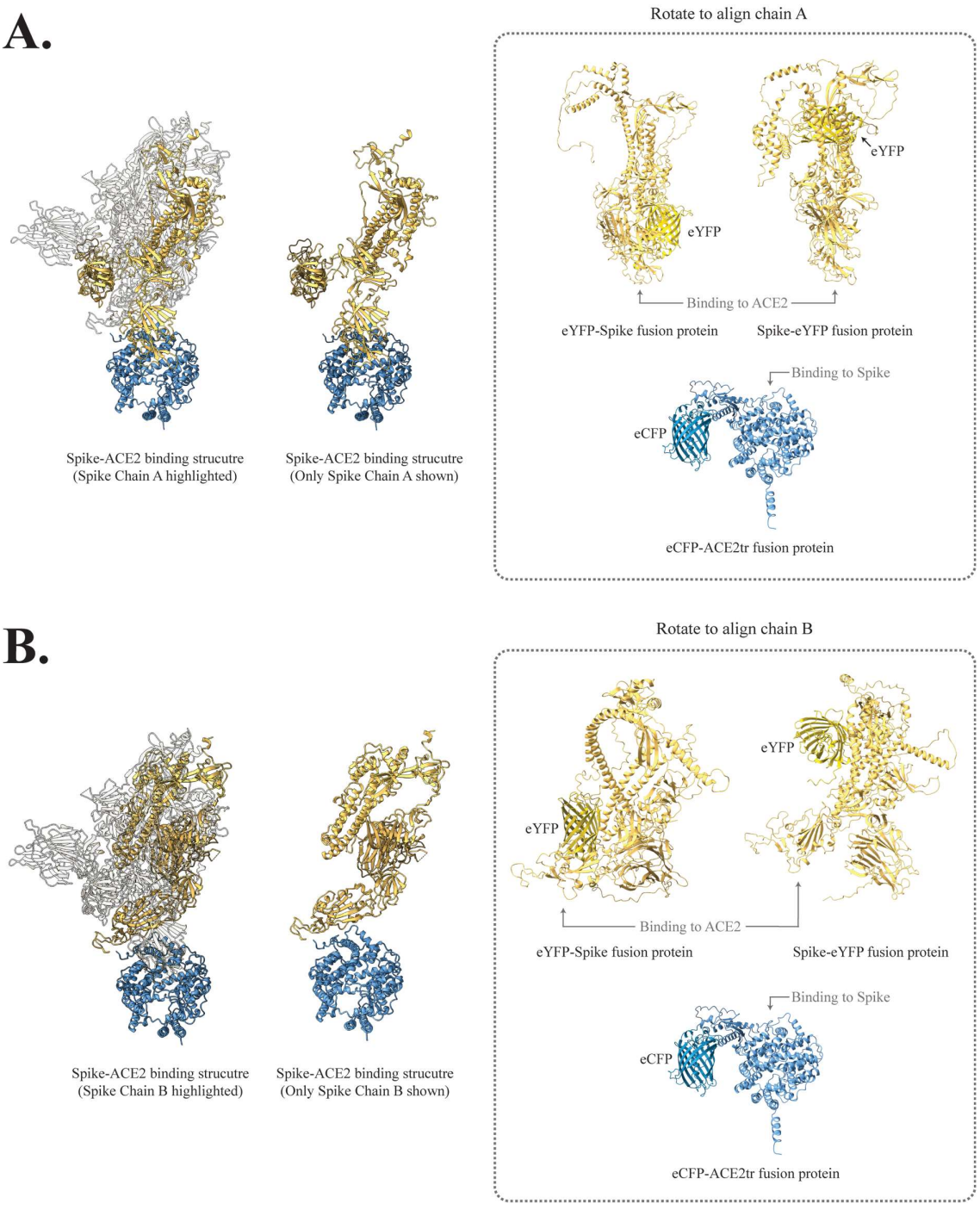


Figure 42. Demonstrating predictive impacts of fused fluorescence reporters on ACE2-Spike interactions. The Spike protein s form trimers during virion assembly, which is notated as Chain A, B, and C in reference structural model. Chain A and Chain B are involved in ACE2-Spike interactions. (A) The Spike-eYFP fusion protein with either 5' or 3' manner are rotated to align Chain A; (B) The Spike-eYFP fusion protein with either 5' or 3' manner are rotated to align Chain B.

To evaluate potential structural differences and the effects of fluorescence reporter fusion on PPI between fusion proteins and their wild-type counterparts, the three-dimensional structures predicted by AlphaFold2 were compared with the previously established binding configuration of the SARS-CoV-2 Spike protein and human ACE2. Utilising the crystal structure from the RCSB Protein Data Bank (PDB ID: 7WPA), which depicts the interface between a SARS-CoV-2 Spike trimer and human ACE2, I conducted a detailed alignment of the predicted fusion proteins' structures to this reference. In the reference structure, the Spike protein naturally forms trimers, labelled as Chains A, B, and C. The interaction surface with ACE2 primarily involves portions of Chains A and B. For my analysis, I aligned the predicted structures of the fusion proteins to correspond with the orientations of Chains A and B, to specifically assess how the integration of fluorescent reporters might influence the interaction interface. Figure 42 illustrates these alignments and indicates that the fluorescent tags do not significantly disrupt the overall folding of the proteins. Importantly, the spatial arrangement ensures that the energy transfer critical for FRET assays remains effective, as the fluorescent reporters (eCFP and eYFP) are positioned laterally relative to the core interaction domains of ACE2tr and Spike. However, a notable concern emerged regarding the proximity of eYFP in the eYFP-Spike fusion to the ACE2 binding interface. This closeness in both alignment scenarios, whether aligned with Chain A or Chain B, suggests that the 5' fusion of eYFP might interfere with the binding efficiency and stability of the Spike-ACE2tr interaction. Such a positional relationship could potentially impact the binding kinetics, and by extension, the biophysical properties observed in FRET assays.

4.3.3. Optimising intracellular overexpression of fusion proteins and FRET signal detections

To confirm the intracellular expression of the fusion proteins and determine the optimal post-transfection time point at which the fluorescence intensity is at its peak in transfected cells, I carried out a series of experiments with HEK293T cell transfection, which was transfected with one of the various constructed plasmids encoding the fusion proteins. After transfection, I monitored the fluorescence emitted by these cells using live-cell imaging over a period of 120 hours (5 days). This approach allowed for real-time observation of the expression dynamics of

the fusion proteins and facilitated the identification of the time point when fluorescence intensity was maximised, indicating the most efficient expression period for each construct. The detailed observations and results of these experiments are presented in Figure 43 and further elaborated in Appendix 15, providing a comprehensive overview of the fluorescence profiles and their implications for subsequent experimental applications.

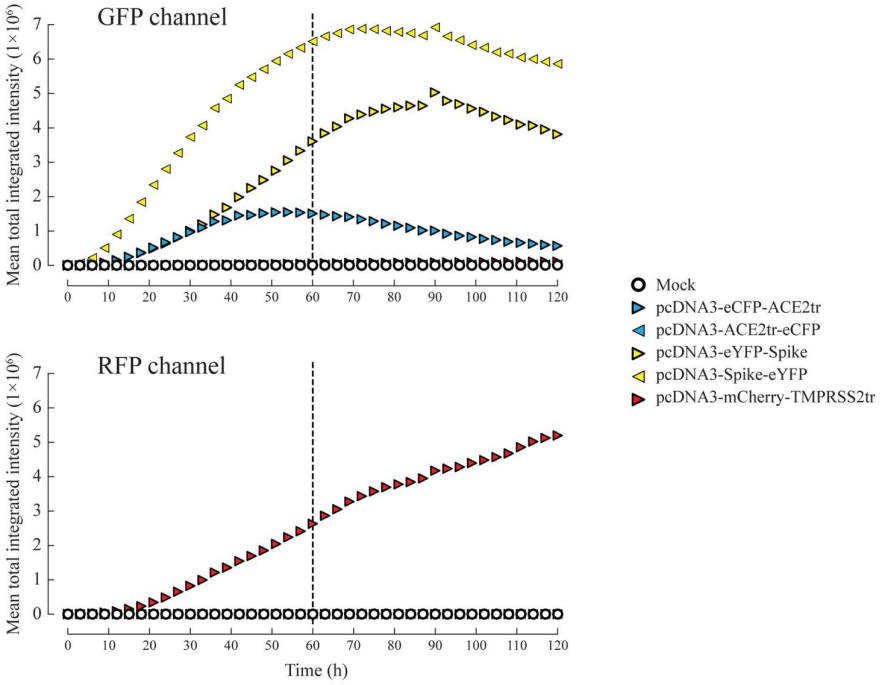


Figure 43. Overview of mean fluorescence intensity collected by live-cell imaging for plasmids transfection optimisation. The mean total integrated intensity is collected fluorescence intensity normalised by area. The GFP channel sub-optimally but efficiently detects fluorescence signals emitted from eCFP and eYFP fluorescence reporters. The RFP channel optimally and efficiently detects fluorescence signal emitted from mCherry fluorescence reporters. The time point of 60 hours post-transfection is considered the best to harvest cell for subsequent flow cytometry analysis.

The mean total integrated intensity, defined as the fluorescence intensity normalised by the well area, was employed to assess and compare the fluorescence intensity among cells transfected with different fusion protein plasmids. Unexpectedly, cells transfected with the pcDNA3-ACE2tr-eCFP construct exhibited virtually no detectable fluorescence emission. This observation suggests potential instability in the ACE2tr-eCFP fusion protein structure during

its folding and maturation processes. Alternatively, suboptimal transcriptional efficiency due to miRNA target sites embedded within the coding sequence might also account for this deficiency. In contrast to the pcDNA3-ACE2tr-eCFP, cells transfected with other fusion protein plasmids demonstrated significant fluorescence emission, aligning with expectations. Notably, the fluorescence emission curve for cells transfected with pcDNA3-mCherry-TMRPSS2tr did not display a distinct peak, indicating that the mCherry-TMRPSS2tr fusion protein likely exhibits substantial stability throughout its expression and maturation phases. Considering the fluorescence emission profiles of all fusion proteins, 60 hours post-transfection was identified as the optimal time point for harvesting cells for subsequent FRET detection experiments using flow cytometry.

4.3.4. FRET-based assay detects PPI among fusion proteins of ACE2, Spike and TMPRSS2

To elucidate PPI using FRET signals as an indirect measure, HEK293T cells were co-transfected with combinations of overexpression plasmids encoding fusion proteins (details of experimental setup provided in Appendix 16). These cells were analysed using flow cytometry to assess FRET signals (Figure 44). The pure FRET signals, along with the FRET efficiency, were quantitatively analysed using a bioinformatic tool developed for this purpose, the Cytometry Utilities Box Expansion (CUBE)[390]. This pipeline processes flow cytometry data to calculate FRET efficiency, which is indicative of the proximity between the fluorescent reporters of the FRET donor and acceptor molecules (detailed methodology in Section 4.2.3.4). FRET efficiency is essentially correlated with the physical distance between these fluorescence reporters, thereby providing a surrogate measure of the spatial closeness and interaction strength between the co-expressed fusion proteins. This approach allows for the indirect but effective quantification of PPI, offering insights into the dynamics of molecular interactions.

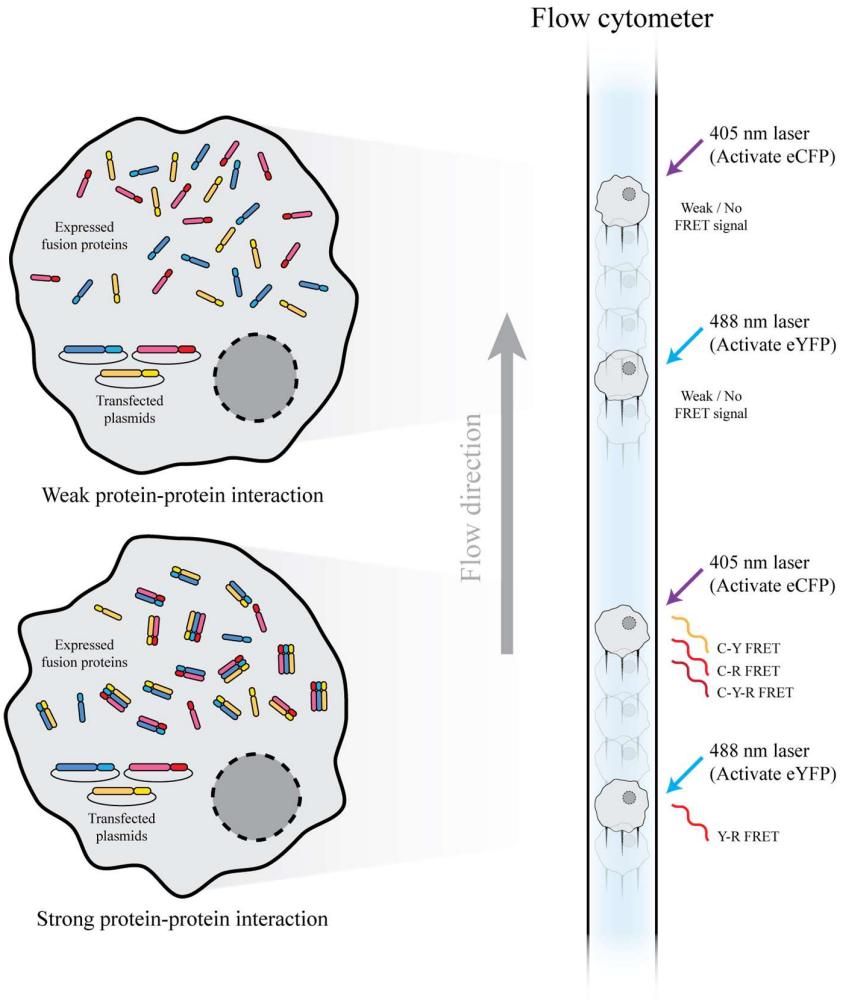


Figure 44. Schematic drawing of using flow cytometry to examine FRET signals from HEK293T cells transfected with fusion proteins, and the PPI among fusion proteins. The HEK293T cells express fusion proteins of eYFP-Spike, eCFP-ACE2tr and mCherry-TMRPSS2tr, and the higher FRET signals detected in flow cytometry suggest the higher PPI among of them.

Figure 45 presents the FRET efficiencies computed using the CUBE software from flow cytometry data. The analysis confirms successful detection of FRET signals across various transfection combinations, though with varying efficiencies. Notably, the C-Y FRET analysis shows that the eCFP-ACE2tr and Spike-eYFP combination yields higher FRET efficiency compared to the combination of eCFP-ACE2tr with eYFP-Spike. This difference suggests that the former combination is more sensitive in detecting the PPI between ACE2tr and Spike.

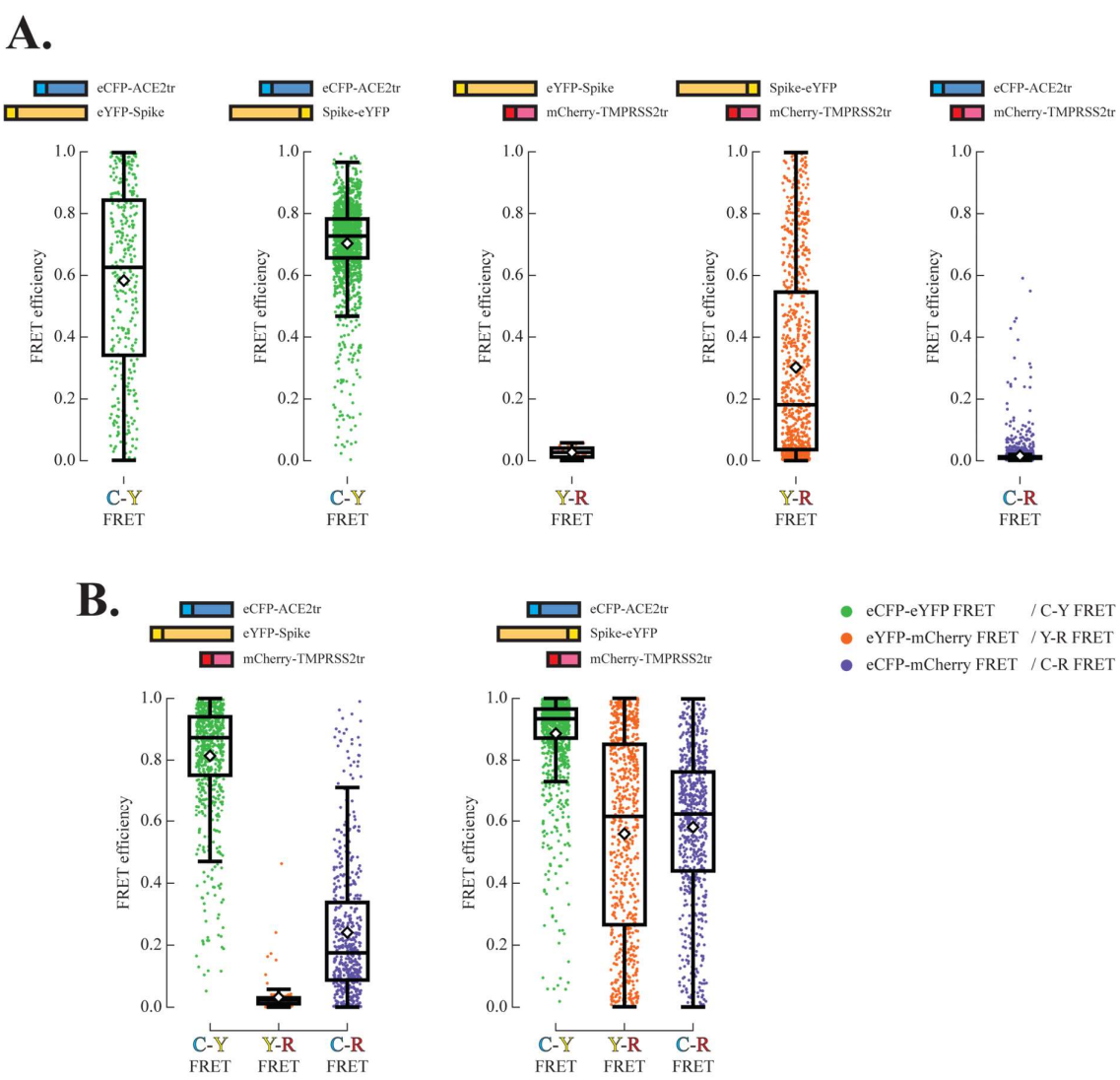


Figure 45. Single-cell FRET efficiencies of HEK293T cells transfected with various plasmid combinations. The FRET efficiencies of different cells are computed through CUBE with flow cytometry data of transfected cells. C-Y FRET is detected in cells transfected with eCFP and eYFP encoding plasmids. Y-R FRET is detected in cells transfected with eYFP and mCherry encoding plasmids. C-R FRET is detected in cells transfected with eCFP and mCherry encoding plasmids. All three C-Y FRET, Y-R FRET and C-R FRET are detected in cells transfected with eCFP, eYFP and mCherry encoding plasmids.

Interestingly, this finding corroborates earlier structural predictions indicating potential interference in Spike-ACE2 binding due to the proximity of the eYFP reporter on the Spike-ACE2 binding surface. Furthermore, in the Y-R FRET analysis, the pairing of Spike-eYFP

with mCherry-TMPRSS2tr demonstrates superior FRET efficiency over the eYFP-Spike and mCherry-TMPRSS2tr combination, enhancing detection sensitivity of the PPI between Spike and TMPRSS2tr. Conversely, the C-R FRET analysis between eCFP-ACE2tr and mCherry-TMPRSS2tr exhibits relatively low FRET efficiency, implying weaker interaction between ACE2tr and TMPRSS2tr. Figure 45.B showcases results from the three-way FRET assay, confirming the successful detection of C-Y, Y-R, and C-R FRET in both sample setups. The findings align with those from the two-way FRET assays, with all three FRET efficiencies notably higher when Spike-eYFP is used instead of eYFP-Spike, particularly in Y-R and C-R FRET analysis. Remarkably, the C-R FRET efficiency in the three-way assay is significantly elevated compared to its two-way counterpart, suggesting that the presence of Spike might enhance the interaction between ACE2tr and TMPRSS2tr.

4.3.5. Searching known RNA editing events on ACE2 and TMPRSS2 genes from the public database REDIportal

In the above section, we effectively detected FRET signals and quantified FRET efficiencies among various fusion protein combinations, confirming the viability of using intracellular overexpression of fusion proteins to indirectly assess protein-protein interactions (PPI). This plasmid-based approach permits the exploration of potential changes in PPI resulting from mutations in proteins of interest by manipulating the plasmid sequences accordingly. Focusing on the effects of RNA editing within the scope of this research, particular attention is directed toward mutations in ACE2 and TMPRSS2 induced by ADAR1L, an IFN-inducible RNA editing enzyme widely up-regulated in diverse virus infections. Future studies will aim to map RNA editing events attributable to ADAR1L expression on ACE2 and TMPRSS2, which are the major contributors in SARS-CoV-2 entry.

I initially consulted the REDIportal database for A-to-I RNA editing events to investigate recorded modifications in the ACE2 and TMPRSS2 genes[398]. Surprisingly, while no A-to-I editing events were recorded for the ACE2 gene, several sites were noted in the TMPRSS2 gene (Appendix 12). It is plausible that under conditions of ADAR1L overexpression, such as during viral infection, ACE2 and TMPRSS2 transcripts may exhibit a higher incidence of RNA editing events. Additionally, the editing events documented in public datasets might be

influenced by specific transcript variants or discrepancies between the ADARs CDS and the sequences of ACE2/TMPRSS2 found in the NCBI RefSeq database, potentially leading to variant editing outcomes.

4.3.6. Constructions of ADAR1L, ACE2, TMPRSS2 over-expression plasmids and examinations of their intracellular expressions

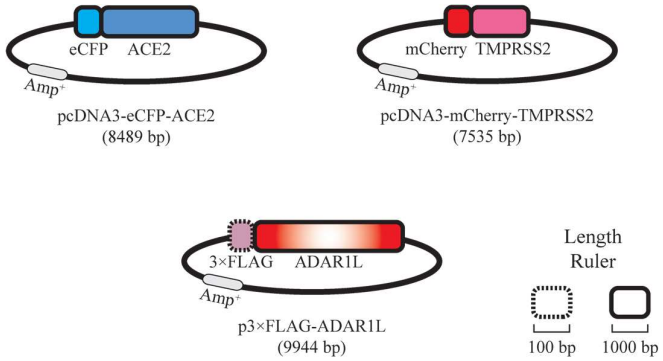


Figure 46. Structure overview of constructed plasmids expressing gene-tags or gene-reporter fusion proteins, consist of RNA editing events detection assay. Coding sequences of ACE2, TRMPSS2 and ADAR1L are wild-type. Plasmids include pcDNA3-eCFP-ACE2, pcDNA3-mCherry-TMPRSS2, and p3×FLAG-ADAR1L.

To validate or identify RNA editing events on ACE2 and TMPRSS2, I have engineered overexpression plasmids incorporating fusion tags for ADAR1L (3×FLAG at the N-terminus), wild-type ACE2 (eCFP at the N-terminus), and wild-type TMPRSS2 (mCherry at the N-terminus). The cloning strategies employed, and the relevant information have been documented in Appendix 30, which details all the methods and steps involved. The summary of all the plasmids created in this thesis is provided in Appendix 34. The successful construction of three plasmids, namely pcDNA3-eCFP-ACE2, pcDNA3-mCherry-TMPRSS2, and p3×FLAG-ADAR1L, is illustrated in Figure 46. Predictive modelling suggested that the

structures of the fusion proteins would remain independent from the native structures of the target proteins, thus preserving the functional integrity of both the fluorescent tags and the proteins of interest (Appendix 35). These fusion tags not only facilitate the microscopic visualisation of protein expression but also enable the verification of overexpression levels via Western blot analysis with anti-3×FLAG antibody.

To confirm the expression of the constructed plasmids, they were transfected into HEK293T cells individually and in combination. The expression of the transfected cells was then visualised using EVOS imaging technology (Figure 47.A). Fluorescence signals corresponding to either eCFP or mCherry were successfully detected in all transfected HEK293T cells, with eCFP signals observed in the CFP channel and mCherry signals in the RFP channel. This confirms the functional expression of the pcDNA3-eCFP-ACE2 and pcDNA3-mCherry-TMRPSS2 plasmids. Additionally, to assess the overexpression of ADAR1L, Western blot analysis was conducted using a specific antibody targeting the 3×FLAG tag. The results revealed distinct bands, indicating the successful overexpression of ADAR1L in cells transfected with the p3×FLAG-ADAR1L plasmid (Figure 47.B). Notably, the presence of multiple bands suggests the existence of several isoforms of ADAR1L within the cells. The predicted molecular weight of the 3×FLAG-ADAR1L fusion protein is 139 kDa, based on its amino acid sequence[391]. Given that the plasmid sequence was confirmed through Sanger sequencing and no internal stop codons were found within the ADAR1L coding sequence, the formation of these isoforms could be attributed to self-editing by ADAR1L. Because RNA editing cannot create stop codons, this self-editing may introduce rare codons into the plasmid transcripts, potentially causing premature termination of translation.

I could detect RNA editing events on ACE2 by comparing RNA sequencing results from cell line transfected either with pcDNA3-eCFP-ACE2 or with both pcDNA3-eCFP-ACE2 and p3×FLAG-ADAR1L. The same scenario is applicable for detecting RNA editing events on TMRPSS2.

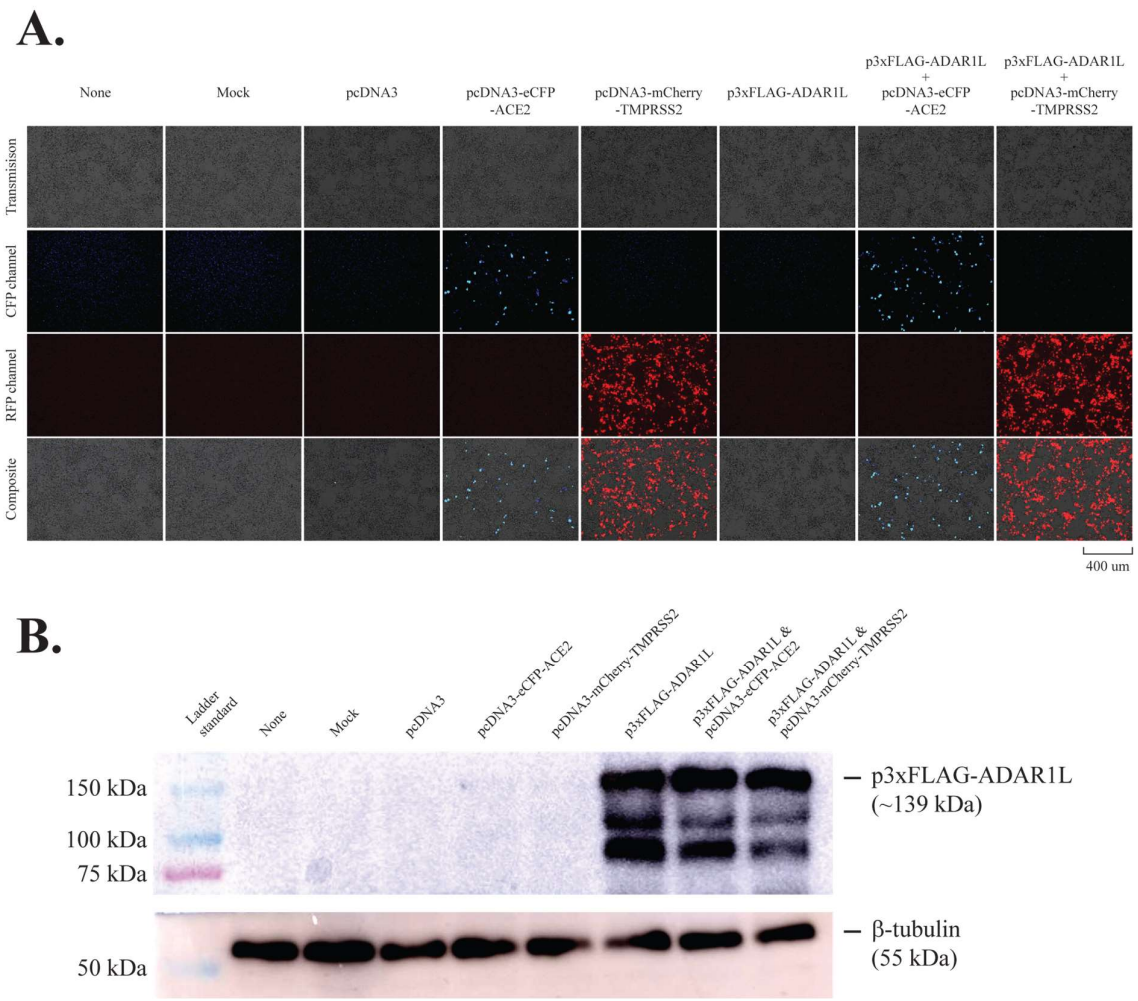


Figure 47. Verifying expression of constructed plasmids before identifying ADAR1L editing events. (A) Verifying expression of eCFP-ACE2 and mCherry-TMPRSS2 fusion proteins through imaging of transfected cells. CFP emission is detected only in pcDNA3-eCFP-ACE2 transfected cells and RFP emission is detected only in pcDNA3-mCherry-TMPRSS2 transfected cells. (B) Verifying expression of 3×FLAG-ADAR1L fusion protein through western-blotting. WB bands are only observed with p3×FLAG-ADAR1L transfected cells, and multiple bands are observed suggesting potential isoforms.

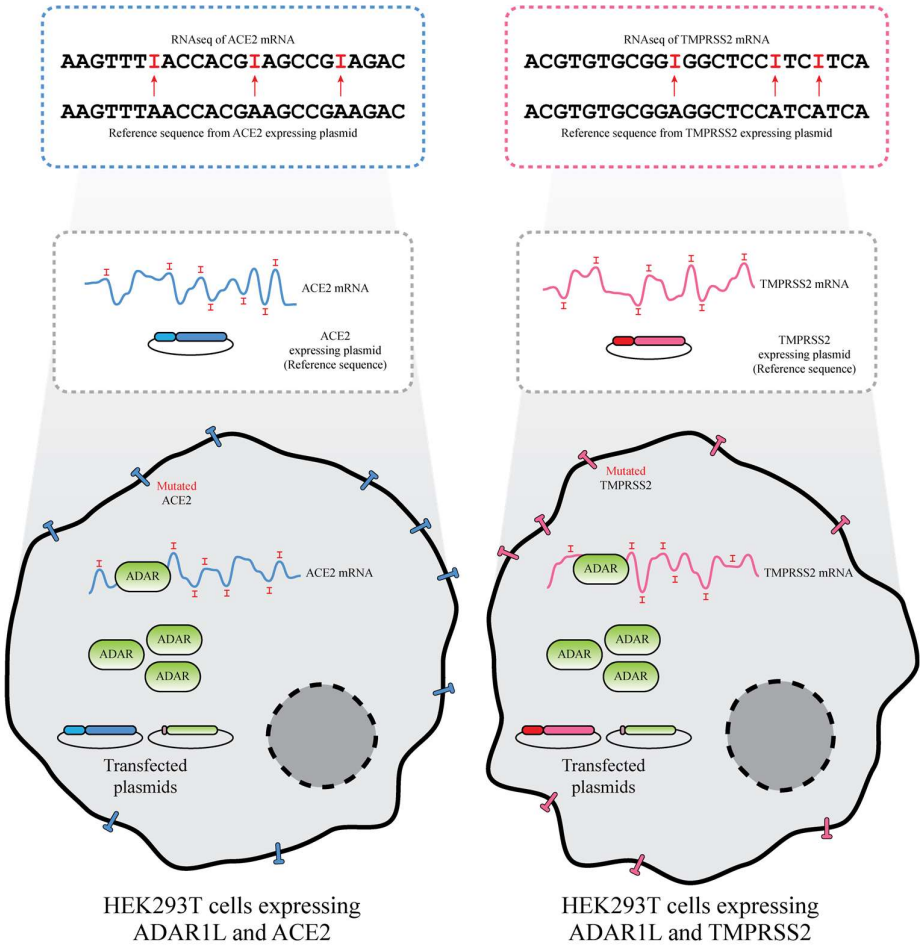


Figure 48. Schematic drawing of RNA editing events on ACE2 and TMPRSS2 transcripts mediated by ADAR1L overexpression. The ACE2 and TMPRSS2 transcripts are expressed by transfected plasmids pcDNA3-eCFP-ACE2 and pcDNA3-mCherry-TMPRSS2. And the ADAR1L protein is expressed by transfected p3×FLAG-ADAR1L plasmid. The RNA editing events will be detected by next-generation RNA sequencing with RNA samples extracted from cells transfected with different plasmids.

4.3.7. General simulation analysis of A-to-I editing events reveals altered codon and amino acid usage

Prior to identifying RNA editing events on ACE2 and TMPRSS2 mediated by overexpressed ADAR1L, all potential outcomes of RNA editing events resulting in codon or amino acid alterations were examined through simulation approaches. This provided an initial impression

of potential mutations that could be generated. Specifically, each adenosine (A) nucleotide within a codon (pre-edited codon) was individually substituted with guanine (G, since I is read as G) nucleotide (post-edited codon), resulting in pre-to-post mutation scenarios. This allowed for the examination of these mutations to understand which types of mutations could be expected in codons encoding various amino acids, taking into account that some amino acids might possess a higher frequency of A-enriched codons.

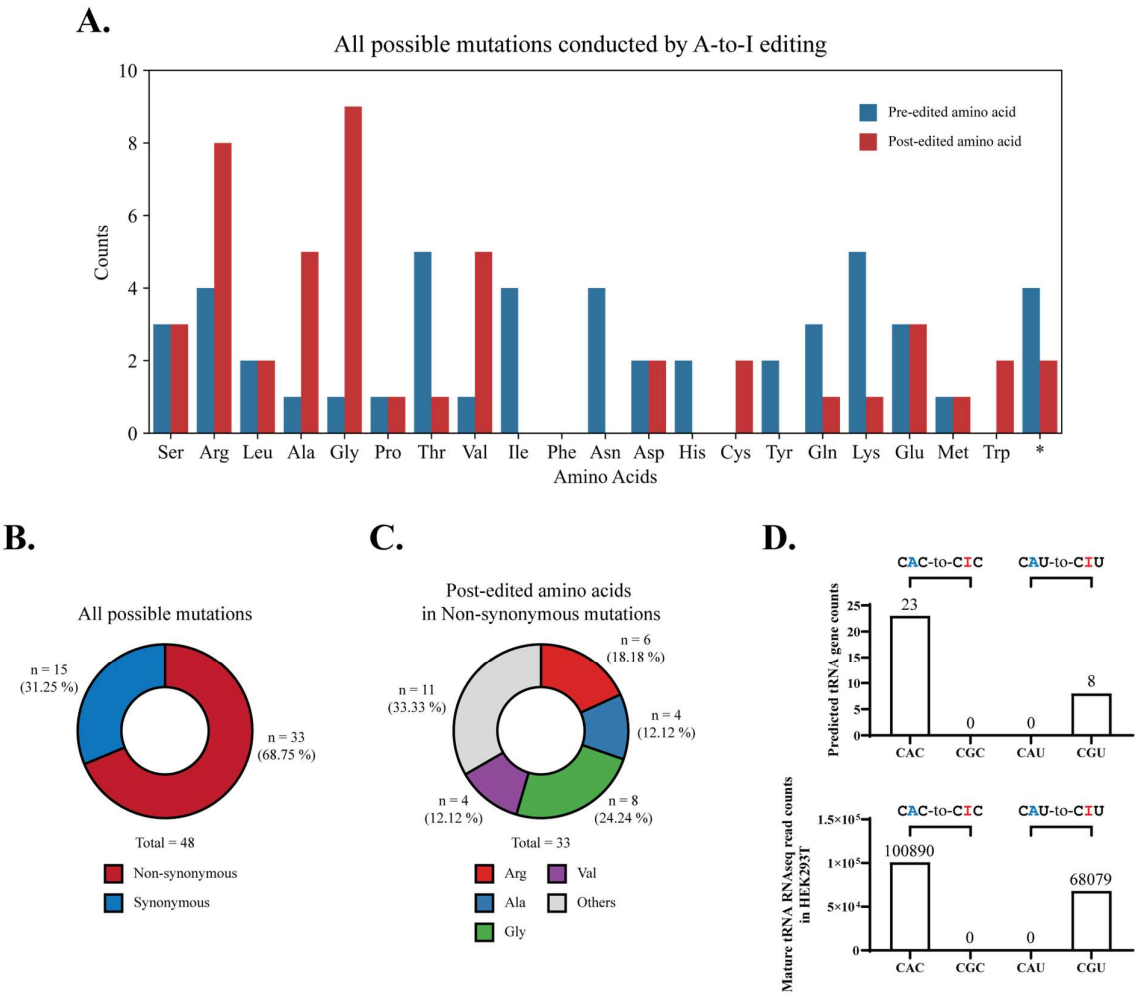


Figure 49. Simulation analysis of A-to-I editing events altering codon and amino acid usage. (A) Counts of pre-edited amino acids and post-edited amino acids are changed after applying A-to-I editing; (B) Majority of A-to-I editing generates non-synonymous mutations; (C) Majority of the post-edited amino acids in non-synonymous mutations are one of Arginine (Arg), Alanine (Ala), Glycine (Gly) and Valine (Val); (D) The data of predicted tRNA gene counts and mature tRNA RNAseq read counts in HEK293 reveal could identified the scratch the tRNA supplies of pre-edited codons and post-edited codons. The CAC-to-CIC mutation is only example of mutating non-rare codon to rare codon, whereas the CAU-to-CIU mutation is only example of mutating rare codon to non-rare codon.

Initially, I examined the overall counts of amino acids derived from pre-edited codons and post-edited codons as a result of the generated A-to-G mutations, and compared the two sets of pre-edited and post-edited amino acids (Figure 49.A). This analysis demonstrated that amino acids such as Arg, Ala, Gly, and Val are enriched in the post-edited amino acids compared to the pre-edited ones. This suggests that hyper-edited coding sequences may have a higher abundance of these amino acids. Conversely, amino acids like Thr, Ile, Asn, His, Gln, and Lys are less prevalent in the post-edited amino acids, indicating a reduction in their presence in hyper-edited coding sequences. Furthermore, RNA editing cannot generate stop codons, but it can convert UAG or UGA stop codons into the Trp-encoding codon UGG. Regarding mutation types, the majority of possible mutations are non-synonymous, implying a high likelihood of hyper-edited coding sequences producing mutated proteins with potentially altered functions (Figure 49.B). Additionally, the predominant post-edited amino acids in non-synonymous mutations resulting from A-to-G transitions are Arg, Ala, Gly, and Val (Figure 49.C, Appendix 36). In contrast, the pre-edited amino acids are more diverse in non-synonymous mutations (data not shown).

Another important aspect of RNA-editing-derived codon mutations is their relationship with rare codons, or codons with low tRNA abundance. I classified codons with tRNA levels below 10,000 (as measured by mature tRNA RNA-seq read counts in HEK293 cells, Figure 10). Subsequently, I isolated all mutations involving pre-edited or post-edited rare codons for further analysis. Surprisingly, nearly all mutations involving rare codons resulted in a transition from one rare codon to another, except for C<u>A</u>C-to-C<u>I</u>C and C<u>A</u>U-to-C<u>I</u>U transitions (Figure 49.D). Both of these transitions represent His-to-Arg mutations, where C<u>A</u>C-to-C<u>I</u>C is a transition from a non-rare codon to a rare codon, and C<u>A</u>U-to-C<u>I</u>U is a transition from a rare codon to a non-rare codon. But the biological implications of these RNA-editing-derived His-to-Arg mutations need further investigations.

4.3.8. Examining the quality of RNAseq data from different samples

Before identifying RNA editing events from aligned BAM files generated by STAR alignment with RNA-seq data and the Homo sapiens reference genome GRCh38.p14, I first evaluated the expression levels of various genes of interest. High expression levels of ACE2, TMPRSS2, and ADAR1L (or ADAR) were observed only in samples harvested from cells transfected with the corresponding plasmids (Figure 50). Furthermore, there were no significant variations in the expression levels of housekeeping genes, including GAPDH, ACTB, SDHA, and PPIA (Appendix 38). Additionally, there were no notable increases in the expression levels of other genes related to RNA editing, such as ADAR2, ADAR3, ADAT1, ADAT2, and ADAT3 (Appendix 38).

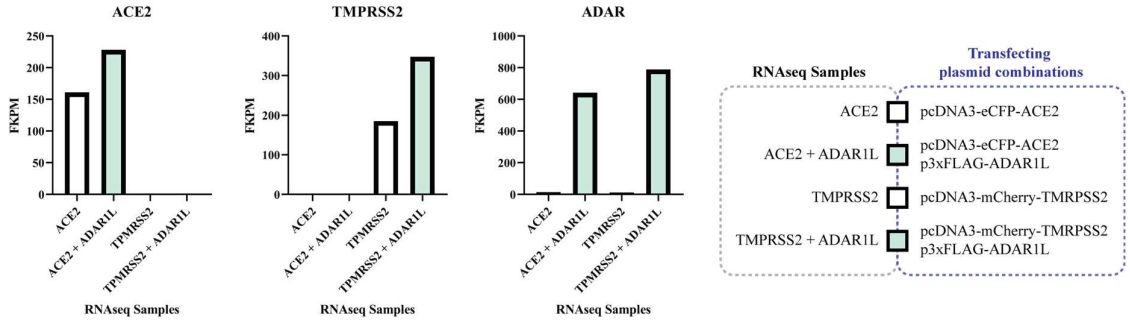


Figure 50. Expression levels (FKPM) of ACE2, TMPRSS2 and ADAR1L in various RNAseq data. RNAseq data was acquired from four RNA samples harvested from HEK293T cells transfected with various plasmid combinations. High ADAR1L expression levels are only detected in p3×FLAG-ADAR1L transfected samples. High ACE2 expression levels are only detected in pcDNA3-eCFP-ACE2 transfected samples, whereas high TMPRSS2 expression levels are only detected in pcDNA3-mCherry-TMPRSS2 transfected samples.

Subsequently, I analysed the counts of sequence reads aligned to ACE2 and TMPRSS2 (Appendix 39.A). The aligned read counts were significantly abundant, with each sample, whether for identifying ACE2 or TMPRSS2 RNA editing events, averaging at least 20,000 read counts aligned to different positions. This abundance ensures that the identified RNA editing events are highly reliable. Interestingly, the aligned read counts were notably higher in samples with ADAR1L overexpression. A potential explanation for this observation is the use of the empty pcDNA3 vector as a negative control. This vector was co-transfected with ACE2

or TMPRSS2 plasmids in the control group without ADAR1L overexpression. Although the empty pcDNA3 vector does not express any proteins, it still efficiently transcribes mRNA driven by the CMV promoter. Additionally, the transcribed mRNAs from the empty pcDNA3 vector are much shorter due to the absence of inserted genes, which may result in higher transcription efficiency. Furthermore, by comparing the counts of sequence reads aligned to various positions with and without ADAR1L overexpression, I observed a significant correlation in the read counts across different positions between samples. This suggests that the subsequent analysis of editing efficiency of A-to-I editing events is unlikely to be affected by imbalances in read alignment (Appendix 39.B).

4.3.9. Identifying RNA editing events on ACE2, TMPRSS2

The RNA transcripts were transcribed from plasmids with inserted ACE2 and TMRPSS2 coding sequences, driven by the CMV promoter. Consequently, the RNA editing events on the coding sequences of ACE2 and TMRPSS2 were exclusively included in the subsequent analysis. To further analyse the editing efficiency of various RNA editing events, the I/A ratio was used as a critical metric. This ratio is calculated by dividing the G counts (since I is read as G) by the A counts among all aligned sequences at each position. By comparing the I/A ratio at the same position between samples with and without ADAR1L overexpression, RNA editing events mediated by ADAR1L could be identified through significantly altered I/A ratios. Initially, all I/A ratios at various positions were compared between samples with and without ADAR1L overexpression using the U-test (Figure 51). Notably, there was a significant elevation in the I/A ratios in the coding sequences (CDS) of TMPRSS2 transcripts when ADAR1L was expressed. In contrast, the changes in the ACE2 transcripts were not as significant.

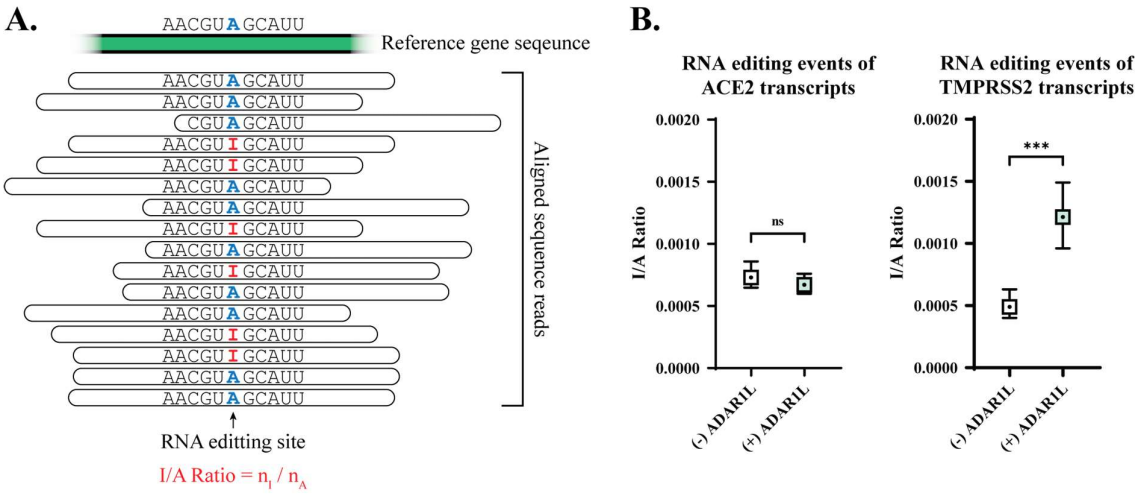


Figure 51. Demonstration of the I/A Ratio of RNA editing events identified from ACE2 and TMPRS22 transcripts. (A) Demonstration of calculation of I/A Ratio, which is calculated by dividing counts of I by counts of A in the same position from all aligned sequence reads; (B) Boxplot demonstrates the I/A Ratio of RNA editing events identified from ACE2 and TMPRS22 transcripts. The I/A Ratio are compared between samples with or without ADAR1L overexpression through Mann Whitney U-test. The I/A Ratio of RNA editing events on ACE2 transcripts are not significantly changed under ADAR1L expression, whereas the I/A Ratio on TMPRSS2 are significantly increased under ADAR1L expression.

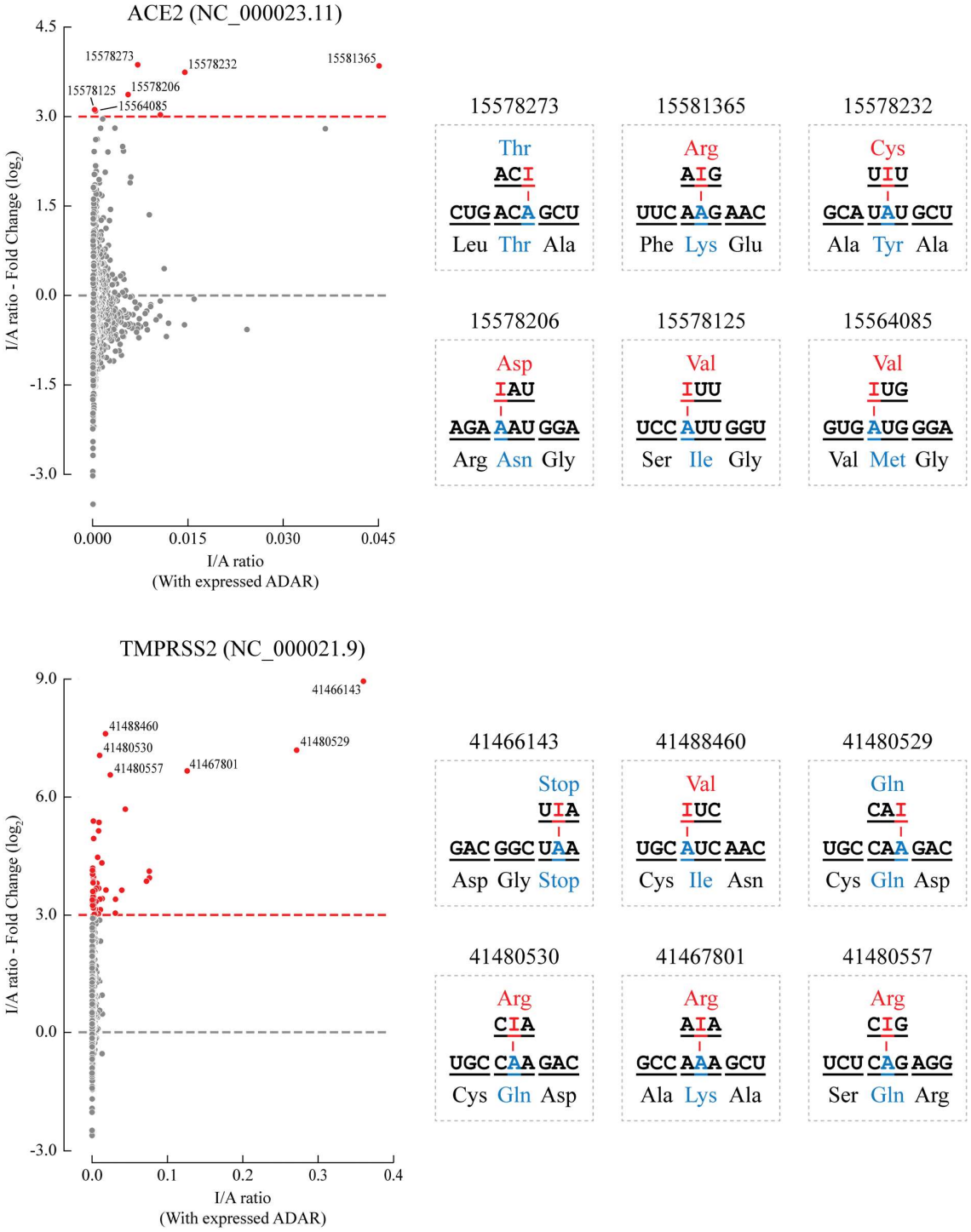


Figure 52. Identifying significant RNA editing events based on log-transformed fold change (log_2) comparing I/A Ratio between samples with or without ADAR1L overexpression. The RNA editing events with log fold change higher than 3 are considered as significant RNA editing events. The top 6 RNA editing events with highest I/A Ratio are highlighted on the right for both ACE2 and TMPRSS2.

The I/A ratio at the same position between samples with and without ADAR1L overexpression was compared using log fold change (log_2) calculation. Positions with an I/A ratio log fold change exceeding 3 were classified as RNA editing events mediated by ADAR1L (Figure 52). The threshold of an I/A ratio of 3 is determined because the lowest detected values are approximately -3. Therefore, the range between -3 and 3 is considered potentially unreliable. As anticipated, more RNA editing events were found in TMPRSS2 compared to ACE2. Specifically, 55 RNA editing events were identified in the coding sequence (CDS) of TMPRSS2, whereas only 7 were found in the CDS of ACE2. Notably, there were no RNA editing events previously identified in the CDS of ACE2 and TMPRSS2 according to the REDIportal database, with only a few events identified in the UTR of TMPRSS2 (Appendix 12). The highest I/A ratio recorded in the ACE2 CDS was 0.045, representing a 3.851-fold increase compared to the sample without ADAR1L overexpression. In contrast, the highest I/A ratio recorded in the TMPRSS2 CDS was 0.360, which is almost a 9-fold increase (8.944) compared to the sample without ADAR1L overexpression. Subsequently, I analysed all codon alterations caused by the identified RNA editing events. I highlighted the 6 identified RNA editing events with the highest I/A ratio fold change for both ACE2 and TMPRSS2 in Figure 52. Interestingly, the majority of the RNA editing (A-to-I editing) events resulted in non-synonymous mutations leading to amino acid alterations in subsequent translations. Specifically, 6 out of 7 identified editing events in ACE2 were non-synonymous, and 37 out of 55 identified editing events in TMPRSS2 were non-synonymous. It would be intriguing to investigate whether these amino acid alterations influence binding interactions among the SARS-CoV-2 Spike protein and human ACE2 and TMPRSS2. Additionally, some RNA editing events resulted in synonymous mutations without amino acid alterations. However, it remains questionable whether such synonymous mutations could affect tRNA availability and thereby alter translational efficiency.

Interestingly, the stop codon UAA of TMPRSS2 was found to be edited into either UA<u>I</u> (UAG) or U<u>I</u>A (UGA) with a high I/A ratio, suggesting that both A nucleotides in the UAA stop codon are editable by ADAR1L. Additionally, C<u>A</u>C-to-C<u>I</u>C mutations were identified in both the 138th and 279th CAC codons on TMPRSS2 CDS. It is uncertain whether this type of mutation will lead to translational interruption when the corresponding tRNA supplies are insufficient.

4.3.10. Self-editing of ADAR1L

As self-editing has been previously documented with ADAR2[28], there is a high likelihood that ADAR1L, which has broader substrate specificity, also exhibits similar biological activity. Moreover, the presence of multiple bands observed in the western blotting experiments of ADAR1L overexpressing cell samples (Figure 47) may indicate self-editing reactions of ADAR1L as well.

Here, I further analysed the RNAseq reads from samples with ADAR1L overexpression to identify RNA editing events on ADAR1L transcripts, setting the I/A Ratio threshold higher than 0.001 for RNA editing sites. A total of 387 RNA editing events were identified, with the highest I/A Ratio recording >0.35, suggesting approximately one-third of the transcript is edited at this site (Figure 53.A). Subsequently, I examined the domain locations of non-synonymous mutations, finding that the majority of RNA editing events are situated within various functional domains (Figure 53.B). Interestingly, almost one-third of the non-synonymous mutations are located in the A-to-I editase domain, possibly due to its longer length compared to other domains.

Regarding the multiple bands observed in the previous western blotting experiments, I hypothesised that the editing-derived mutation CAC-to-CIC may lead to premature termination of translation due to insufficient tRNA supply. I identified all instances of CAC-to-CIC mutations and predicted the protein sizes using online bioinformatics tools based on the coding sequences from the start site to the CIC mutation sites (Figure 53.C). After translating the coding sequences into amino acid sequences, I found that the 780th CIC and 875th CIC mutations correspond to protein sizes of 86.05 kDa and 96.45 kDa, potentially explaining the lower bands observed in the western blotting. Similarly, the 1014th CIC and 1129th CIC mutations correspond to protein sizes of 111.78 kDa and 124.75 kDa, which may account for the middle bands observed. However, these are speculative assumptions that require further experimental validation.

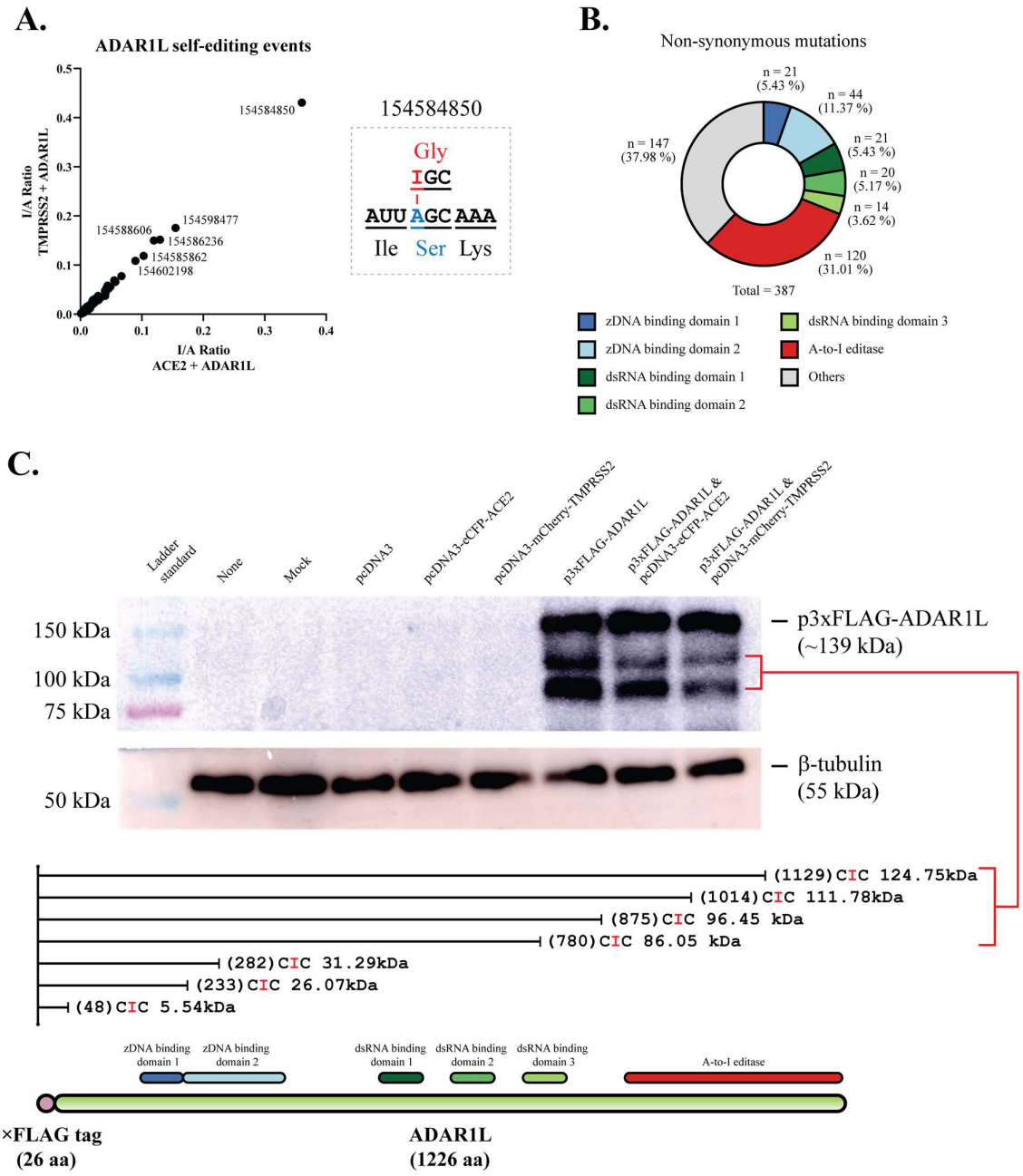


Figure 53. Analysis of RNA editing events on ADAR1L transcripts, which are mediated by ADAR1L self-editing. (A) Self-editing events of ADAR1L with high I/A Ratio identified in RNAseq data; (B) Proportional distribution of ADAR1L self-editing events in ADAR1L domains; (C) Demonstrating the assumed multiple isoforms created by the CAC-to-CIC mutations, potentially leading to premature termination in translation due to insufficient tRNA supplies. These events corresponding different bands observed in previous western-blotting experiment.

4.4. Discussion

In this chapter, I explore the hypothesis that RNA editing, facilitated by antiviral signalling, may introduce mutations in host genes that influence virus entry. This investigation focuses on a model involving the interaction of the SARS-CoV-2 Spike protein with human ACE2 and TMPRSS2 receptors. I hypothesised that mutations in human ACE2 and TMPRSS2, driven by RNA editing, could potentially alter the protein-protein interactions (PPI) between Spike, ACE2, and TMPRSS2, thereby impacting viral entry. To assess changes in PPI caused by RNA editing, I employed a FRET-based fusion protein assay. This method successfully captured clear FRET signals indicative of the interaction between the SARS-CoV-2 Spike protein and human ACE2 and TMPRSS2, with the detected FRET signals serving as a quantitative readout of PPI among these proteins of interest. Furthermore, I identified specific RNA editing events on the ACE2 and TMPRSS2 transcripts by comparing RNA sequencing data from HEK293T cell line with and without ADAR1L expression.

In this research, I have evaluated the capabilities of an intracellular FRET-based PPI detection system, which is designed to assess the impact of RNA editing-derived mutations on PPI dynamics. Additionally, I successfully identified RNA editing events on ACE2 and TMPRSS2 transcripts within mammalian cells overexpressing ADAR1L. Based on these foundational results, I have developed a series of planned experiments aimed at elucidating the effects of RNA editing-induced mutations in ACE2 and TMPRSS2 on their respective interactions with the SARS-CoV-2 Spike protein. While I am unable to execute these experimental plans within the current timeframe, the intention is to determine how specific mutations influence the binding efficiency and interaction dynamics between ACE2 and Spike, as well as TMPRSS2 and Spike.

Before studying the impacts of individual RNA editing events on PPI between Spike, ACE2, and TMPRSS2, I could also find out the global PPI changes among Spike, ACE2 and TMRPSS2 under ADAR1L overexpression, which is appliable if I add p3×FLAG-ADAR1L in the transfection combinations of FRET assay (Spike-eYFP, eCFP-ACE2tr, mCherry-TMPRSS2tr). I was planning to compare the FRET signal varies between cells expressing fusion proteins with and without ADAR1L overexpression. Based on the increases or decreases

of the FRET signals, I would have evidence of ADAR1L editing impacting on PPI among Spike, ACE2 and TMPRSS2.

Thus, I planned to mutate the fusion protein over-expression plasmid through site-directed mutagenesis according to the RNA editing events detected in RNA sequencing. The mutations containing plasmids will be used in transfection again to detect any changes in FRET signals compared to un-mutated plasmids. This experiment will help identify the critical RNA editing events that may have important impacts on SARS-CoV-2 entry.

To further evaluate PPI changes caused by RNA editing derived mutations, pseudo-virus assay may be used. The SARS-CoV-2 pseudo-virus could be constructed with published method similar to authentic lentivirus construction method with plasmids psPAX2, pMD2.G, pBOB-CAG-GFP and pBOB-CAG-SARS-CoV2-Spike-HA[399]. This method will generate SARS-CoV2 pseudo-typed lentiviral vectors that incorporated Spike GP as their envelope protein carrying a payload expressing eGFP under the CAG promoter. The pseudo-typed lentiviral vectors will 'infect' and enter ACE2 expressing cells and the 'infected' cells will emit detectable green fluorescence due to released eGFP expressing payload. Thus, I could over-express either wild-type ACE2 or RNA editing derived mutations contained ACE2 before application of SARS-CoV2 pseudo-typed lentiviral vectors. By examining changes in eGFP emissions, the SARS-CoV-2 host entry efficiency could be quantitatively examined.

In fact, previously published research has demonstrated various FRET-based techniques for detecting binding interactions among Spike, ACE2, and TMPRSS2[400, 401]. However, these methods rely on purified proteins and are limited in their ability to detect binding affinity variations with high throughput or achieve the level of accuracy offered by the method I employed.

There are some concerns and potential improvements regarding research of this chapter. In this chapter, I used truncated ACE2 (ACE2tr) and TMPRSS2 (TMRPSS2tr) having only extracellular domain to study PPI among Spike, ACE2 and TMPRSS2. The exclusion of these domains might alter the native folding structures of ACE2 and TMPRSS2, potentially leading to structural discrepancies between the full-length and truncated variants. Not to mention the addition of fluorescence reporter sequences, which also have the potential to alter structures of interested proteins although the chances are small. Although the FRET-based method

employed provides an indirect measure of interaction by detecting the proximity of fluorescent tags attached to the binding proteins, it does not directly assess the interaction between the proteins themselves. Therefore, it is advisable to conduct follow-up experiments, such as pseudo-virus entry efficiency assays, to confirm the functional consequences of any observed changes in protein interactions. These assays could provide direct evidence on how mutations in ACE2 and TMPRSS2 impact their binding with the Spike protein and, consequently, the viral entry process of SARS-CoV-2. Unfortunately, due to time constraints within my PhD program, I was unable to complete this aspect of the experimental setup.

Another important concern is other possible aspects that could affect detected FRET signals. Because the FRET signals are detected at single-cell level, the changes in detected FRET signals are not purely caused by changes in distances between fluorescence reporters. Another crucial contributing factor is the association rate between FRET donors and acceptors. For instance, the decrease or increase of detected FRET signal in cells expressing mutated proteins may be caused by decrease or increase of the association rate instead of the distances between fluorescence reporters. In another word, the detected FRET signals are possible to be altered with changed association rate when the distances between fluorescence reporters remain the same. Moreover, because different fusion proteins are expressed with individual plasmids, the uptake rates of different plasmids would not remain the same in different cells, making the expression levels of fusion proteins distinct in single-cell level. This loophole is difficult to avoid in single-cell level assay, which the actual amount of intracellularly expressed proteins is difficult to quantify. One possible solution is to use a linker sequence linking both fusion proteins into one using cloning method. For example, I could clone a linker sequence between eCFP-ACE2tr coding sequence and Spike-eYFP coding sequence to construct pcDNA3-eCFP-ACE2tr-Linker-Spike-eYFP plasmid (Figure 54). Because the expression cassette containing both fusion proteins is on the same plasmid, both proteins will be equally expressed intracellularly in each cell after transfections. This improved assay will help keep detected FRET signal proportional to the distance between fluorescence reporters, which may have significant correlation to the PPI of two fused proteins. The (GGGGS)n linker sequence is one of the candidates due to its significantly high flexibility, which is likely to have minimal impact on the folding and interactions of fusion proteins[402]. This flexibility allows the linked fusion proteins to interact in a manner similar to that of free-flowing proteins, akin to tethering the fusion proteins at the ends of a string.

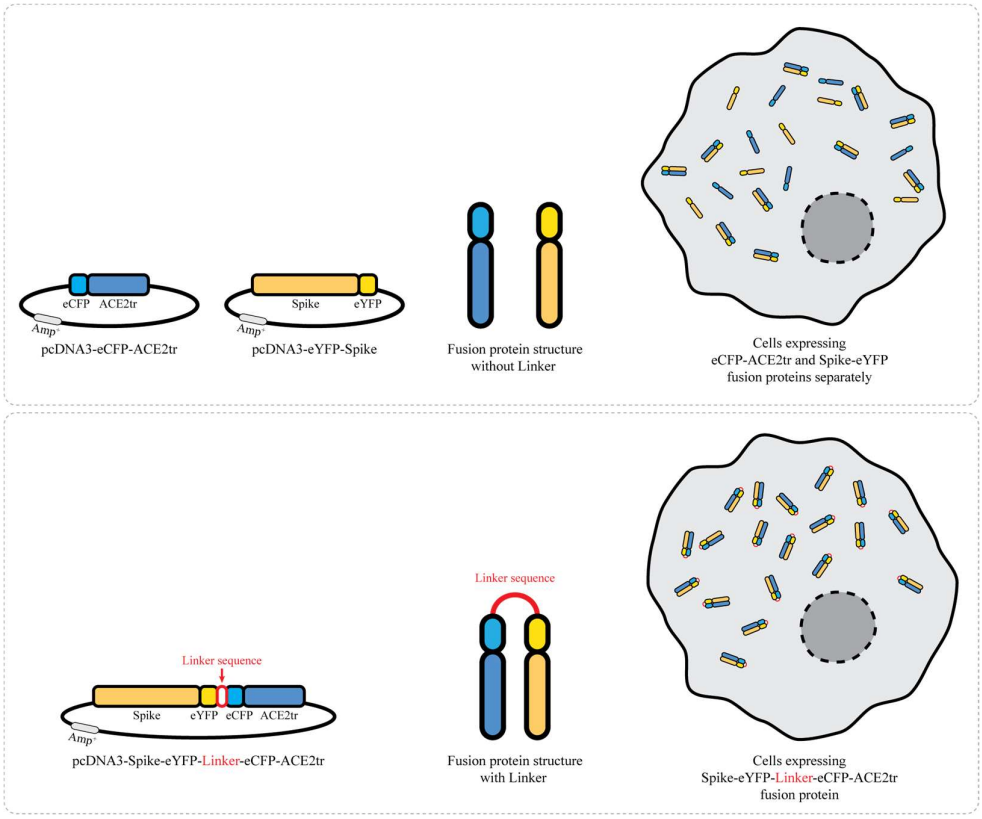


Figure 54. Demonstration of applying a linker sequence between fusion proteins. The fusion proteins of Spike-eYFP and eCFP-ACE2tr are shown as an example. The top box is showing the method expressing fusion protein separately, which the expression levels between the two fusion proteins are not controllable. The lower box is showing the method of adding linker sequence between the fusion proteins to construct a complete ORF, which expressions of Spike-eYFP and eCFP-ACE2tr will remain as 1:1 ratio. This will result a better accuracy and consistence of calculated FRET efficiencies.

Another nuanced scenario involves the orientation of the bound proteins potentially influencing the energy transfer efficiency between the FRET donor and acceptor. If the binding orientation causes the proteins to align in such a way that the bound structure obstructs the energy transfer pathway, the resulting FRET signals could be diminished or even absent, despite the proteins being effectively bound. This alignment factor is crucial as it can lead to misinterpretations of the FRET data, suggesting a lack of interaction when, in fact, a binding interaction is present but obscured by the physical configuration of the protein complex.

When exploring RNA editing events induced by ADAR1L on ACE2 and TMPRSS2 using the transfection of pcDNA3-ACE2 and pcDNA3-TMPRSS2 alongside p3×FLAG-ADAR1L, it's important to note that these plasmids only contain the coding sequences (CDS) of ACE2 and TMPRSS2, excluding the 5' and 3' untranslated regions (UTRs). ADAR1L typically recognises approximately 20 bp of substrate sequences, which suggests that the absence of the native UTRs in these constructs might influence the editing profile, particularly near the start and stop codons of the CDS. This could lead to a potential underestimation or misrepresentation of the editing events. Similarly, this consideration is also important if investigating self-editing events in ADAR1L transcripts, where modifications in UTR sequences or lack thereof could affect the detection and characterisation of editing sites on ADAR1L. Therefore, any identified editing events near the CDS boundaries should be validated with additional experiments to ensure their accuracy. Furthermore, the editing events identified through combinational overexpression can be validated using infected samples to determine whether similar RNA editing events occur. Detecting consistent editing events across diverse infected samples would provide stronger evidence for the importance of RNA editing in virus infections.

In the current experimental setup, another significant consideration is the potential expression of endogenous ADAR1L or other ADAR genes such as ADAR1S and ADAR2 within the HEK293T cells, alongside the overexpression of the exogenous 3×FLAG-ADAR1L fusion protein driven by plasmid transfection. This scenario complicates the attribution of observed RNA editing events specifically to the transfected ADAR1L, as they could also be derived from the cell's native ADAR1L expression. To address this, further experiments could involve the use of ADAR1L-knockout HEK293T cells, ideally generated using CRISPR-Cas9 technology, to eliminate background editing activity. Or those endogenous editing could included in the future PPI experiments because they may also contribute to the outcomes of virus infections. Additionally, considering the mild expression of ACE2 and TMPRSS2 in HEK293T cells, there is a possibility that some sequence reads attributed to these genes could originate from their endogenous expression rather than the transfected plasmids. This could be clarified by comparing the RNA sequences obtained from experimental assays to the sequences of the transfected plasmids, ensuring that only plasmid-derived transcripts are analysed. This approach would help refine the data analysis and enhance the reliability of attributing observed effects specifically to the experimental manipulations, thereby minimising confounding influences from native gene expressions.

To identify RNA editing events on ACE2 and TMPRSS2 mediated by ADAR1L, we employed next-generation RNA sequencing, which generated tens of thousands of sequences reads for each A nucleotide. This approach ensures robust detection of RNA editing events. Importantly, we confirmed that the identified RNA editing events were attributable to the reference isoform of ADAR1L by comparing ADAR1L-overexpressed samples with control samples. This comparison allowed us to exclude RNA editing events mediated by endogenous ADAR1L isoforms, which may have different substrate selections compared to the reference isoform. However, given that this experiment was conducted once, replication experiments are necessary to validate these findings.

Another significant observation is the self-editing phenomenon of ADAR1L, where numerous RNA editing events occur within its functional domains. Given ADAR1L's broad substrate specificity, these self-editing events may expand its substrate range, potentially generating isoforms with altered functionalities. Specifically, non-synonymous RNA editing events observed in the zDNA binding domains and dsRNA binding domains suggest potential functional alterations. During western blotting experiments, I also detected two shorter isoforms of ADAR1L. I hypothesise that these isoforms result from rare codon CIC mutations caused by insufficient tRNA supply, derived from CAC-to-CIC RNA editing events. Since the functional A-to-I editase is located near the 3' end of the ADAR1L coding sequence, it raises questions about whether these smaller isoforms retain A-to-I editing capabilities due to the possible truncation of the A-to-I editase domain. Additionally, it remains unclear whether these smaller forms of ADAR1L compete with the full-length ADAR1L for substrates, potentially modulating RNA editing activities and protecting RNA substrates from excessive editing. Further investigations are required to elucidate these hypotheses.

CHAPTER 5

Summary and Perspective

by Shuquan (Steve) Su

Keywords:

Insights, Future work

Chapter 5. Summary and Perspective

Identifying the host range of viruses, especially those previously unknown, is a crucial area of research for global virologists. The significant loss of human lives and economic disruptions caused by historical virus pandemics are often the result of unpredictable shifts in virus host ranges, particularly towards human hosts. Consequently, understanding virus host range or fitness in host is essential for effective pandemic prevention. Recent advancements in high-throughput sequencing techniques, particularly for RNA molecules, have made virus genome sequences much more accessible. As a result, these sequences are often among the first types of data generated from samples collected from infected patients during the onset of a pandemic[374]. Additionally, numerous virus genome sequences from wild fields have also been generated for research and public health monitoring purposes. Using this extensive virus genome sequencing data to predict or evaluate virus host range, especially concerning human hosts, is critical for pandemic preparedness. This is increasingly important as the potential for human-infectious viruses grows with the rising global population. The primary research focus of this thesis is to contribute to the prediction of virus host range using virus genome sequences derived data.

Although the infection processes vary among different viruses, there are several general steps in viral infection, including virus entry, viral gene expression, viral genome replication, virion assembly, and the release of new virions (Figure 55). In this thesis, my research primarily focuses on two key infection steps: viral gene expression and virus entry. The additional objective is to evaluate the potential outcomes of RNA editing on these processes. To contribute to the study of viral gene expression, I investigated virus codon usage biases and employed machine learning methods to assess the codon fitness of viral genomes in specific hosts. This approach aims to enhance our understanding of how virus codon usage affects viral gene expression and host adaptation. For the study of virus entry, I developed a FRET-based protein-protein interaction (PPI) detection assay to efficiently evaluate changes in PPI caused by various mutations. This assay can potentially help predict virus entry efficiency based on PPI predictions between the viral Spike protein and human receptors. This method provides a

framework for assessing how mutations in viral proteins may impact their interactions with host cell receptors and, consequently, the efficiency of virus entry into host cells.

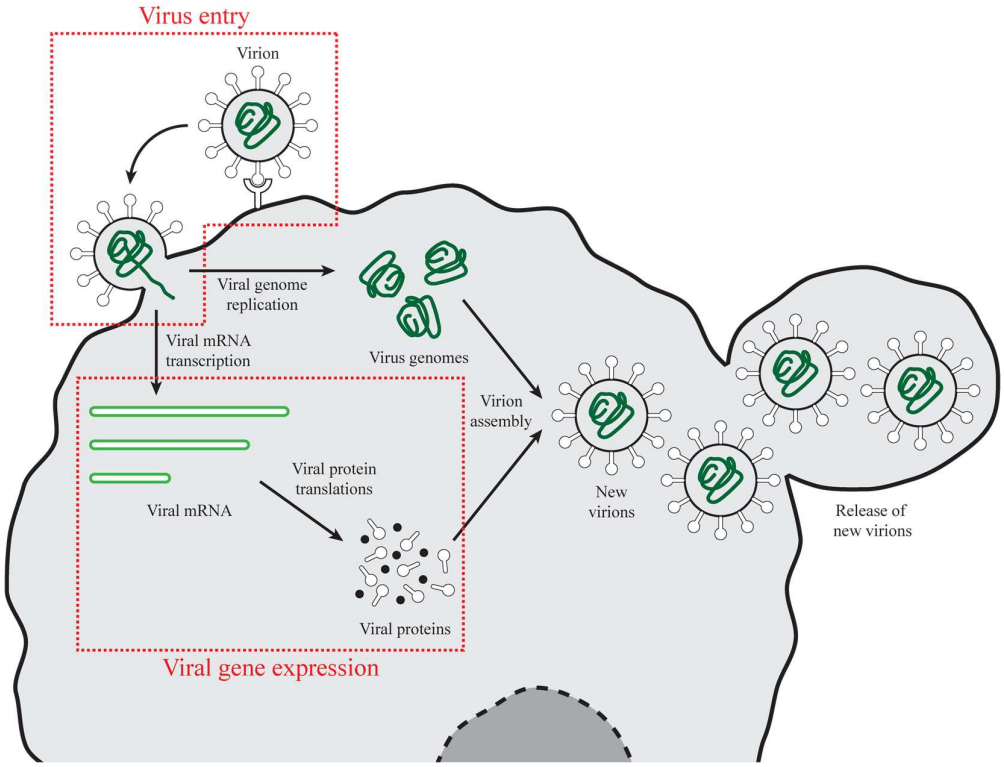


Figure 55. Demonstration of general steps in virus infection processes. The general steps include virus entry, viral gene expression, viral genome replication, virion assembly, and the release of new virions. In this thesis, research focus is paid at the steps of viral gene expression and virus entry.

5.1. Insights and subsequent work to investigate virus host codon fitness

For investigations into viral gene expression, my focus is on virus codon usage biases and virus-host codon fitness. In Chapter 2, I analyse the characteristics of amino acid and codon usages in human viruses by statistically comparing AminoAcid%, Codon%, and RSCU datasets against those of non-human viruses. The statistical tests reveal that the coding sequences of human viruses are more AU-rich compared to other viruses, particularly at the

third nucleotide, or wobble position. This suggests that human viruses are naturally more susceptible to ADAR editing due to the abundance of editable nucleotides. Additionally, the coding sequences of human viruses are enriched with ADAT-related amino acids and codons. The expression of ADAT2 and ADAT3 is predicted to enhance the expression of viral genes, or it could be a consequence of human viruses adapting to the human translational machinery. The study of codon usage biases concerning different ADAT relations, including ADAT-benefit, ADAT-suppress, and others, reveals that many ADAT-related amino acids are positively biased towards both ADAT-benefit and ADAT-suppress categories. Arginine exhibits a distinct pattern in codon bias and its relationship with ADAT. Specifically, it demonstrates a positive bias toward ADAT-non-related codons. This observation suggests that the translation of Arginine is likely not substantially influenced by ADAT2/3 expression, highlighting the potential unique roles of Arginine in viral infections.

The correlation between codon usage biases and tRNA supplies has been extensively studied. However, in this research, I did not investigate the correlation between codon usage biases of human viruses and the tRNA supplies of human hosts. Although I acquired data regarding predicted human tRNA gene counts and tRNA expression levels in HEK293 cells (Figure 10), I had concerns about the reliability of these data and could not find a better database to describe human tRNA supplies accurately. The predicted tRNA gene counts were obtained using the tRNAscan-SE algorithm applied to the human reference genome, but many of these predictions have not been experimentally verified. Additionally, tRNA gene counts may not correlate well with their capacity to decode corresponding codons due to variations in regulatory networks, promoter sequences, miRNA target sequences, and other factors. Thus, the expression levels or intracellular abundance of different tRNAs may vary significantly, meaning that a higher predicted tRNA gene count does not necessarily equate to higher tRNA abundance. Moreover, different tRNA isodecoders, which are tRNA molecules with the same anticodon but different body sequences, may have varying translational efficiencies, further complicating the relationship between gene count and functional tRNA abundance. Conversely, tRNA expression levels obtained from HEK293 cells through specialised RNA sequencing techniques provide a more accurate measure of available tRNA molecules and their sequences. However, there is currently no comprehensive database summarising this kind of data, making it difficult to analyse correlations with codon usage biases on a larger scale. Another concern is that tRNA gene expression may be regulated differently in the context of viral infections,

potentially altering tRNA abundance. There is limited research on changes in tRNA abundance during viral infections using similar sequencing techniques, making it challenging to conduct species-level comparisons between human and non-human viruses at this time, although this area holds promise for future studies.

Another significant finding in this chapter is the proposal and subsequent study of multi-codon usage biases in human viruses, adapted from the single-codon usage biases equation. Although statistical tests using RSMCU-n and NZP-n revealed significant variations between human and vertebrate viruses, the follow-up analysis was limited. A major issue with multi-codon usage biases is the high prevalence of zero values as the codon stretch length (CSL) increases, resulting in sparse RSMCU-n data matrices. The possible combinations of codons increase exponentially with longer codon stretches, leading to a higher percentage of zero values. This high zero-value rate prompted me to focus on non-zero RSMCU-n values for subsequent U-Test studies, as statistical analyses would be more reliable using actual RSMCU-n values rather than zero value abundance. When the CSL exceeds three, the zero value rates for different codon stretches exceed 90%, and the RSMCU-n values are often minimal. Therefore, I conducted additional analyses using NZP-n values to identify codon stretches with significant usage biases. Interestingly, the codon stretches identified using RSMCU-n differed significantly from those identified using NZP-n, suggesting that these analyses might require further refinement to better understand multi-codon usage biases in human viruses. Alternatively, developing entirely new computational metrics might be necessary for studying multi-codon usage biases effectively. This discrepancy underscores the complexity of multi-codon usage analysis and highlights the need for improved methodologies or novel approaches to achieve a comprehensive understanding of viral multi-codon usage patterns.

Chapter 3 successfully demonstrates the feasibility of using machine learning models to systematically evaluate codon usage biases in specific groups of coding sequences (i.e. virus genome of a certain group of viruses containing different CDS). The predicted probability generated by the Random Forest (RF) model, trained with datasets including codon usage biases data (RSCU) of all codons and other features, serves as a readout of codon fitness in specific hosts. This metric is denoted as the Virus Codon Fitness (VCF) score. I later utilised the human virus codon fitness score (HVCF score) in several case studies. These included comparisons between human-sourced and non-human-sourced viruses, monitoring HVCF

scores of SARS-CoV-2 during the COVID-19 pandemic, and simulating codon-based mutation paths between SARS-CoV-2 and other betacoronaviruses. These case studies underscore the practical applications of the HVCF score in understanding viral adaptation especially in the context of human hosts. By applying this machine learning approach, this chapter provides a robust framework for analysing codon usage biases and predicting host codon fitness, offering valuable insights into viral host adaptation and informing strategies for monitoring and managing viral outbreaks.

Although significant variations in HVCF scores exist between human and non-human viruses, these scores should not be over-interpreted. Specifically, HVCF scores represent the adaptation level of viral codon usage biases to the human host translational machinery. They do not correlate with viral lethality or infection outcomes. Furthermore, HVCF scores cannot determine whether a virus is capable of infecting a human host, as other factors such as the binding affinity between viral Spike proteins and human receptors, and viral mechanisms for immune escape, contribute to infection outcomes. However, it is safe to argue that codon fitness predictions, like HVCF scores, have potential as important contributors to host range prediction for unknown viruses. For instance, a genome sequence from a wild-field virus with a high HVCF score suggests a high probability of adaptation to human translational machinery, though it does not guarantee human infectivity. The Hepatitis D virus (HDV) provides an ideal example. HDV uses the Hepatitis B virus (HBV) surface antigen (HBsAg) as its envelope protein, proliferating only in patients infected with HBV[152, 153]. Thus, HDV may exhibit high codon fitness to the human host but requires specific mechanisms to enter host cells. Conversely, the efficiency of viral entry into host cells, determined by the binding interactions between viral Spike proteins and host receptors, is not the sole determinant of the viral host range. Conclusively, predicting virus host range is complex and involves multiple factors, including codon fitness, binding interactions between viral Spike proteins and host receptors, and identifying host miRNA binding sites. Systematically evaluating codon fitness in specific hosts through machine learning can significantly contribute to predicting virus host range.

Using RF models to predict host labels has several disadvantages, one of the most significant being the discontinuous nature of the predicted probability, which is used as the VCF score. This discontinuity arises from the inherent voting mechanism of the RF model, where each decision tree produces a binary True or False prediction, and the overall predict probability is

a fraction based on the number of trees that vote for True. Since the number of decision trees is fixed, the resulting HVCF score is inherently dis-continuous. Due to this discontinuity, multiple optimal mutations are often predicted when using HVCF scores as a gradient, as seen in the codon-based mutation simulations described in Chapter 3. This necessitates the use of additional metrics, such as the correlation coefficient to the codon counts of the target virus, to guide the simulations effectively. The primary reason for using an RF model was its ability to adapt to the RSCU metrics of codon usage biases, despite the challenging or possibly unknown data distribution of RSCU. Given these limitations, I suggest developing new metrics to better characterise codon usage biases before focusing on training improved models for evaluating codon fitness in hosts with greater accuracy.

The nature of the problem lies in extracting codon usage biases from viral gene coding sequences, which doesn't necessarily require a mathematical approach. Instead, I propose using natural language processing (NLP) to encode the coding sequences from a codon-based perspective. For instance, employing codon-based or amino-acid-based one-hot encoding can effectively encode the sequences. By leveraging an NLP model such as the Transformer encoder, we can process the encoded data for subsequent classification. This method encapsulates the codon and amino acid sequences into a sequence-based data format, which inherently includes characteristics of codon usage biases. Furthermore, as this method is fundamentally a sequence embedding technique, it not only comprehends the biases of single codons or amino acids but also the usage biases of codon stretches comprising multiple codons or amino acids. Consequently, this approach may provide a more comprehensive understanding of virus codon usage biases from various perspectives.

In the context of a virus genome, which comprises multiple viral genes, it's crucial to consider how to aggregate the encoded data from different genes to represent the entire virus genome accurately. The method I employed simply summed all the codon counts from different viral genes without accounting for their distinct functionalities, which is overly simplistic. Moreover, this approach tends to be biased towards genes with longer lengths rather than those that are crucial determinants of virus proliferation. To address this limitation, I propose encoding the gene names along with the sequences. This additional information could help the algorithm differentiate critical genes and assign different weights to them. As previously mentioned, various factors determine virus proliferation in specific host cells. In future work, deep learning

models could incorporate additional features related to these factors. For instance, complete sequences of different viral genes could be encoded to investigate potential miRNA regulations. Furthermore, predicting the spatial structure of the Spike protein based on the coding sequence using AlphaFold could provide valuable structural characteristics to enhance the model's performance.

In this thesis, I aim to predict virus codon fitness in various hosts using the model trained with RSCU data. However, the case study of HVCF scores from human-sourced and non-human-sourced virus genomes reveals distinct patterns among different viruses (Figure 32). Viruses such as MERS-CoV, West Nile virus, and Orthohantavirus show no significant differences between human-sourced and non-human-sourced virus genomes. This raises concerns about the biological variations among different viruses, especially those that are fundamentally different in classifications. Therefore, developing specific models tailored to predict specific viruses or classes of viruses could lead to more accurate and reliable predictions. Such models may be more practical in real-world scenarios, as researchers often focus on a single virus within specific contexts. For example, given the intense focus on SARS-CoV-2 during the COVID-19 pandemic, researchers are keen to understand how the virology of SARS-CoV-2 varies when infecting different hosts. In practical terms, if the goal is to predict the proliferation or infection efficiency of a specific virus, it may be beneficial to use all sequence-based data mined from that viral genome sequence. Additionally, incorporating data outlining specific cell lines or infecting subjects could significantly enhance, or even be a determining factor in, model training for predicting the viral proliferation efficiency in various settings.

To achieve this, various biological features outlining virus-host interacting activities could be considered. As mentioned earlier, the binding interactions between virus Spike proteins and human receptors are critical for virus entry into host cells and subsequent infections. For instance, in the case of studying a specific virus like SARS-CoV-2, which binds to human ACE2 and TMPRSS2, it would be practical to include the amino acid sequences of these human receptors using embedding methods such as Transformers. This addition would reflect the protein structure of the human receptors, enabling the model to identify critical protein structures with significant binding affinity to the Spike protein, indicating potential infection efficiency. Moreover, human host cells possess other anti-viral mechanisms such as exogenous RNA sensing and miRNA regulations. The amino acid sequences of genes involved in these

pathways can also be embedded using Transformer models. This gene-sequence-embedding approach has been successful in many research fields, demonstrating excellent performance in training models for various purposes[403, 404]. Many of these mechanisms are sequence-specific, meaning different virus genome sequences may encounter varying levels of anti-viral efficiency. For example, MDA5 (Melanoma Differentiation-Associated protein 5), a major player in the exogenous RNA sensing pathway, may exhibit varied efficiency in binding viral RNA depending on the sequences of the viral RNA. Therefore, including amino acid sequence data in the models can help them understand the interplay between virus genome sequences and human host anti-viral mechanisms, potentially enhancing the performance of predictions. A similar method can be applied to human host miRNAs, which have specific binding targets due to their seeding sequences correlating virus transcript sequences.

With the method mentioned above, various virus-infection-related factors could contribute cooperatively to model training. Most of this data, from either the target virus or human cell, can be acquired using sequencing technologies such as Sanger sequencing, RNA sequencing, and others, followed by computational processing. However, data on the prediction targets to determine the efficiency of virus proliferation or infection in various cell lines are relatively difficult to obtain. In traditional virology research, virus proliferation rates are often studied using cell-culture-based approaches such as virus plaque assays, which involve serial dilution of virion stocks to evaluate the virus load after infecting cell cultures[405]. Additionally, proliferation rates can be studied using recombinant virus construction methods by inserting fluorescence reporter coding sequences into the virus genome, which allows for the study of virus proliferation rates using fluorescence-based detection techniques[406]. These methods enable the evaluation of virus proliferation rates in different cell lines. Unfortunately, there is currently no database summarising such data, which would require significant effort to compile. Moreover, the data may be difficult to organise and potentially unreliable due to variations in experimental conditions across different studies. Therefore, it would be ideal to set up reliable experimental conditions and directly collect data from various cell lines infected by the target virus.

5.2. Insights and subsequent work to investigate protein-protein interaction among Spike, ACE2 and TMPRSS2

In my research on virus entry, I focused on studying protein-protein interactions (PPI) involving the virus Spike protein and human receptors. In Chapter 4, I successfully established a FRET-based PPI detection assay to efficiently study the impacts of protein mutations on PPI changes. I then identified several RNA editing events on mRNA encoding ACE2 and TMPRSS2, conducted by ADAR1L. These events were studied to understand the PPI changes in binding interactions among the SARS-CoV-2 Spike protein and human ACE2 and TMPRSS2 resulting from these RNA editing events. However, some of the designed experiments have not been completed yet, particularly regarding whether the RNA editing events could affect the PPI among Spike, ACE2, and TMPRSS2.

The FRET signals predominantly correlate with the distance between fluorescence molecules. However, intracellularly, the expression levels of these molecules are not equal, as they are expressed from different plasmids, which may have independent cell entry rates and expression levels. Additionally, when mutations are applied to the plasmids through site-directed mutagenesis, the binding rates of FRET donors and acceptors could change. Therefore, this plasmid-transfection-based assay requires further improvement in intracellular quantitative normalisation to achieve more accurate computational results. Improvements could be made in the experimental setup before flow cytometry. One method is to link both fusion proteins' coding sequences by applying linker sequence (e.g., Spike-eYFP-Linker-eCFP-ACE2tr), where the complete expression cassette is cloned into a single plasmid (e.g., pcDNA3 vector). There are various linker sequences designed for various purposes, some of which have both good flexibility and high translation efficiency[402]. With this method, both fusion proteins will be equally expressed because they are from the same plasmid entering cells and expressed from the same coding sequence sharing the same promoter sequence.

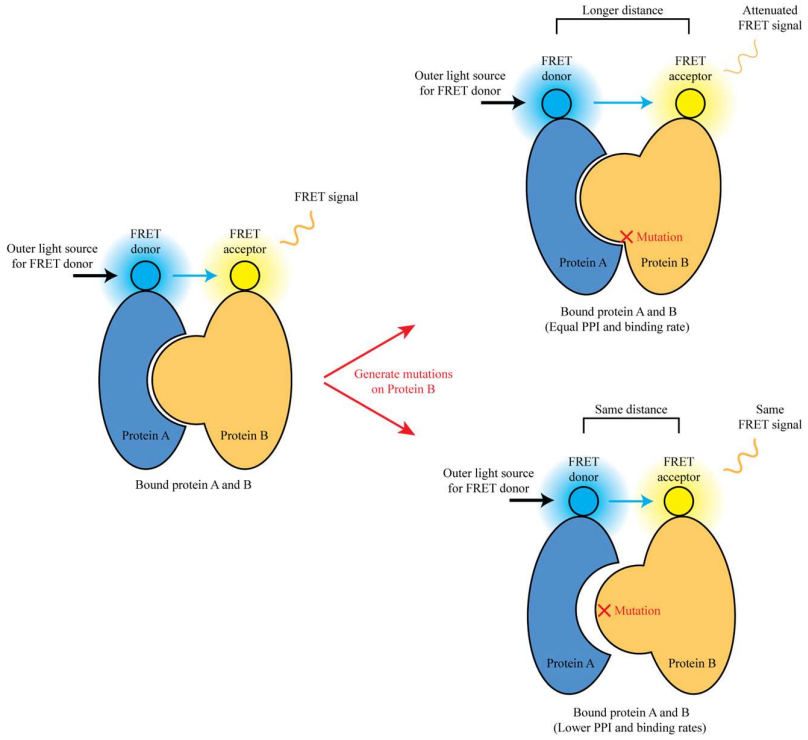


Figure 56. Demonstrating potential impacts of generated protein mutations on PPI and FRET efficiency in some scenario.

In the discussion section of Chapter 4, I mentioned a defect of the FRET efficiency computed from the FRET-based PPI detection assay. It is affected by two relatively independent factors: the rotation angles of two fluorescence molecules and the binding rates between two fusion proteins. The FRET efficiency is primarily correlated with the distance between two fluorescence molecules, but this distance is not only influenced by the binding affinity between two fusion proteins but also by the rotation angle of two fusion proteins. It is possible that the fusion proteins have better binding affinity but the distance between fluorescence molecules is longer due to altered rotation angle due to mutation-derived conformational changes (Figure 56). Another concern is the varied binding rates or association rates between two fusion proteins. If the distance between two fluorescence molecules remains the same but the applied mutations (e.g. side-direct mutagenesis on expression plasmid according to RNA editing events)

reduce the binding rate of the two fusion proteins, the FRET efficiency is also attenuated (Figure 56). This adjustment with applying linker sequence may sacrifice the flexibility of the binding interaction between two fusion proteins because the two fusion proteins are not binding in the situation of free flowing. The linker sequence will bring a certain level of stiffness and potentially affect actual binding dynamics, altering the binding structure formed with the two fusion proteins. Additional disadvantage of linking coding sequences of two fusion proteins is the difficulty in studying the binding interactions among more than two proteins of interest (e.g., Spike, ACE2, and TMPRSS2) because adding more linker sequences may significantly influence structure formation. Therefore, this method is advantageous in studying a large number of mutations and their derived impacts on protein-protein interactions (PPI) between two proteins of interest. The data correlating mutations and PPI changes are ideal for training better deep learning models for PPI optimisation purposes, which PPI optimisation is fundamentally required for applications in various fields of research, medication, and industry[407].

Another potential method to minimise the influence of unequal transfection rates and expression rates of different fusion protein-expressing cassettes is to establish stable expression cell lines. Many excellent methods exist for establishing stable expression cell lines with multiple genomic inserts of expressing cassettes, particularly those based on CRISPR-Cas9 genome editing and lentivirus transfection[408]. By integrating the expressing cassettes of fusion proteins inside the cell genome and constitutively expressing them, the intracellular expression levels of different fusion proteins are not expected to vary significantly. Although different cell events may exhibit minor variations in intracellular fusion protein abundance due to varying actual cell status, the ratios between intracellular fusion proteins' abundance may remain very similar. Therefore, the single-origin cell events are expected to generate accurate and stable FRET signals, which eventually leads to highly accurate computed FRET efficiency, thus subsequently lead to PPI prediction model with excellent performance.

Although the FRET-based PPI detection assay can efficiently assess PPI changes caused by numerous protein mutations, further experimental verification is advisable. This is due to potential structural variations between wildtype proteins and fusion proteins tagged with fluorescence molecules. PPI changes observed in fusion-protein-based assays may differ from those in wildtype-protein assays. For instance, a pseudovirus experiment could serve as

additional verification. Following the identification of a mutation that significantly alters Spike-ACE2 binding, the effect of this mutation on PPIs can be confirmed by infecting ACE2-expressing cells with pseudovirus particles encapsulating Spike proteins[399]. The efficiency of pseudovirus entry into mutated ACE2-expressing cells can then serve as confirmation of the PPI changes. Furthermore, additional validation of PPI can be conducted by directly studying spatial structures of bound protein, using purified protein complexes followed by methods such as X-ray crystallography or cryogenic electron microscopy.

Appendix and References

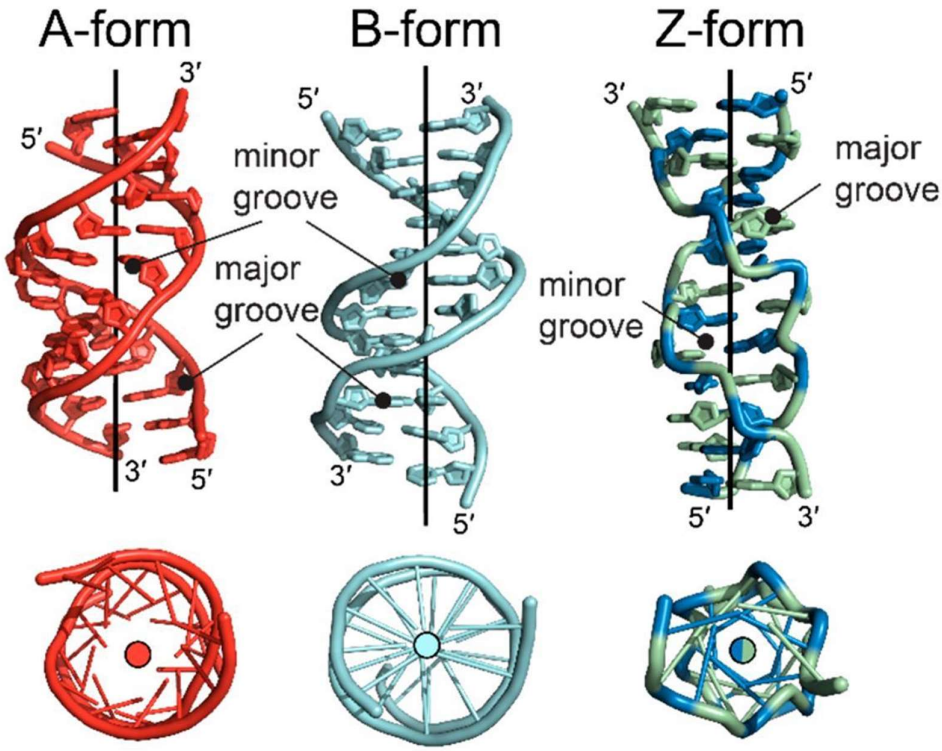
by Shuquan (Steve) Su

Appendix

Chapter 1 appendix

Gene	Other name	Reference or isoform	NCBI transcript accession ID	Title
ADAR1	ADAR1L ADAR1-p150	Reference	NM_001111.5	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 1, mRNA
ADAR1	-	Isoform	NM_015840.4	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 2, mRNA
ADAR1	-	Isoform	NM_015841.4	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 3, mRNA
ADAR1	-	Isoform	NM_001025107.3	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 4, mRNA
ADAR1	ADAR1S ΔDAR1-p110	Isoform	NM_001193495.2	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 5, mRNA
ADAR1	-	Isoform	NM_001365045.1	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 6, mRNA
ADAR1	-	Isoform	NM_001365046.1	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 7, mRNA
ADAR1	-	Isoform	NM_001365047.1	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 8, mRNA
ADAR1	-	Isoform	NM_001365048.1	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 9, mRNA
ADAR1	-	Isoform	NM_001365049.1	Homo sapiens adenosine deaminase RNA specific (ADAR), transcript variant 10, mRNA
ADAR2	-	Reference	NM_001112.4	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 1, mRNA
ADAR2	-	Isoform	NM_015833.4	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 2, mRNA
ADAR2	-	Isoform	NM_015834.4	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 3, mRNA
ADAR2	-	Isoform	NM_001160230.2	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 7, mRNA
ADAR2	-	Isoform	NM_001346687.2	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 9, mRNA
ADAR2	-	Isoform	NM_001346688.2	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 10, mRNA
ADAR2	-	Isoform	NM_001410722.1	Homo sapiens adenosine deaminase RNA specific B1 (ADARB1), transcript variant 12, mRNA
ADAR3	-	Reference	NM_018702.4	Homo sapiens adenosine deaminase RNA specific B2 (inactive) (ADARB2), mRNA
ADAT1	-	Reference	NM_012091.5	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 1, mRNA
ADAT1	-	Isoform	NM_001324444.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 2, mRNA
ADAT1	-	Isoform	NM_001324445.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 3, mRNA
ADAT1	-	Isoform	NM_001324446.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 4, mRNA
ADAT1	-	Isoform	NM_001324448.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 5, mRNA
ADAT1	-	Isoform	NM_001324449.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 6, mRNA
ADAT1	-	Isoform	NM_001324450.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 7, mRNA
ADAT1	-	Isoform	NM_001324451.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 8, mRNA
ADAT1	-	Isoform	NM_001324452.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 9, mRNA
ADAT1	-	Isoform	NM_001324453.2	Homo sapiens adenosine deaminase tRNA specific 1 (ADAT1), transcript variant 10, mRNA
ADAT2	-	Reference	NM_182503.3	Homo sapiens adenosine deaminase tRNA specific 2 (ADAT2), transcript variant 1, mRNA
ADAT2	-	Isoform	NM_001286259.2	Homo sapiens adenosine deaminase tRNA specific 2 (ADAT2), transcript variant 2, mRNA
ADAT3	-	Reference	NM_138422.4	Homo sapiens adenosine deaminase tRNA specific 3 (ADAT3), transcript variant 1, mRNA
ADAT3	-	Isoform	NM_001329533.2	Homo sapiens adenosine deaminase tRNA specific 3 (ADAT3), transcript variant 2, mRNA

Appendix 1. General NCBI information of ADAR and ADAT transcripts.



Appendix 2. Different conformational forms of DNA and RNA. This includes the spatial views of the nucleotides from different angles.

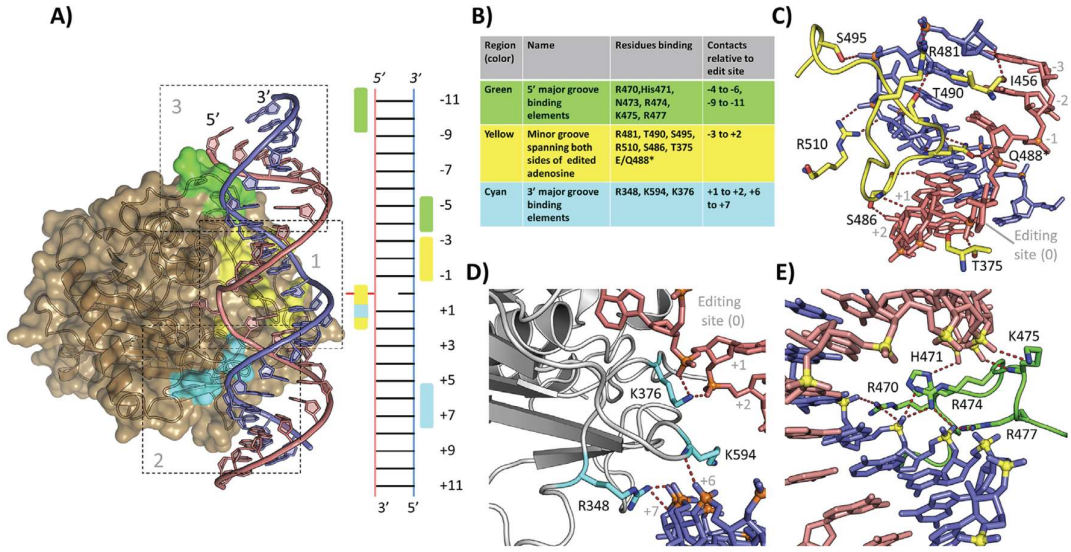
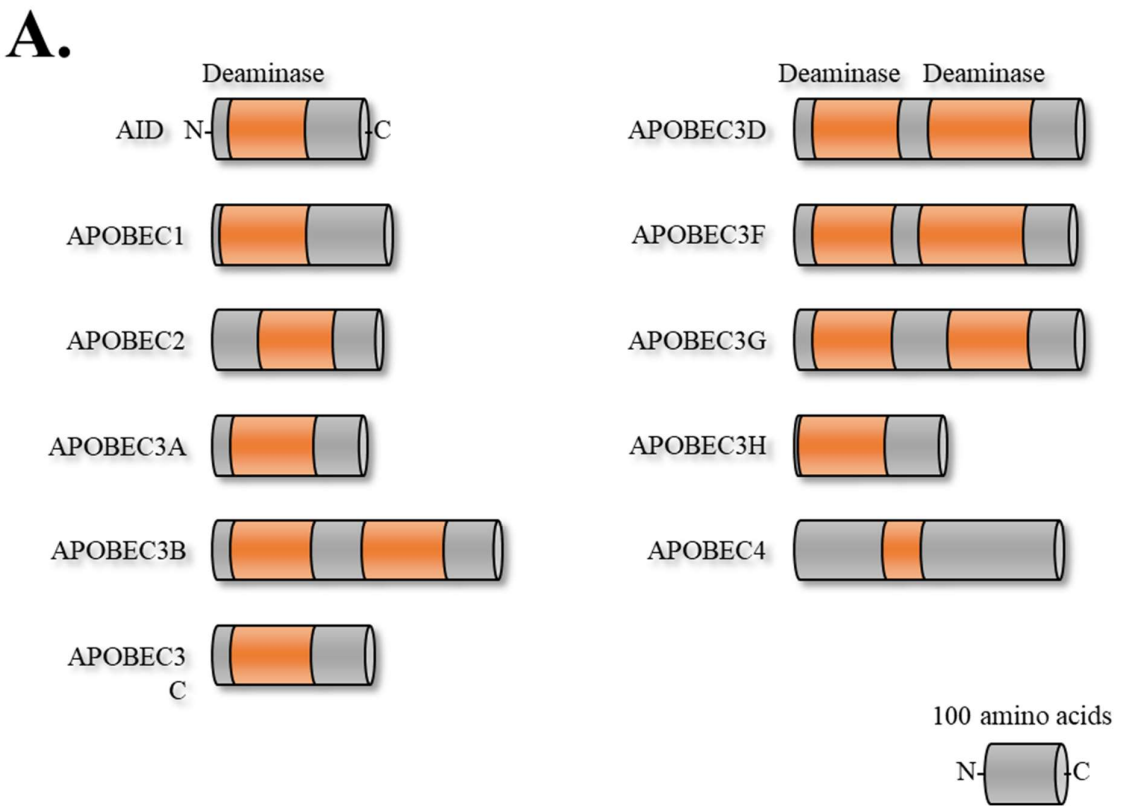
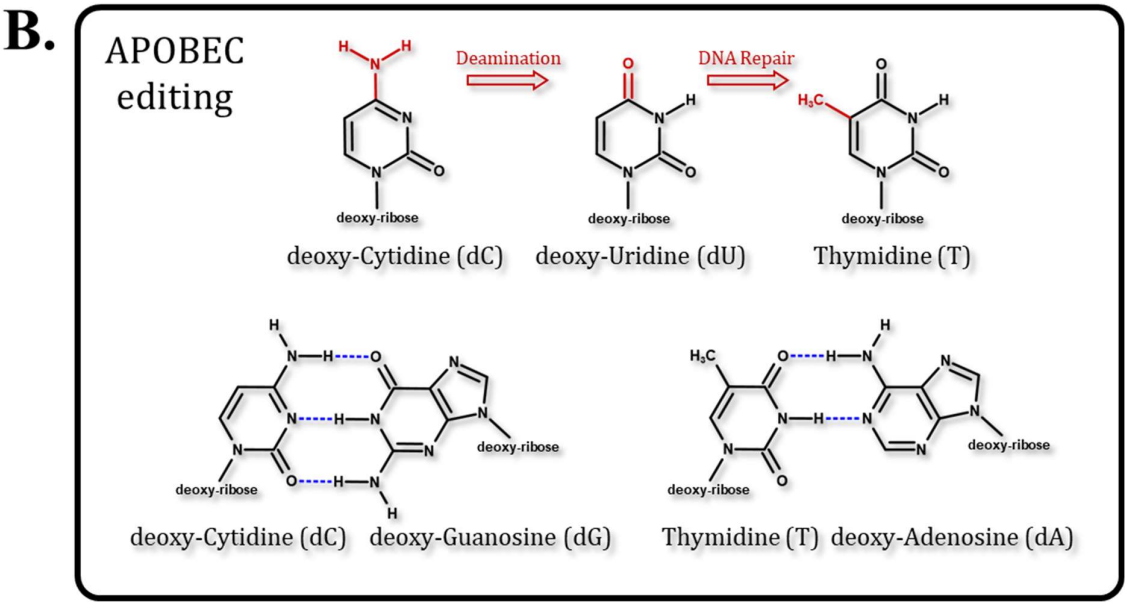


Figure 2. Interactions of hADAR2d with dsRNA beyond the active site. **A:** Overview of hADAR2d-RNA structure (5ED1, hADAR2d E488Q with Bdf2 derived 23-mer 8-azanebularane containing RNA) showing three main regions of contact. Edited strand in salmon, complementary stand in blue. Region 1 residues highlighted in yellow, region 2 in cyan, region 3 in green. Ladder diagram shows a secondary structure representation of the protein RNA contact interface (editing site shown in red flipped from the helix) **B:** Table summarizing the protein residues of each region and the RNA registers bound by each. **C:** Detail view of region 1. **D**: Detail view of region 2 (NOTE Fig. 2C shows 5ED1 hADAR2d E488Q). **E:** Detail view of region 3 (NOTE Fig. 2E is from 5ED2, crystal structure of hADAR2D E488Q with hGli1 derived 23mer 8-azanebularne containing RNA. This RNA extends 1 bp farther in the 5′ direction than the Bdf2 derived RNA); (5ED1 and 5ED2 [22]).

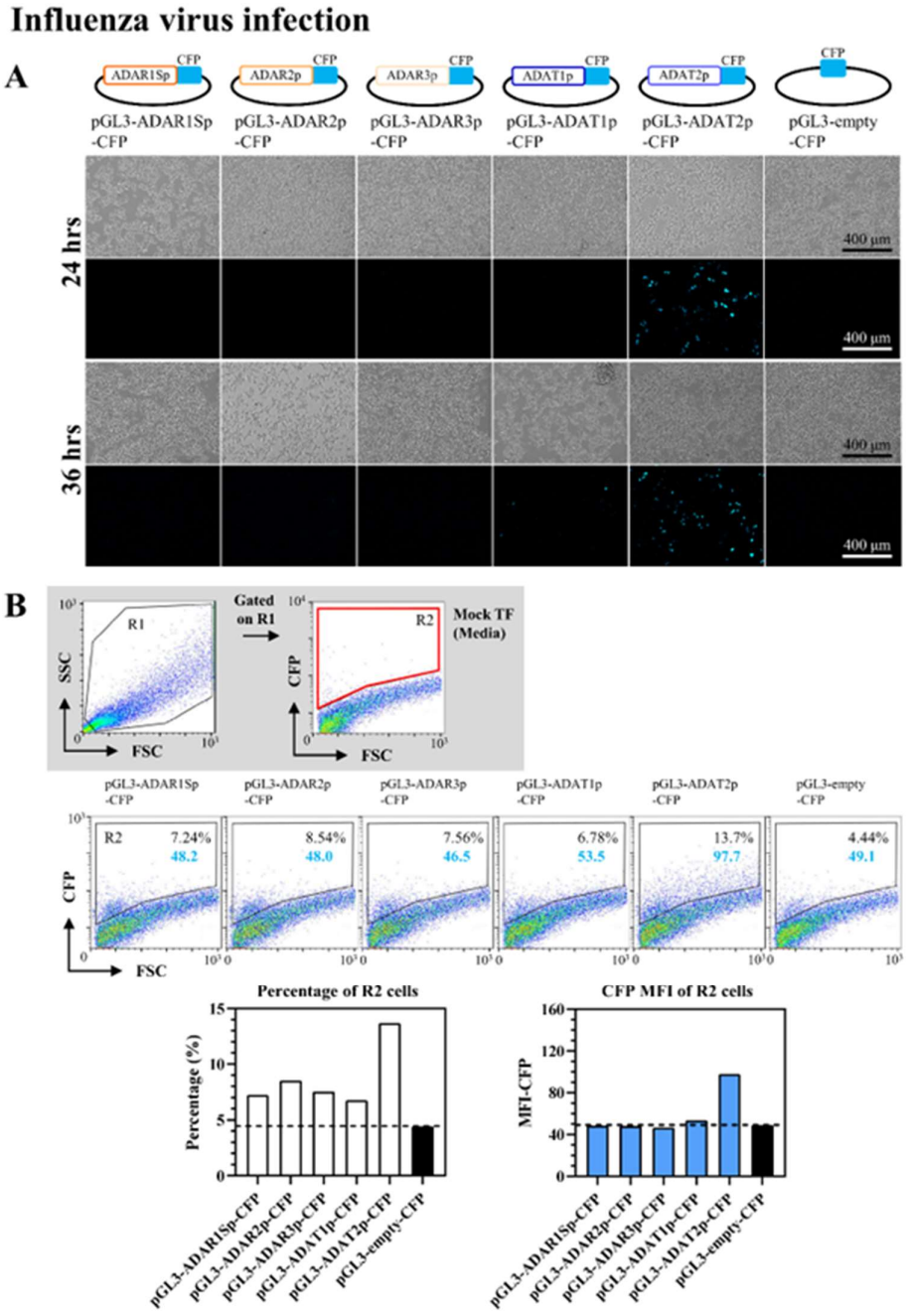
R481, respectively. Finally, the backbone carbonyl oxygens of T490 and I456 H-bond with the 2′-hydroxyl groups of the −2 and −3 positions of the complementary strand (Fig. 2A, yellow; Fig. 2C). The interactions in region 1 are likely to be critical for efficient deamination of most substrates.

Appendix 3. Figure of regional features of ADAR deaminase domain The regional features are Region 1, 2, and 3. This figure is original from figure 2 of publication titled ‘How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs’ with PMID 28217931[4].





Appendix 4. Information of APOBEC and dC-to-dU editing.



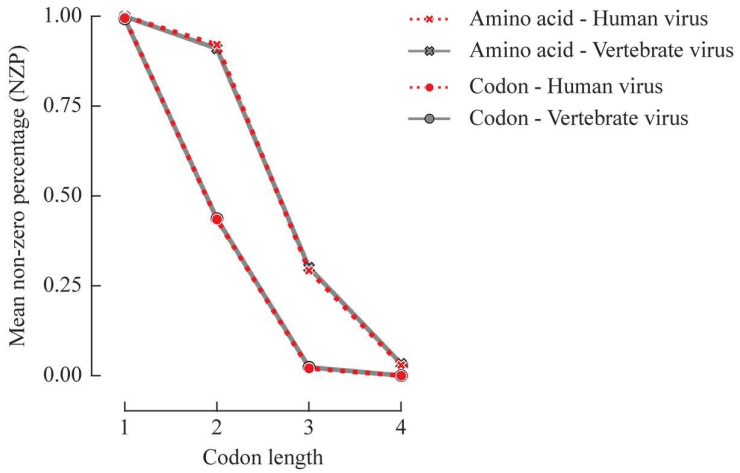
Appendix 5. Plasmid pGL3-ADARp-CFP (with 2kb ADAR promoter) and pGL3-ADATp-CFP (with 2kb ADAT promoter) transfection into Influenza virus infected HEK293. A. Fluorescence microscopy analysis of CFP emissions in different groups. B. Flow cytometry analysis of CFP emissions in different groups. Other infections with various viruses and activation of those promoters could be found in the previous thesis titled 'Defining Roles of Adenosine Deaminase Acting on RNA (ADAR) in Virus Infection'[3].

Appendix 6. Demonstration of data and annotations from NCBI accession ID.

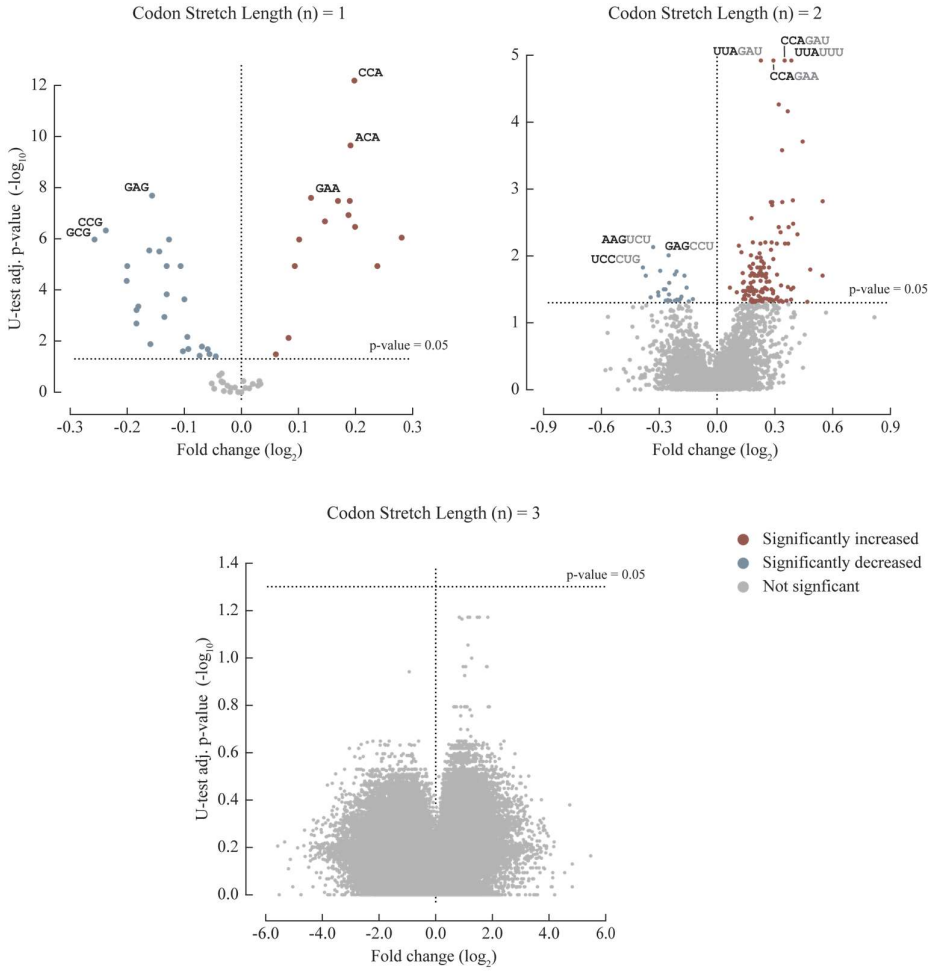
Amino acid / Codon	*Applied Group*	*Properties of Amino Acid / Codon*	*Ranking order (encoding)*	*Explanation*
Amino acid	AA%	Essential	Essential (2), Conditionally essential (1), Not-essential (0)	Whether the amino acid is essential or conditionally essential
Amino acid	AA%	Essential (General)	Essential (1), Not-essential (0)	Whether the amino acid is essential
Amino acid	AA%	Polar (General)	Polar (1) Not-polar (0)	Whether the amino acid is polar
Amino acid	AA%	Codon box	Six (4), Four (3), Three (2), Two (1), One (0)	Encoded codon counts
Amino acid	AA%	Codon box > 2	Yes (1), No (0)	Whether the amino acid has more than 2 codons
Amino acid	AA%	A present	Yes (1), No (0)	Whether the amino acid has A-present codons
Amino acid	AA%	U present	Yes (1), No (0)	Whether the amino acid has U-present codons
Amino acid	AA%	G present	Yes (1), No (0)	Whether the amino acid has G-present codons
Amino acid	AA%	C present	Yes (1), No (0)	Whether the amino acid has C-present codons
Amino acid	AA%	Total AU/GC-rich	GC-rich (2), Balanced (1), AU-rich (0)	Whether the amino acid has more GC-rich or AU-rich codons
Amino acid	AA%	ADAT relate	Yes (1), No (0)	Whether the amino acid is related to ADAT-related codons
Amino acid	AA%	Sum of predicted tRNA gene counts	-	Sum of predicted tRNA gene counts for all available codons
Amino acid	AA%	Sum of mature tRNA expression levels	-	Sum of mature tRNA expression levels for all available codons
Codon	Codon%, RSCU	Essential	Essential (2), Conditionally essential (1), Not-essential (0)	Whether the codon is encoding essential or conditionally essential amino acid
Codon	Codon%, RSCU	Essential (General)	Essential (1), Not-essential (0)	Whether the codon is encoding essential amino acid
Codon	Codon%, RSCU	Polar (General)	Polar (1) Not-polar (0)	Whether the codon is encoding polar amino acid
Codon	Codon%, RSCU	Codon box	Six (4), Four (3), Three (2), Two (1), One (0)	Codon counts of encoded amino acid
Codon	Codon%, RSCU	Codon box > 2	Yes (1), No (0)	Whether the codon is encoding the amino acid which has more than 2 codons
Codon	Codon%, RSCU	A present	Yes (1), No (0)	Whether the codon has A
Codon	Codon%, RSCU	U present	Yes (1), No (0)	Whether the codon has U
Codon	Codon%, RSCU	G present	Yes (1), No (0)	Whether the codon has G
Codon	Codon%, RSCU	C present	Yes (1), No (0)	Whether the codon has C
Codon	Codon%, RSCU	A/U present	Yes (1), No (0)	Whether the codon has A or U
Codon	Codon%, RSCU	G/C present	Yes (1), No (0)	Whether the codon has G or C
Codon	Codon%, RSCU	A count	-	A count in codon
Codon	Codon%, RSCU	U count	-	U count in codon
Codon	Codon%, RSCU	G count	-	G count in codon
Codon	Codon%, RSCU	C count	-	C count in codon
Codon	Codon%, RSCU	A/U count	-	Sum of A count and U count in codon
Codon	Codon%, RSCU	G/C count	-	Sum of G count and C count in codon
Codon	Codon%, RSCU	First nucleotide as A/U	Yes (1), No (0)	Whether the 1st nt is A or U in the codon
Codon	Codon%, RSCU	First nucleotide as G/C	Yes (1), No (0)	Whether the 1st nt is G or C in the codon
Codon	Codon%, RSCU	Second nucleotide as A/U	Yes (1), No (0)	Whether the 2nd nt is A or U in the codon
Codon	Codon%, RSCU	Second nucleotide as G/C	Yes (1), No (0)	Whether the 2nd nt is G or C in the codon
Codon	Codon%, RSCU	Third nucleotide as A/U	Yes (1), No (0)	Whether the 3rd nt is A or U in the codon
Codon	Codon%, RSCU	Third nucleotide as G/C	Yes (1), No (0)	Whether the 3rd nt is G or C in the codon

Codon	Codon%, RSCU	Total AU/GC-rich	GC-rich (2), Balanced (1), AU-rich (0)	Whether the codon is GC-rich or AU-rich
Codon	Codon%, RSCU	ADAT relate	Yes (1), No (0)	Whether the codon is related to ADAT editing
Codon	Codon%, RSCU	ADAT suppress/benefit	Benefit (2), Others (1), Suppress (0)	Whether the codon is supressed or benefited from ADAT editing in translation
Codon	Codon%, RSCU	Predicted tRNA gene counts	-	Predicted tRNA gene counts for the codon
Codon	Codon%, RSCU	Mature tRNA expression levels	-	Mature tRNA expression levels for the codon

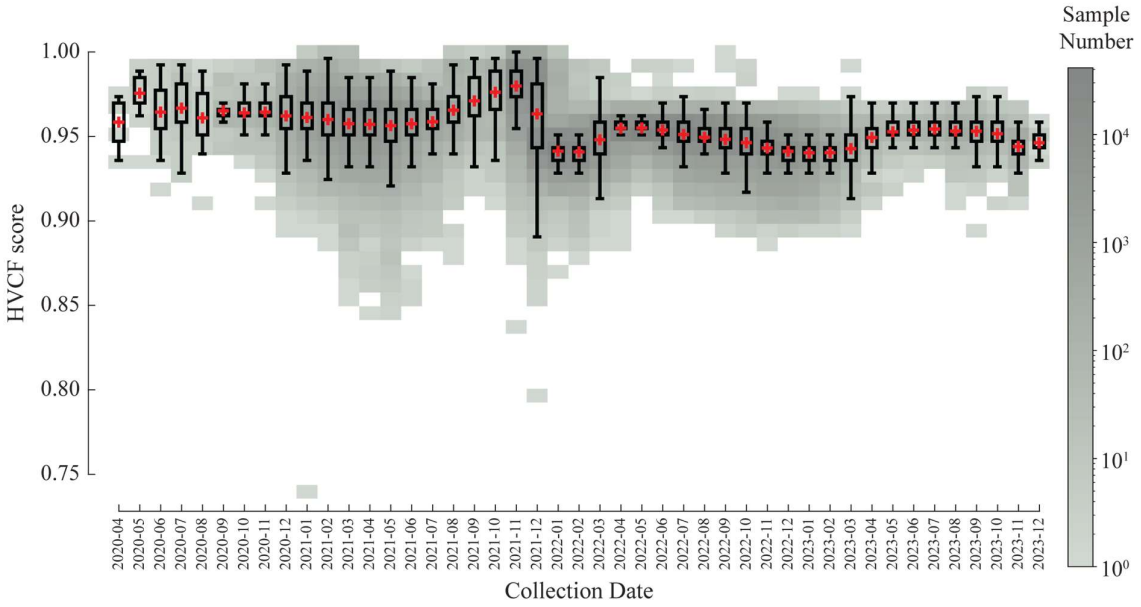
Appendix 7. Interested properties of amino acids (AA) and codons that are used to find correlations.



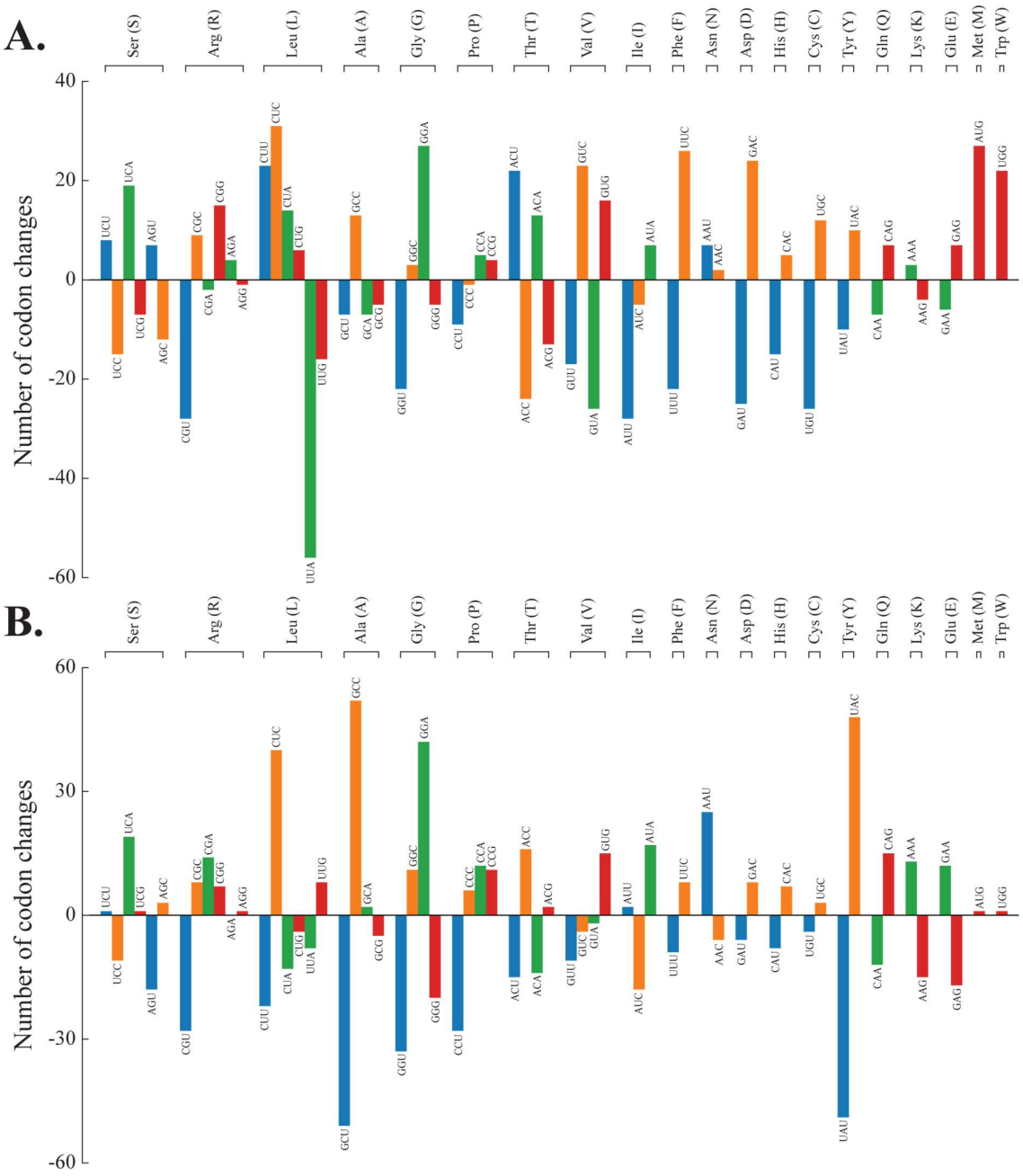
Appendix 8. Lineplot demonstrating non-zero percentages of all virus genome when codon stretch lengths (CSL) increase.



Appendix 9. Volcanoplot demonstrating transformed BH-adjusted p-values ($-\log_{10}$) from U-test and fold changes in RSMCU-n analysis.

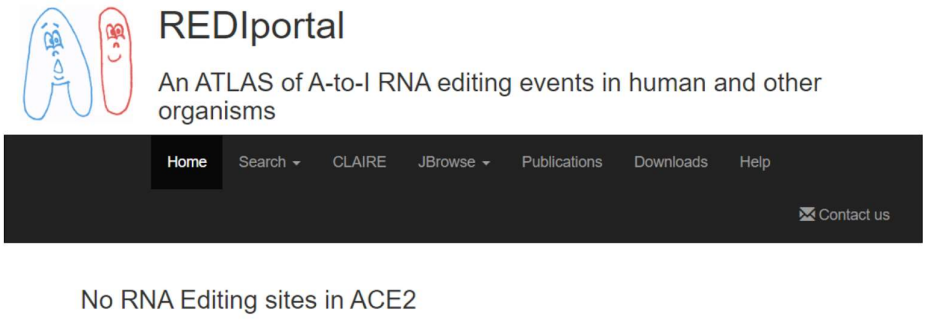


Appendix 10. Predicted HVCF scores of SARS-CoV-2 in USA across timeline from April 2020 to December 2023, all the data points are shown.

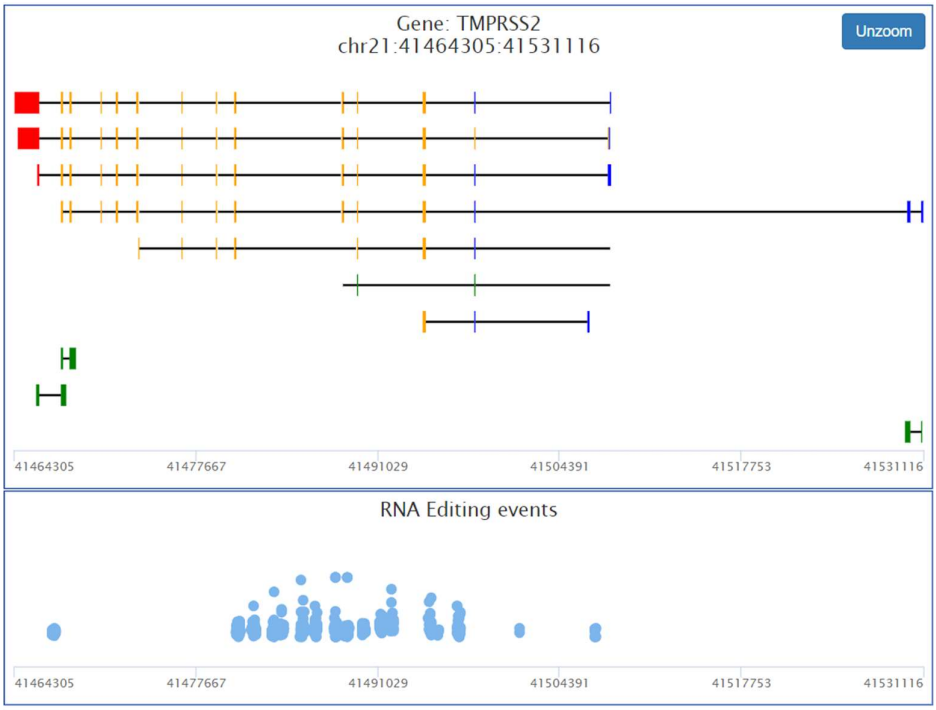


Appendix 11. Changes of codon numbers in predicted evolutionary path from Tylonycteris bat coronavirus HKU4 to SARS-CoV-2. (A) Low-to-High mutation path simulation. (B) Reversed High-to-Low mutation path simulation.

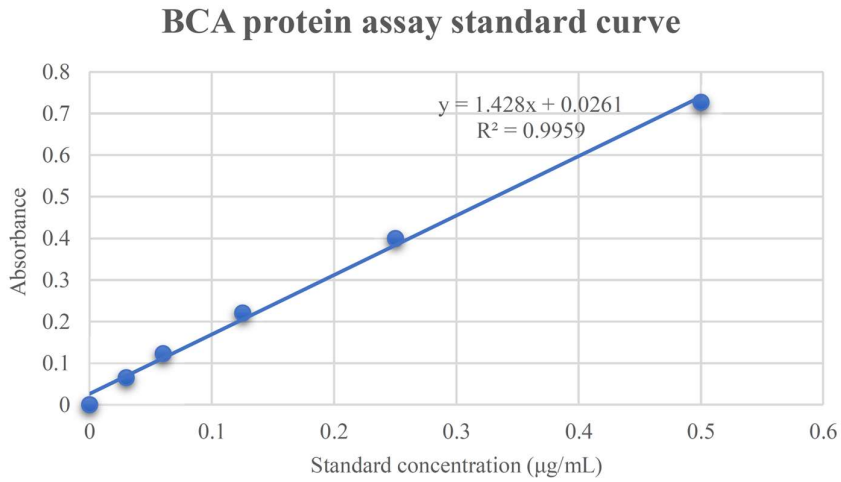
A.



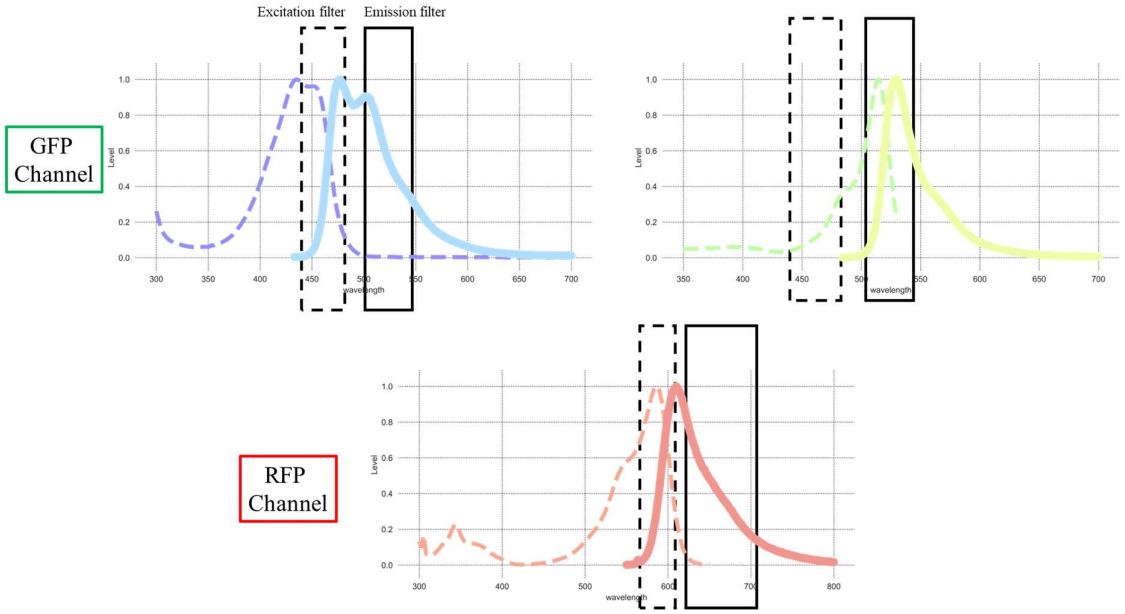
B.



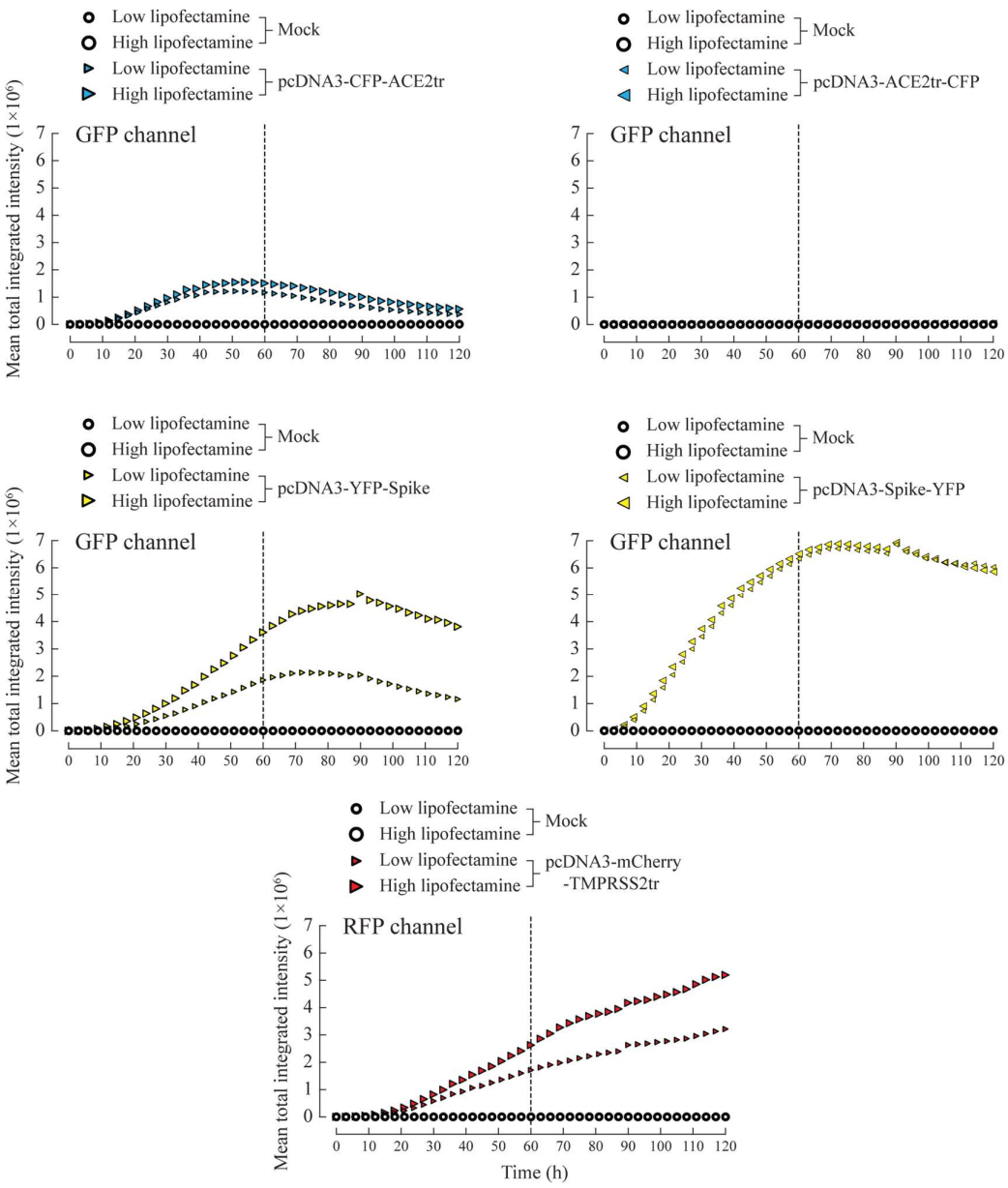
Appendix 12. RNA editing events in human genes ACE2 and TMRPSS2 from REDIportal database. (A) RNA editing events of ACE2 gene; (B) RNA editing events of TMPRSS2 gene.



Appendix 13. Standard curve of BCA assay to evaluate protein concentration before using in Western-blotting.



Appendix 14. Spectra of fluorescence reporters eCFP, eYFP, and mCherry, which signals could be detected in the corresponding channels in Incucyte S3 for live-cell imaging.



Appendix 15. Optimisation of lipotransfection of HEK293T cells with different plasmids.

Laser	Detectors	Filters	Note
488 nm (100 Mw)	PMT1		FSC
	PMT2	488/10	SSC
	PMT3	530/30	eYFP
	PMT4	575/26	
	PMT5	610/20	eYFP-mCherry FRET
	PMT6	670/30	mCherry
	PMT7	780/60	
405 nm (50 Mw)	PMT8	450/50	eCFP
	PMT9	525/50	eCFP-eYFP FRET
	PMT10	610/20	eCFP-mCherry FRET
	PMT11	670/30	
	PMT12	710/50	
	PMT13	780/60	
635 nm (40 Mw)	PMT14	670/30	
	PMT15	730/45	
	PMT16	780/60	

Appendix 17. Flow cytometry channel setting of BD LSR Fortessa X20, including the primary channels to detect different reporters' emission and FRET signals.

Sample #	Purpose	Transfected plasmids	FRET type	FRET donor-acceptor pair
1	Mock transfection	None	None	None
2	Empty vector control	pcDNA3	None	None
3	eCFP single color control	pcDNA3-eCFP-ACE2tr	None	None
4	eYFP single color control	pcDNA3-eYFP-Spike	None	None
5	eYFP single color control	pcDNA3-Spike-eYFP	None	None
6	mCherry single color control	pcDNA3-mCherry-TMPRSS2tr	None	None
7	PPI between eCFP-ACE2tr and eYFP-Spike	pcDNA3-eCFP-ACE2tr, pcDNA3-eYFP-Spike	C-Y FRET	eCFP→eYFP
8	PPI between eCFP-ACE2tr and Spike-eYFP	pcDNA3-eCFP-ACE2tr, pcDNA3-Spike-eYFP	C-Y FRET	eCFP→eYFP
9	PPI between eCFP-ACE2tr and mCherry-TMPRSS2tr	pcDNA3-eCFP-ACE2tr, pcDNA3-mCherry-TMPRSS2tr	C-R FRET	eCFP→mCherry
10	PPI between eYFP-Spike and mCherry-TMPRSS2tr	pcDNA3-eYFP-Spike, pcDNA3-mCherry-TMPRSS2tr	Y-R FRET	eYFP→mCherry
11	PPI between Spike-eYFP and mCherry-TMPRSS2tr	pcDNA3-Spike-eYFP, pcDNA3-mCherry-TMPRSS2tr	Y-R FRET	eYFP→mCherry
12	PPI between eCFP-ACE2tr, eYFP-Spike and mCherry-TMPRSS2tr	pcDNA3-eCFP-ACE2tr, pcDNA3-eYFP-Spike, pcDNA3-mCherry-TMPRSS2tr	C-Y-R FRET	eCFP→eYFP, eYFP→mCherry, eCFP→mCherry, eCFP→eYFP→mCherry
13	PPI between eCFP-ACE2tr, Spike-eYFP and mCherry-TMPRSS2tr	pcDNA3-eCFP-ACE2tr, pcDNA3-Spike-eYFP, pcDNA3-mCherry-TMPRSS2tr	C-Y-R FRET	eCFP→eYFP, eYFP→mCherry, eCFP→mCherry, eCFP→eYFP→mCherry

Appendix 16. Summary of plasmid combinations used in HEK293T transfections to study different FRET signals.

Appendix 18. Detailed cloning methods used in the chapter 4. The methods including (A.4.1.1) Molecular biology methods; (A.4.1.2) Microbiology methods.

A.4.1.1 Molecular biology methods

A.4.1.1.1. Genes synthesis. The coding sequences of human genes ADAR1L, ACE2, TMPRSS2 recorded from dominant transcripts were synthesised with Gene Universal company, and pre-cloned into pcDNA3 plasmids. The synthesised plasmids were received as desalted lyophilised powders, which were handled according to manufacturer's instructions. The dominant transcripts (or reference) information is listed in Appendix 19. Besides, the coding sequences of eCFP, eYFP were also synthesised with Gene Universal company, and pre-cloned into pcDNA3 with configurations of either 5'tag fusion or 3'tag fusion by Gene Universal company. The plasmids containing various synthesised sequences were received as desalted lyophilised powders, which were reconstituted into a volume of Nuclease-Free MilliQ water to achieve concentration of 100 ng/μL.

Gene	Species	Reference or isoform	NCBI transcript accession ID	Title
ACE2	Homo Sapiens	Reference	NM_021804.3	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 2, mRNA
ACE2	Homo Sapiens	Isoform	NM_001371415.1	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 1, mRNA
ACE2	Homo Sapiens	Isoform	NM_001386259.1	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 3, mRNA
ACE2	Homo Sapiens	Isoform	NM_001386260.1	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 4, mRNA
ACE2	Homo Sapiens	Isoform	NM_001388452.1	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 5, mRNA
ACE2	Homo Sapiens	Isoform	NM_001389402.1	Homo sapiens angiotensin converting enzyme 2 (ACE2), transcript variant 6, mRNA
TMPRSS2	Homo Sapiens	Reference	NM_005656.4	Homo sapiens transmembrane serine protease 2 (TMPRSS2), transcript variant 2, mRNA
TMPRSS2	Homo Sapiens	Isoform	NM_001135099.1	Homo sapiens transmembrane serine protease 2 (TMPRSS2), transcript variant 1, mRNA
TMPRSS2	Homo Sapiens	Isoform	NM_001382720.1	Homo sapiens transmembrane serine protease 2 (TMPRSS2), transcript variant 3, mRNA

Appendix 19. Summary of ACE2 and TRMPSS2 transcripts. The Reference transcript sequence was used in this thesis.

A.4.1.1.2. DNA oligos synthesis. The DNA oligos, including those ultilised as primers for either sub-cloning PCR or bacteria colony PCR, and those ultilised for DNA fragments annealing, were synthesised with Sigma-Aldrich. All the DNA oligos were received as desalted lyophilised powders, which were reconstituted into a volume of Nuclease-Free MilliQ water to achieve a 10× stock primers with concentration of 100 μM. Reconstituted primers were stored as 10× stock solution at 4 ℃ until using.

A.4.1.1.3. DNA fragments annealing. DNA oligos were designed and synthesised as parts of sub-cloning human TMPRSS2 gene and SARS-CoV-2 Spike gene respectively into

pcDNA3 over-expression plasmids. DNA oligos are prepared based on manufacturer's instructions and annealed together through temperature protocol in PCR shown below. The annealed DNA oligos were later digested and sub-cloned with other DNA fragments to reconstruct gene coding sequences in different forms.

A.4.1.1.4. Sub-cloning PCR. The DNA fragments containing fluorescence reporter protein coding sequences of eCFP, eYFP, mCherry, and mRFP were amplified by PCR using designed primers that flanked the 5' to the ATG initiating codon or first codon of those coding sequences, or 3' to the stop codon. Forward primers contained either a *NheI* 'G/CTAGC' restriction site or a *NotI* 'GC/GGCCGC' restriction site (depends on 5'-tag fusion or 3'-tag fusion manner), which were 5' to the ATG initiating codon plus several random nucleotides for subsequential restriction enzymes binding. Reverse primers contained either a *NotI* 'GC/GGCCGC' restriction site 3' to the last codon before the stop codon for stop-codon-exclusive 5'-tag fusion, or a *ApaI* 'GGGCC/C' restriction site which were 3' to the stop codon for stop-codon-inclusive 3'-tag fusion, which both have several random nucleotides for subsequential restriction enzymes binding.

To obtain DNA fragments of fluorescence reporter eCFP, eYFP, mCherry, and mRFP coding sequences for subsequent cloning into pcDNA3 overexpression plasmids, Polymerase Chain Reaction (PCR) amplifications were performed with Platinum Taq DNA Polymerase Kit (Thermofisher). PCR reaction mixture (50 μL) is prepared with 5 μL of 10× PCR buffer (no Mg^{2+}), 1.5 μL of $MgCl_2$ (50 mM), 1 μL of dNTPs (10 mM), 0.2 μL of Platinum Taq DNA Polymerase (10 units/μL), 1 μL of each forward and reverse primers (10 μM), and either 3 μL of plasmid template (100 ng/μL) with Nuclease-Free Water (Ambion) adding up to 50 μL. The PCR reaction was performed using ProFlex PCR System (Thermofisher) according to the following conditions: 1 cycle of 94 ℃ 2 mins, then 25 cycles of 94℃ for 30 sec (denaturation), 55 ℃ for 30 sec (annealing), and 72℃ for 1 min (elongation), followed by a final single cycle of 72 ℃ for 3 mins. The PCR conditions would be changed in annealing temperature and elongation cycle for optimisation.

A.4.1.1.5. Restriction enzyme digestion. The restriction enzyme digestions were used in multiple scenarios including verify plasmids from colonies and cutting fragments for sub-cloning. The restriction enzyme digestions for screening colonies and verifying were performed in a digestion mixture (total 10 μL) containing 2 μL of 10× digestion buffer (NEB), 1 μL of plasmid DNA solution (100 ng/μL) and 0.2 μL of each restriction enzyme (10 units/μL,

NEB), which Nuclease-Free Water is added up until 10 µL. The digestion was performed using pre-set temperature heat block (VWR International) according to the following conditions: 37 ℃ for 2 hours (digestion), then 80 ℃ for 20 mins (heat-inactivation).

The restriction enzyme digestion for sub-cloning (re-collecting digested fragments) was performed in a digestion mixture (50 µL) containing 5 µL of 10× digestion buffer (NEB), 20 µL of DNA (100 ng/µL) and 1 µL of each restriction enzyme (10 units/µL, NEB). Antarctic Phosphatase (AnP) was used to remove phosphates on 5' or 3' ends of fragments for preventing self-ligations, which additional 5 µL of 10× AnP buffer (NEB) and 1 µL of AnP (5 units/µL, NEB) was added into the reaction mixture. Nuclease-Free Water was added up to 10 µL. The digestion was performed using pre-set temperature heat block (VWR International) according to the following conditions: 37 ℃ for 2 hours (digestion), then 80 ℃ for 20 mins (heat-inactivation).

A.4.1.1.6. DNA quantification. Concentrations of nucleic acid (plasmid DNA or other DNA fragments) were measured by Nanodrop One Spectrophotometer (Thermofisher) according to the manufacturer's instructions. Briefly, 1 µL of the plasmid DNA or other nucleic acid were placed on the instrument after using 1 µL of Nuclease-Free MilliQ water as a blank. The purity of sample DNA/RNA was assessed by A260/280 and A260/230 according to manufacturer's instructions.

A.4.1.1.7. DNA gel electrophoresis. Unless stated, a 1 % Tris-borate-EDTA (TBE) agarose gels were made by 1 g of agarose (Bioline) and 100 mL of TBE through heating of microwave until completely molten. In some circumstances, a 2% agarose gel was prepared by melting 2 g agarose in 100 mL of TBE. 4 µL of GelRed stains (Biotium) was added into 100 mL of melted agarose gel mixture before loading on agarose gel casting tray, assembled with plastic gel well combs. The gels were allowed to solidify at RT for 1 hour before assembling in MGU-252T Horizontal Mini-Gel Systems (VWR International).

DNA samples and 1 kb DNA ladder (Invitrogen) were mixed with 6× Gel Loading Buffer Purple (NEB) according to manufacturer's instructions. DNA electrophoreses were performed at 80 V using PowerPac Basic Power Supply (Bio-Rad) for approximately 1.5 hours or until the loading buffer dye indicator reached approximately 4/5 of the total gel length. The electrophoresed DNA was visualised under 302nm wavelength light using InGenius3 UV

trans-illuminator (Syngene), and images of gels were acquired with GeneSys software (version 1.5.0.0).

A.4.1.1.8. DNA gel purification. To purify DNA fragments from either PCR amplification or restriction enzyme digestions, the DNA is re-extracted from agarose gels after electrophoresis. For cutting out the gel pieces containing DNA bands, DNA bands were visualised under Kodak Electrophoresis Documentation and Analysis System 290. For extracting the gels containing desired DNA bands, a thin slice of the gel was cut with a sterile scalpel blade and placed the gel slice into a sterile 1.6 mL Eppendorf tube. DNA was extracted from the agarose gel using the PureLink Quick Gel Extraction Kit (Invitrogen) according to manufacturer's instructions. Briefly, the gel slices were completely dissolved into Gel Solubilisation Buffer L3 with 1:3 ratio (ex. 400 mg gel slice into 1.2 mL L3 buffer) at 50 ℃. The mixture was later transferred into a Quick Gel Extraction Column inside a Wash Tube and the column was centrifuged at 12,000× g for 1 min (flow-through solution was discarded) in an Eppendorf 5415D micro-centrifuge. The column was washed by adding 500 µL of Wash Buffer (W1) and centrifuging at 12,000× g for 1 min (flow-through solution was discarded). Then, the column was centrifuged for another 12,000× g for 2 min and the DNA was eluted by assembling the column into a collection tube, adding 50 µL of Nuclease-Free Water (Ambion) and centrifuging at 12,000× g for 1 min.

A.4.1.1.9. DNA ligation. DNA ligations were performed with reaction mixtures containing 1 µL of 10× T4 DNA ligase buffer, 1 µL of T4 DNA ligase (NEB), vector/insert fragments and Nuclease-Free Water (Ambion) added up to 10 µL. Mostly, vector-insert ratio was 1:1 (1:3 or 3:1 were used in some circumstances) and total volume of vector and insert solution together was smaller than 5 µL for avoiding influences of potential salts in extracted vector or insert solutions. The reaction mixtures were incubated at 4 ℃ overnight. Alternatively, the reaction mixtures were mounted on a thin layer of ice at room temperature overnight. The incubating temperature slowly increased from 0 ℃ to 25 ℃ during the ice melting process (16 ℃ at certain time point). The reaction mixtures were incubated at 65 ℃ for 10 mins after ligation (heat-inactivation). To increase reaction successful rate of multi-fragments ligations (3-fragments or 4-fragments ligations), which all the fragments were generated by either restriction enzyme digestion or DNA fragments annealing, 2 µL of T4 DNA ligase (NEB) was used instead of 1 µL of T4 DNA ligase.

TA Cloning Kit Dual Promoter (pCRII, Thermofisher) was used for preserving PCR products by T-A cloning ligations according to manufacturer's instruction. The reaction mixture contained 2 μL of 5× ExpressLink T4 DNA Ligase Buffer, 2 μL of pCRII vector (25 ng/μL), 1 μL of ExpressLink T4 DNA Ligase (5 units/μL), and approximately 10 ng of fresh PCR product, which Nuclease-Free Water (Ambion) added up to totally 10 μL. The positive and negative control ligation reactions were prepared as manufacturer's protocol.

A.4.1.2 Microbiology methods

A.4.1.2.1. Bacteria culture. Luria-Bertani (LB) Broth was prepared with LB Broth pre-mixed powder (Difco) according to manufacturer's instructions. 40 μL of 1000× Ampicillin stock (100 mg/mL, Sigma) or 200 μL of 200× Kanamycin stock (50 mg/mL, Sigma or 40 μL 1000× Spectinomycin (100 mg/mL, Sigma) were dissolved into 40 mL of LB if necessary.

LB agar plates were prepared with LB Broth pre-mixed powder (Difco) and Granulated Agar powder (Difco) according to manufacturer's instructions. 40 μL of 1000× Ampicillin stock (100 mg/mL, Sigma) or 200 μL of 200× Kanamycin stock (50 mg/mL, Sigma) or 40 μL 1000× Spectinomycin (100 mg/mL, Sigma) were added into 40 mL of 50 ℃ LB agar solution before pouring the plates. For making X-Gal/IPTG plates, for T-A cloning, 100 μL of 100 mM IPTG (Bioline) and 20 μL of 50 mg/mL X-Gal (Progen) were mixed and spread onto the antibiotic-containing LB agar plates and the LB agar plates were incubated at 37℃ for 30 mins until the IPTG/X-Gal were completely absorbed by the LB agar.

Desired colonies on LB agar plates were picked with sterile pipette tips and those tips were immersed into LB broth (with or without antibiotics). The bacterial LB broth was incubated at 37 ℃ for 16 hours before subsequential plasmids extraction.

A.4.1.2.2. Heat-shock transformation. Heat-shock competent DH5α *E. coli* was used for heat-shock transformation in purpose of amplifying plasmids in *E. coli*. DH5α super-competent *E. coli* was kindly provided by colleague Dr Zhongran Ni (UTS). The frozen *E. coli* (50 μL) was thawed on ice then the DNA ligation products (usually 10 μL) were added directly into the *E. coli* culture. The *E. coli*/DNA mixtures were incubated on ice for 20 mins before exposing

the cells to a heat shock treatment of 42 ℃ for 50 sec using a pre-set temperature heat block. The *E. coli*/DNA mixtures were returned to ice for another 2 mins before mixing with 950 µL of antibiotics-free LB broth and incubating at 37 ℃ for 1 hour. The transformed cells cultures were later centrifuged at 3000× g for 5 mins then the clear supernatant was discarded (and decontaminated in fresh 0.1% bleach solution). The *E. coli* cell pellet was resuspended in 100 µL of LB broth and the entire resuspended culture was spread onto LB agar plates containing antibiotics according to the antibiotic resistance gene of the plasmids. The newly plated *E. coli* LB agar plates were incubated at 37 ℃ overnight for approximately 16 hours or until bacterial colonies were large enough to see for counting and for colony picking to a master plate containing the same antibiotics as master plates.

A.4.1.2.3. Plasmid extractions. Bacterial plasmids prepared from transformed *E. coli* cultures, were extracted using ISOLATE-II Plasmid Mini Kit (Bioline) according to manufacturers' protocol. The extracted plasmids could be used for subsequential experiments such as colony screening, molecular cloning and cell culture transfection (i.e. Lipofectamine transfection). Here is a brief Miniprep protocol. 5 mL transformed *E. coli* cultures were pelleted by centrifuging at 2,300 ×g for 5 mins, which the bacterial pellet was resuspended with 250 µL of Resuspension Buffer P1 after LB supernatant was removed. 250 µL of Lysis Buffer P2 was added to the resuspended bacteria, and the bacteria was lysed for 5 mins at room temperature. The mixture was neutralised and precipitated by adding 300 µL of Neutralization Buffer P3, which the insoluble was precipitated by centrifuging at 11,000 ×g for 5 mins. The supernatant was transferred into the ISOLATE II Plasmid Mini Spin Column, and the plasmid DNA was bound on the silico membrane of the column by centrifuging at 11,000 ×g for 1 min. 600 µL of Wash Buffer PW2 was added to the column and removed by centrifuging at 11,000 ×g for 1 min. After drying the column by centrifuging at 11,000 ×g for 2 mins, the plasmid was eluted with 20 µL of Nuclease-Free MilliQ water by centrifuging at 11,000 ×g for 1 min. The eluted plasmid was often diluted into 100 ng/µL later for subsequential experiment.

The sequences of extracted plasmids were verified with Sanger sequencing service provided by Macrogen (South Korea).

A.4.1.2.4. Colony PCR. The transformations of multi-fragments ligation in competent *E. coli* will lead to significantly lower rates in getting desired ligation products. Thus, colony PCR

with Taq DNA Polymerase (NEB) to screen large amount of transformed *E. coli* colonies. PCR reaction mixture (10 μL) is prepared with 1 μL of 10× Standard Taq Reaction Buffer, 0.2 μL of dNTPs (10 mM), 0.2 μL of each forward and reverse primers (10 μM), and 0.1 μL of Taq DNA Polymerase (5 units/μL) with Nuclease-Free Water (Ambion) adding up to 10 μL. After the mixture is prepared, the *E. coli* colony was gently inoculated and immersed in the reaction mixture. The PCR reaction was performed using ProFlex PCR System (Thermofisher) according to the following conditions: 1 cycle of 95 ℃ 6 mins, then 25 cycles of 95℃ for 20 sec (denaturation), 62 ℃ for 30 sec (annealing), and 68 ℃ for 4.5 mins (elongation), followed by a final single cycle of 68 ℃ for 7 mins. The PCR conditions would be changed in annealing temperature and elongation cycle for optimisation. The PCR product will be examined by DNA gel electrophoresis to allocate correct colonies.

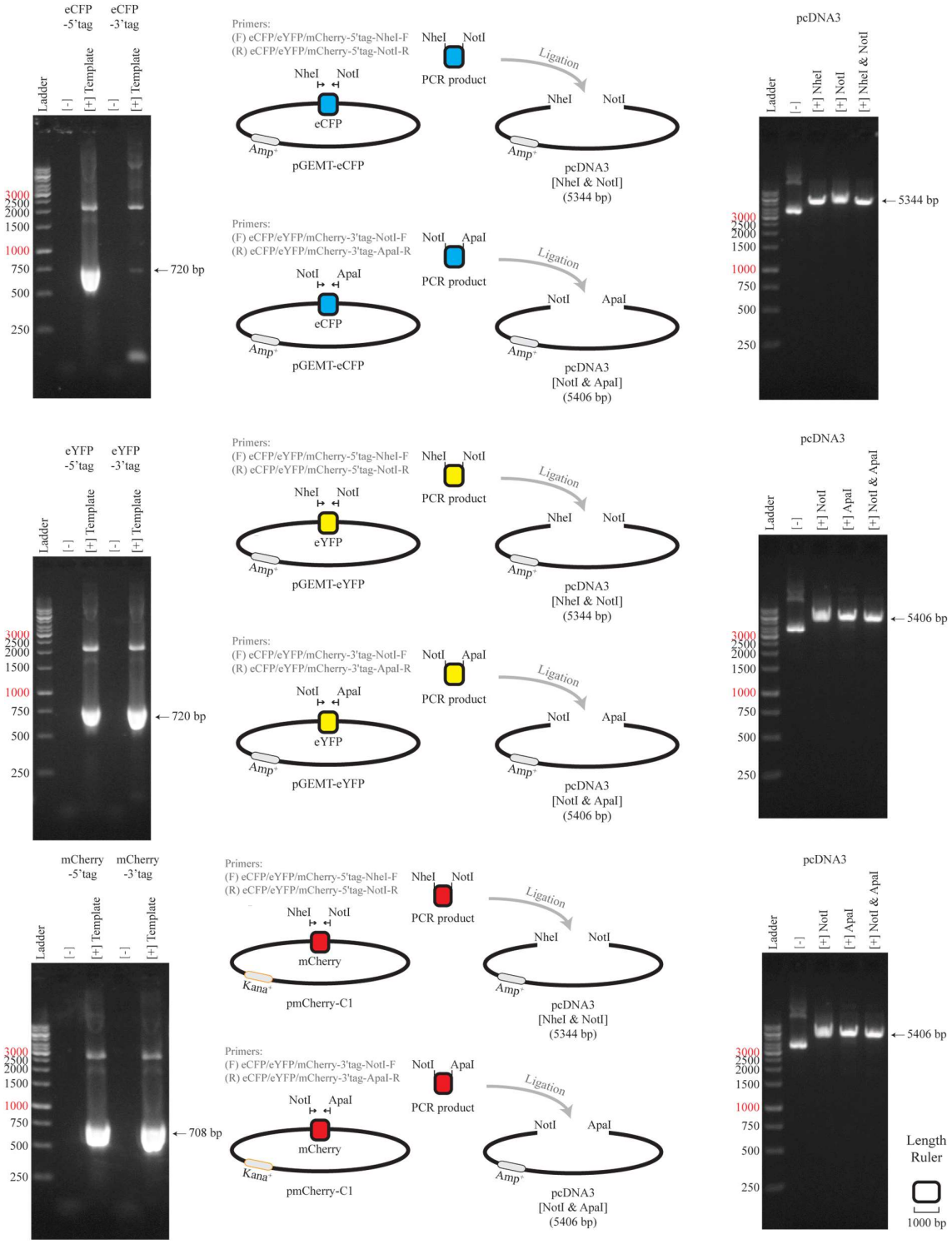
Appendix 20. Step-by-step detailed description of molecular cloning constructions of plasmids of FRET-based PPI detection assay. Please also see method details in Appendix 18.

A.4.1.3. Molecular cloning in constructions of intracellular FRET assay

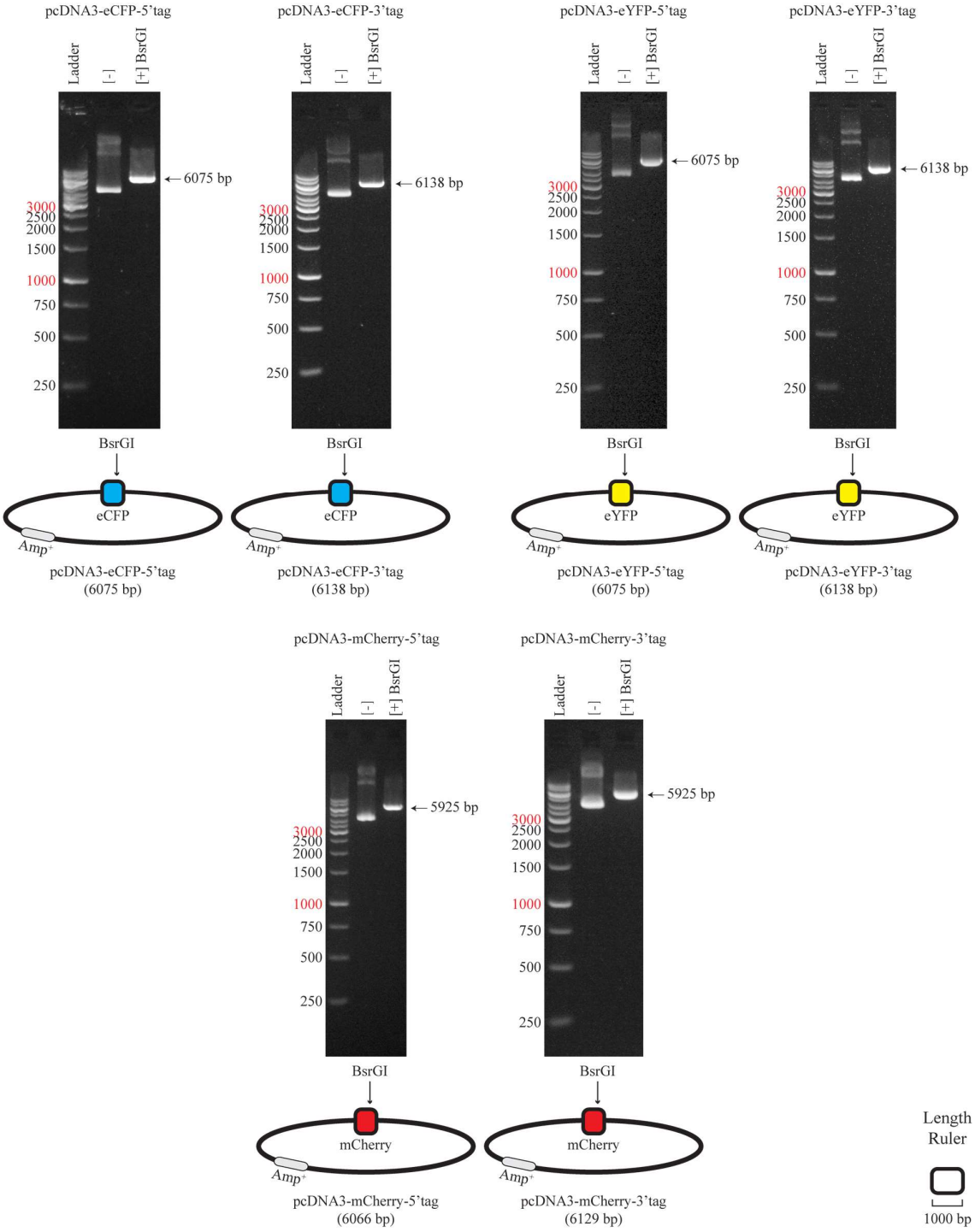
A.4.1.3.1. Establish empty fluorescence-fusion plasmids

To establish intracellular FRET-based protein-protein interaction (PPI) examining system, empty overexpression plasmids (pcDNA3 vector) containing fluorescence reporter tag coding sequences adapting to either 5' fusion or 3' fusion were first established before inserting other genes of interest. In this project, eCFP, eYFP, mCherry were chosen as mainly used fluorescence reporters which mRFP served as alternative backup of mCherry. The eCFP-eYFP-mRFP three-way FRET assay was validated in colleague Dr Zhongran Ni's Ph.D. thesis[390]. Fluorescent protein mCherry has almost identical spectrum profile compared to mRFP but more stable molecular structure[396], and it had approved its functioning as a FRET donor[397].

The coding sequences of eCFP, eYFP, and mCherry were first PCR amplified with desired primers, and the PCR products were restrictions enzyme digested, which coding sequences for 5' tag fusion (reporter-gene fusion) were digested with *NheI* and *NotI*, and coding sequences for 3' tag fusion (gene-reporter fusion) were digested with *NotI* and *ApaI*. The digestion products were later ligated into empty pcDNA3 plasmids, which were previously digested with either *NheI/NotI* or *NotI/ApaI* combinations accordingly. Multiple florescence reporter CDS containing pcDNA3 plasmids in either 5' tag fusion or 3' tag fusion was created successfully (Appendix 21).



Appendix 21. Construction of intermediate plasmids containing only fluorescence reporter coding sequences for subsequential cloning of fusion protein plasmids. The fluorescence reporters include eCFP, eYFP, and mCherry, having both 5'- and 3'- fusion plasmids.



Appendix 22. Verification of intermediate plasmids containing only fluorescence reporter coding sequences via restriction enzyme digestions.

Unfortunately, only pcDNA3-mCherry-5'tag, and pcDNA3-mCherry-3'tag have the 100 % correct insert sequences (data not shown), where 2 non-synonymous mutations were identified in both pcDNA3-eCFP-5'tag and pcDNA3-eCFP-3'tag, and 1 non-synonymous mutation was identified in both pcDNA3-eYFP-5'tag and pcDNA3-eYFP-3'tag in Sanger sequencing results (data not shown). Those sequencing identified mutations were consistently identical in different DNA samples from different colonies, suggesting the mutations were derived from the templates instead of PCR errors. Thus, the plasmids pcDNA3-eCFP-5'tag, pcDNA3-eYFP-5'tag, pcDNA3-eCFP-3'tag, and pcDNA3-eYFP-3'tag were synthesised with commercial gene synthesis service, and the sequencing reports show all the plasmids had 100 % correct insert sequences (data not shown). All the information of the empty fluorescence-fusion plasmids were listed below, and they were verified again with restriction enzyme digestions (Appendix 22). Besides, pcDNA3-mRFP-5'tag, and pcDNA3-mRFP-3'tag are also constructed for alternative backup of pcDNA3-mCherry-5'tag, and pcDNA3-mCherry-3'tag (Data not shown).

A.4.1.3.2. Prepare fragments for multi-way ligation

To study the protein-protein interaction of SARS-CoV-2 cell entry mechanisms, and also identify PPI changes caused by RNA-editing-derived mutations, ACE2, TMRPSS, and Spike were inserted into fluorescence reporter CDS containing pcDNA3 plasmids to create plasmids expressing gene-reporter fusion protein.

The full-length coding sequences of human genes ACE2 and TMPRSS2 were synthesised with commercial gene synthesis service, which were received as pcDNA3-ACE2 and pcDNA3-TMPRSS2 plasmids. The full-length SARS-CoV-2 viral Spike CDS was obtained from the purchased plasmid pDONR223-Spike (Addgene ID #149329). Unlike Spike protein, ACE2 and TMPRSS2 are membrane proteins, which will be translocated to membrane once expressed intracellularly. In this case, the intracellularly expressed ACE2 and TMPRSS2 will be translocated to the membrane while the expressed Spike protein would remain in cytoplasm, which will affect the protein-protein interactions among them. To make expressed ACE2 and TMRPSS2 remained in cytoplasm for increasing protein-protein association rate, transmembrane domains, and extracellular domains of both ACE2 and TMRPSS2 were removed with molecular cloning strategies, which only cytoplasmic domains remained. These truncated versions of ACE2 and TMRPSS2 will be expected to stay in cytoplasm having more

chances to bind with Spike, which will hopefully generate detectable FRET signals. The truncated ACE2 with amino acids 0~740 of wild-type ACE2 (total 805 amino acids) was noted as ACE2tr, and the truncated TMPRSS2 with amino acids 106~492 of wild-type TMPRSS2 (total 492 amino acids) was noted as TMPRSS2tr.

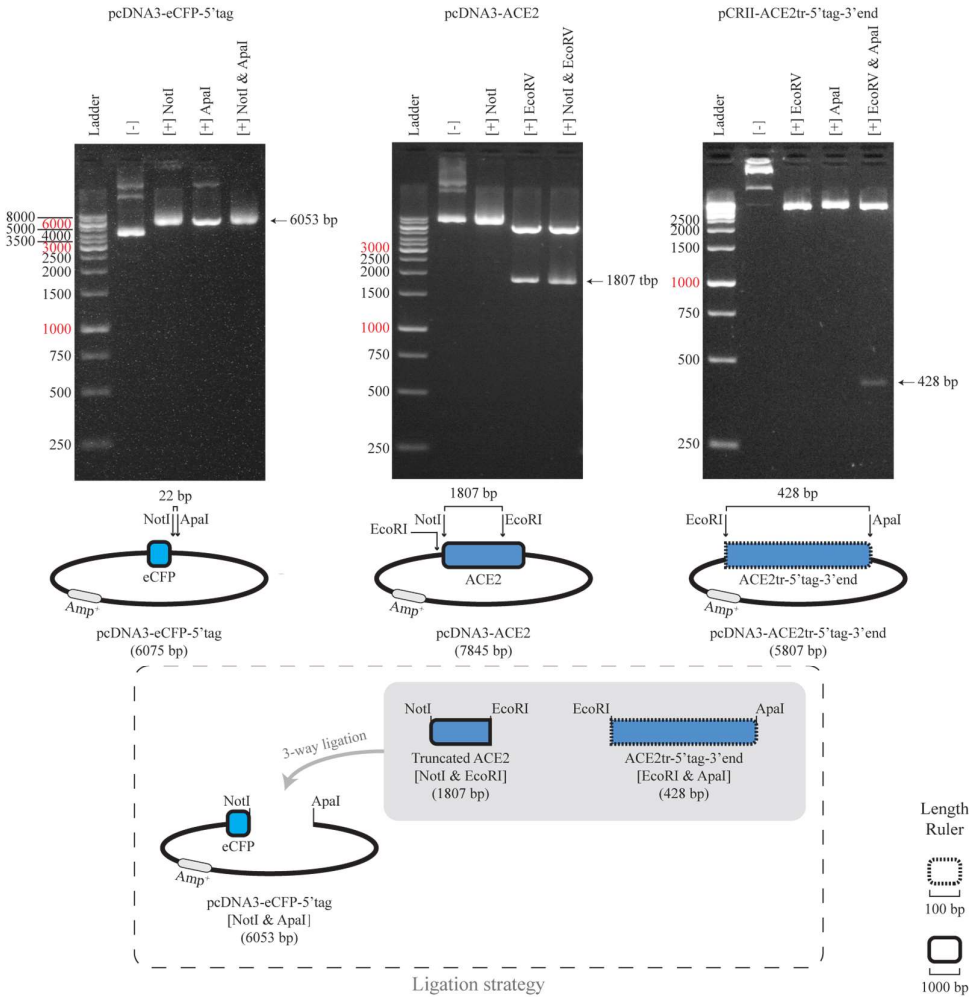
To reduce PCR-derived mutations, small portions of 5'end and 3'end of ACE2, Spike, and TMPRSS2 were obtained by either PCR amplified with designed primers or DNA oligo ligations followed with restriction digestions before using in multi-fragment ligations to re-construct ACE2tr, TMPRSS2tr, and full-length Spike adapting to sub-cloning of either 5' tag fusion or 3' tag fusion. The PCR products were first cloned into pCRII plasmid by T-A cloning, which were sequenced to ensure there is no PCR-derived mutations (data not shown). In summary, plasmid pCRII-ACE2tr-3'tag-5'end was constructed for pcDNA3-ACE2tr-CFP (3'tag fusion) sub-cloning. Plasmid pCRII-Spike-5'tag-3'end was constructed for pcDNA3-YFP-Spike (5'tag fusion) sub-cloning, while plasmid pCRII-Spike-3'tag-3'end was constructed for pcDNA3-Spike-YFP (3'tag fusion) sub-cloning.

Besides, aligned DNA oligos were also used to obtain fragments for sub-cloning. In summary, fragment Spike-5'tag-5'end was aligned for pcDNA3-eYFP-Spike (5'tag fusion) sub-cloning, while fragment Spike-3'tag-5'end was aligned for pcDNA3-Spike-eYFP (3'tag fusion) sub-cloning. Fragments TMPRSS2tr-5'tag-5'end and TMPRSS2tr-5'tag-3'end was aligned for pcDNA3-mCherry-TMPRSS2tr (5'tag fusion) sub-cloning, while fragments TMPRSS2tr-3'tag-5'end and TMPRSS2tr-3'tag-3'end was aligned for pcDNA3-TMPRSS2tr-mCherry (3'tag fusion) sub-cloning. In addition, some plasmids were synthesised for later multi-way ligation also which were longer and easier to have potential PCR errors, which sequences were 100 % correct from sequencing (data not shown). Plasmid pcDNA3-ACE2tr-5'tag-3'end was constructed for pcDNA3-eCFP-ACE2tr (5'tag fusion) sub-cloning, while plasmid pcDNA3-ACE2tr-3'tag-3'end was constructed for pcDNA3-ACE2tr-eCFP (3'tag fusion) sub-cloning.

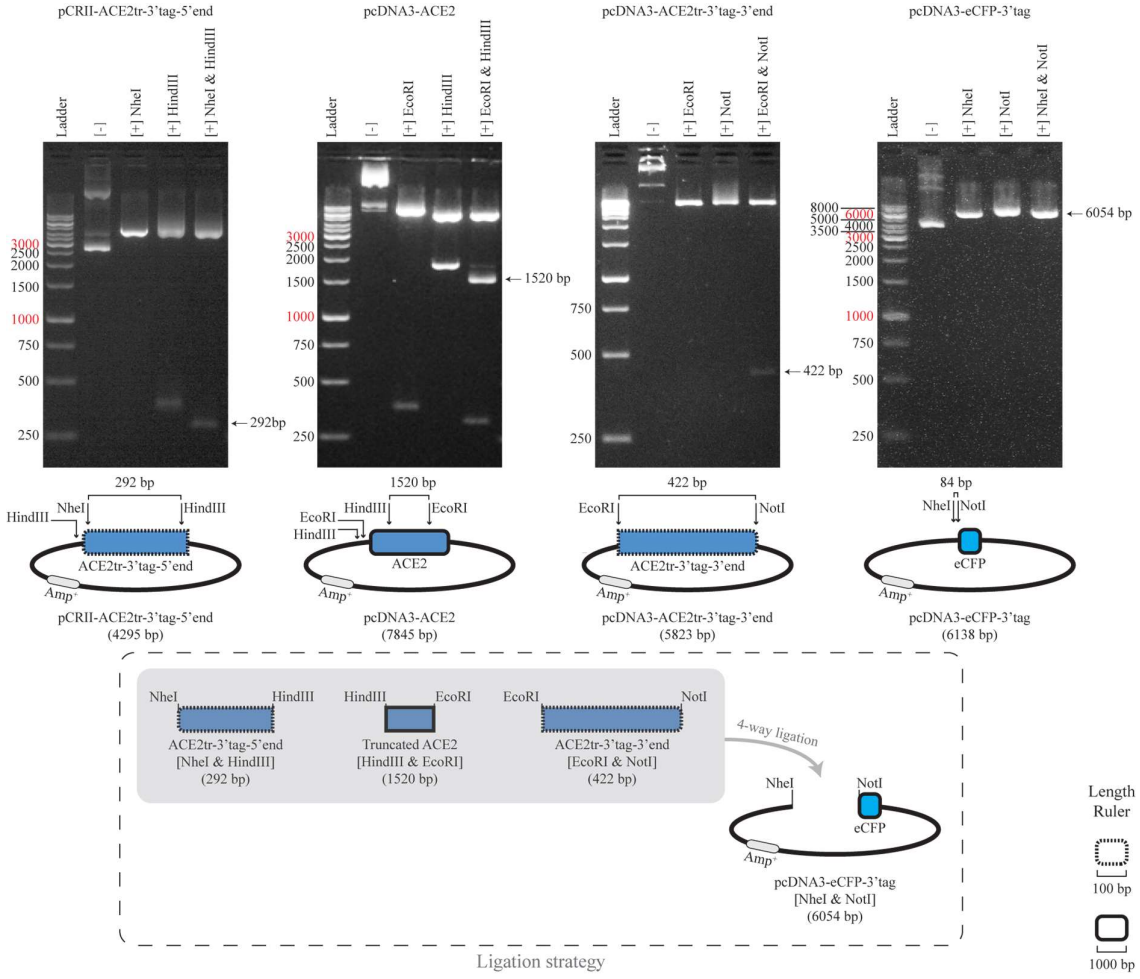
A.4.1.3.3. Construct final gene-fluorescence fusion plasmids

With all T-A cloned plasmids, synthesised plasmids, aligned DNA fragments, multi-fragments ligations were preformed to construct final plasmids of gene-fluorescence fusion for FRET-

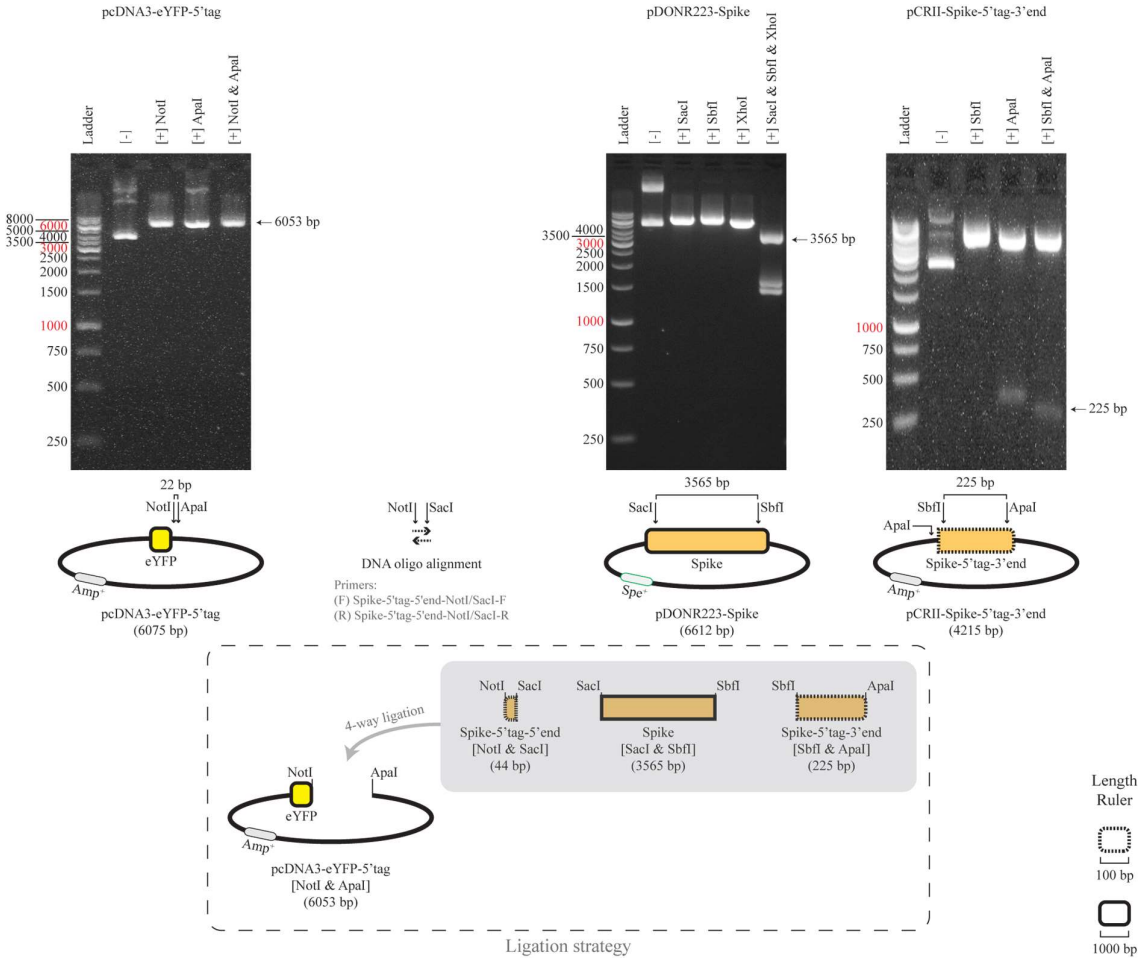
assay. ACE2tr 5’-fused with eCFP plasmid pcDNA3-eCFP-ACE2tr was constructed by three-fragment ligation with *NotI-EcoRI* digested ACE2tr fragment from pcDNA3-ACE2 plasmid, *EcoRI-ApaI* digested ACE2tr-5’tag-3’end fragment from pcDNA3-ACE2tr-5’tag-3’end plasmid, and *NotI-ApaI* digested pcDNA3-eCFP-5’tag vector backbone (Appendix 23). ACE2tr 3’-fused with eCFP plasmid pcDNA3-ACE2tr-eCFP was constructed by four-fragment ligation with *HindIII-EcoRI* digested ACE2tr fragment from pcDNA3-ACE2 plasmid, *NheI-HindIII* digested ACE2tr-3’tag-5’end fragment from pCRII-ACE2tr-3’tag-5’end plasmid, *EcoRI-NotI* digested ACE2tr-3’tag-3’end fragment from pcDNA3-ACE2tr-3’tag-3’end plasmid, and *NheI-NotI* digested pcDNA3-CFP-3’tag vector backbone (Appendix 24). Full-length Spike 5’-fused with eYFP plasmid pcDNA3-eYFP-Spike was constructed by four-fragment ligation with *SacI-SbfI* digested Spike fragment from pDONR223-Spike plasmid, *NotI-SacI* digested Spike-5’tag-5’end fragment from aligned Spike-5’tag-5’end DNA oligo annealed fragments, *SbfI-ApaI* digested Spike-5’tag-3’end fragment from pCRII-Spike-5’tag-3’end plasmid, and *NotI-ApaI* digested pcDNA3-eYFP-5’tag vector backbone (Appendix 25). Full-length Spike 3’-fused with eYFP plasmid pcDNA3-Spike-eYFP was constructed by four-fragment ligation with *SacI-SbfI* digested Spike fragment from pDONR223-Spike plasmid, *NheI-SacI* digested Spike-3’tag-5’end fragment from aligned Spike-3’tag-5’end DNA oligo annealed fragments, *SbfI-NotI* digested Spike-5’tag-3’end fragment from pCRII-Spike-3’tag-3’end plasmid, and *NheI-NotI* digested pcDNA3-eYFP-3’tag vector backbone (Appendix 26). TMPRSS2tr 5’-fused with mCherry plasmid pcDNA3-mCherry-TMPRSS2tr was constructed by three-fragment ligation with *KpnI-XbaI* digested TMPRSS2tr fragment from pcDNA3-TMPRSS2 plasmid, *NotI-KpnI* digested TMPRSS2tr-5’tag-5’end fragment from aligned TMPRSS2tr-5’tag-5’end DNA oligo annealed fragments, *NotI-XbaI* digested pcDNA3-mCherry-5’tag vector backbone (Appendix 27). TMPRSS2tr 3’-fused with mCherry plasmid pcDNA3-TMPRSS2tr-mCherry was attended to construct by four-fragment ligation with *KpnI-HindIII* digested TMPRSS2tr fragment from pcDNA3-TMPRSS2 plasmid, *NheI-KpnI* digested TMPRSS2tr-3’tag-5’end fragment from aligned TMPRSS2tr-3’tag-5’end DNA oligo annealed fragments, *HindIII-NotI* digested TMPRSS2tr-3’tag-3’end fragment from aligned TMPRSS2tr-3’tag-3’end DNA oligo annealed fragments, and *NheI-NotI* digested pcDNA3-mCherry-3’tag vector backbone (Appendix 28). However, the sub-cloning construction of pcDNA3-TMPRSS2tr-mCherry plasmid was unsuccessful, which no correct colonies were spotted.



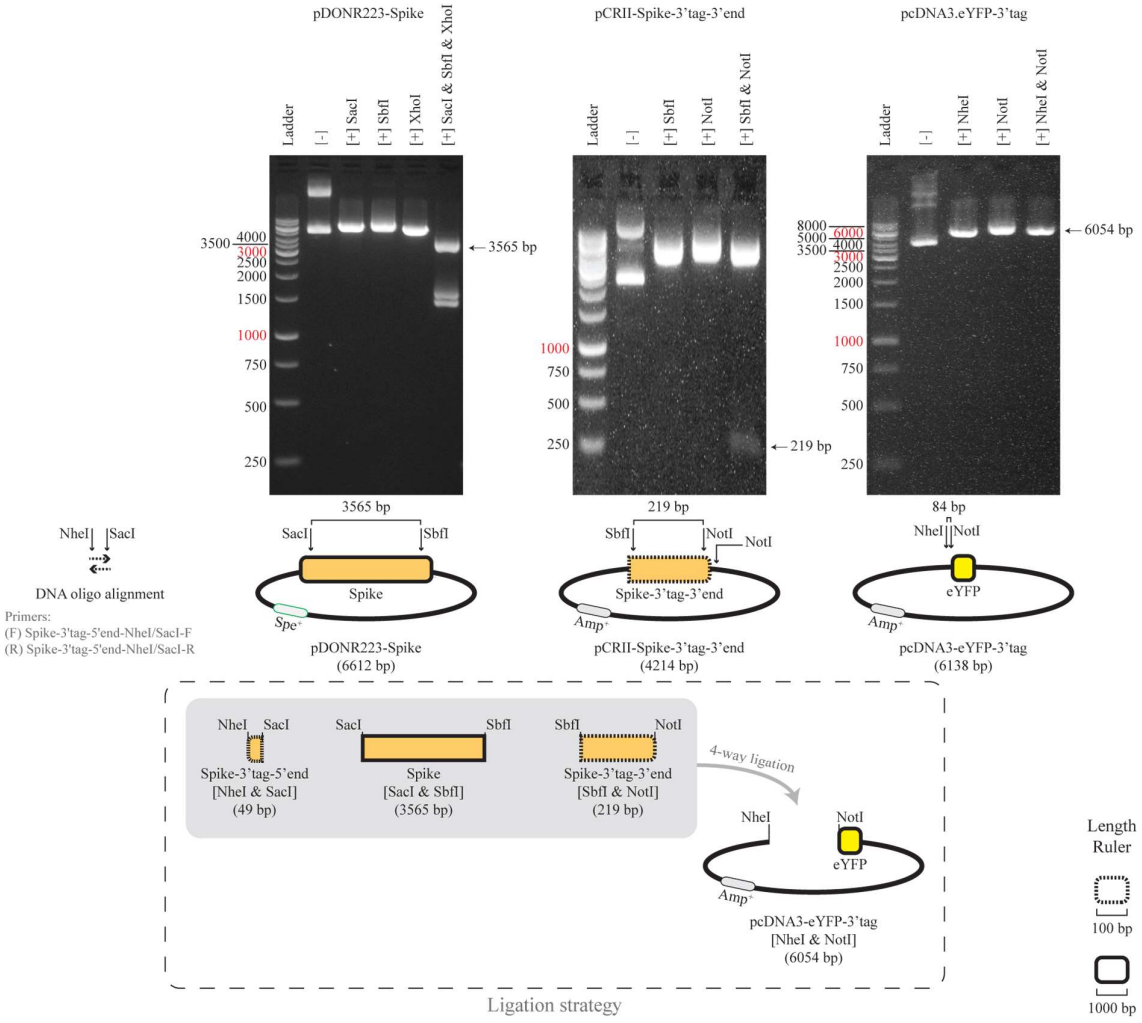
Appendix 23. Construction of pcDNA3-eCFP-ACE2tr. This plasmid is for FRET-based PPI detection assay.



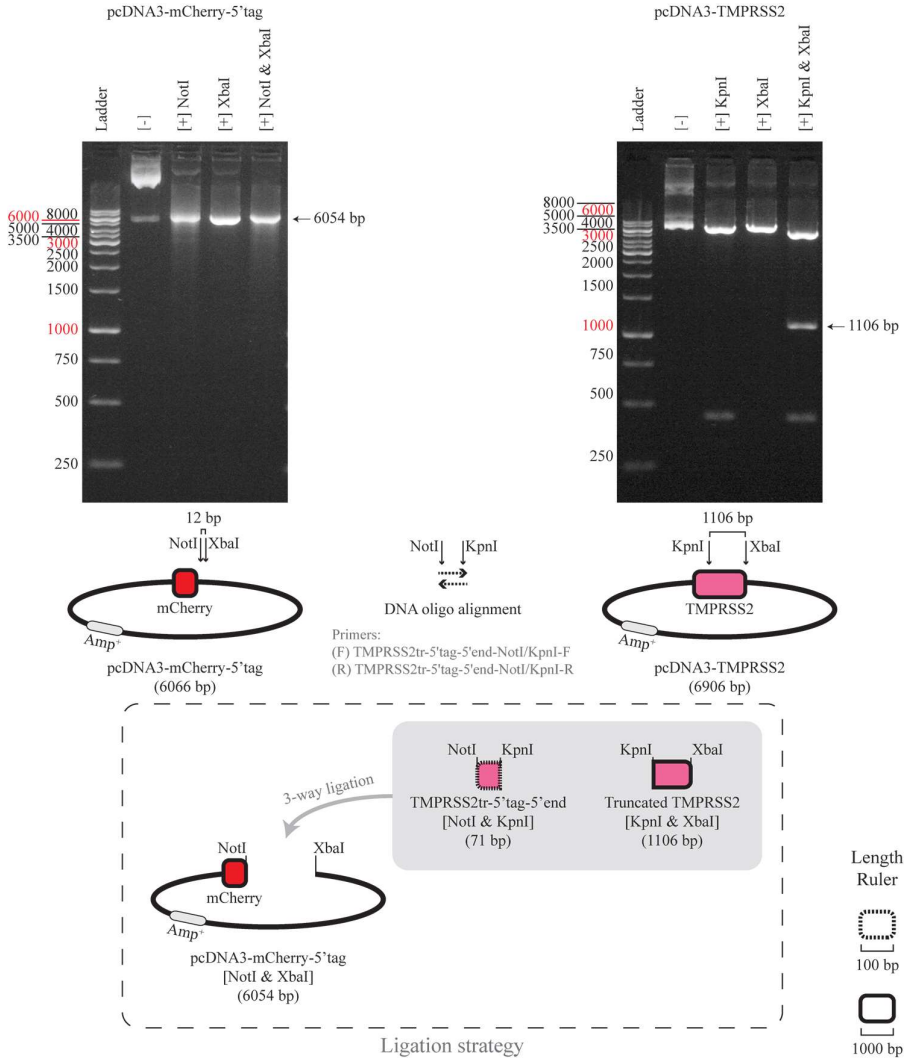
Appendix 24. Construction of pcDNA3-ACE2tr-eCFP. This plasmid is for FRET-based PPI detection assay.



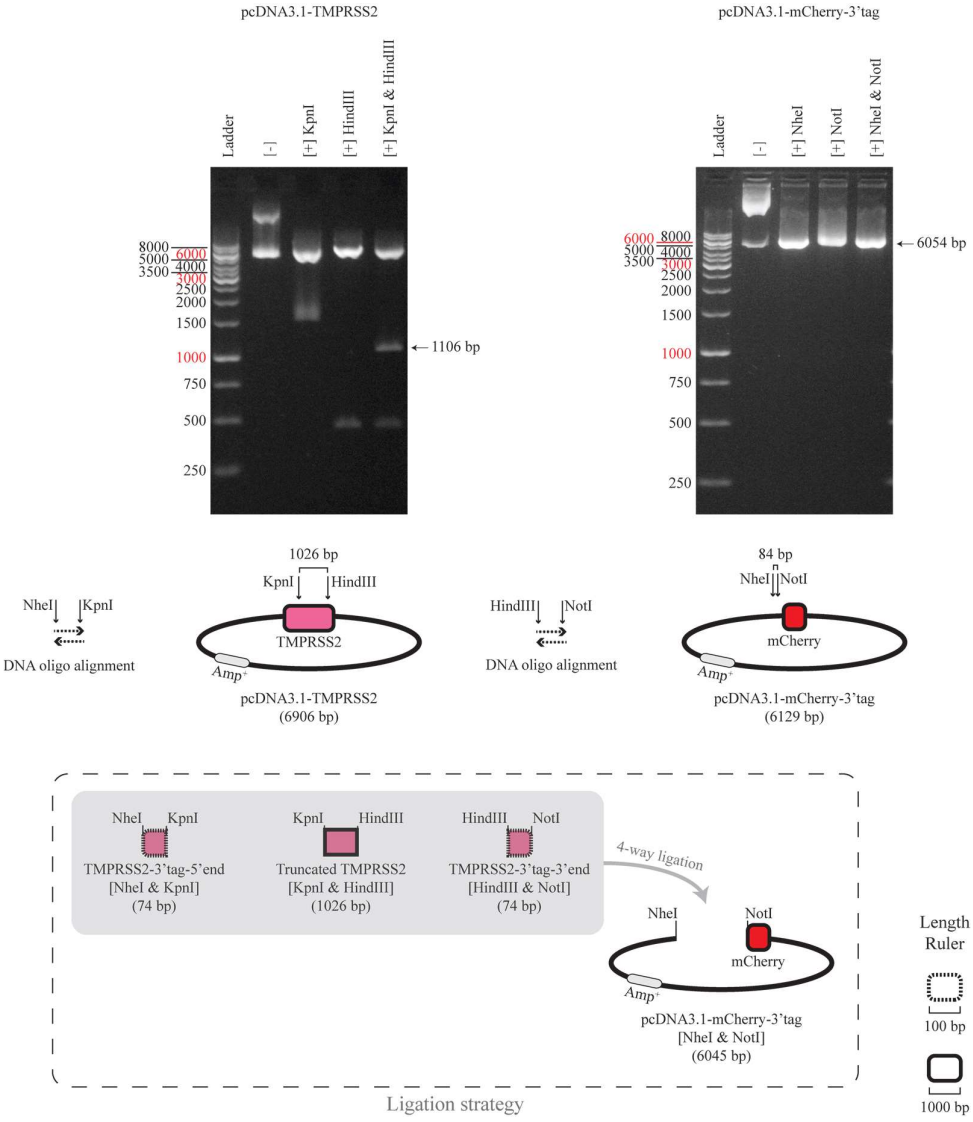
Appendix 25. Construction of pcDNA3-eYFP-Spike. This plasmid is for FRET-based PPI detection assay.



Appendix 26. Construction of pcDNA3-Spike-eYFP. This plasmid is for FRET-based PPI detection assay.



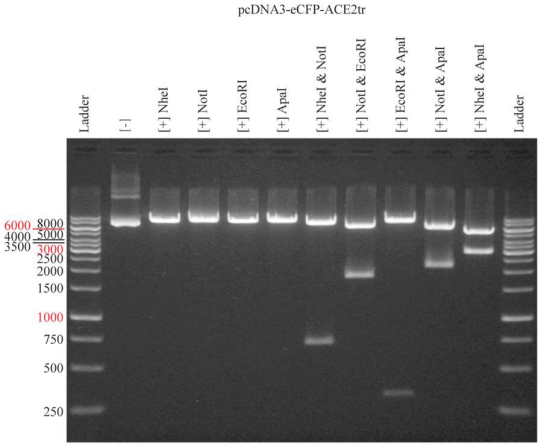
Appendix 27. Construction of pcDNA3-mCherry-TMPRSS2tr. This plasmid is for FRET-based PPI detection assay.



Appendix 28. Construction of pcDNA3-TMPRSS2tr-mCherry. This plasmid is for FRET-based PPI detection assay.

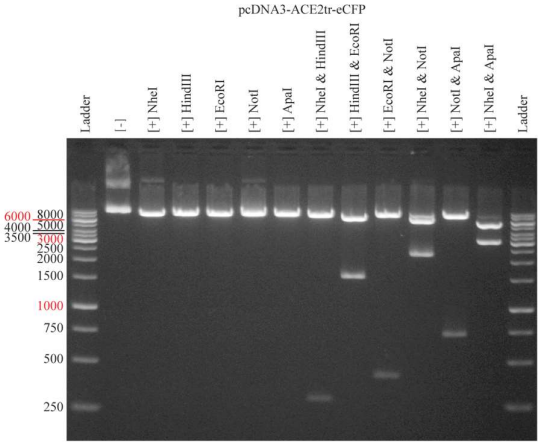
pcDNA3-eCFP-ACE2tr
(8288 bp)

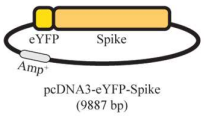
pcDNA3-eCFP-ACE2tr (8288 bp)	
Enzyme	Expected fragment lengths (bp)
NheI	8288
NotI	8288
EcoRI	8288
ApaI	8288
NheI & NotI	731, 7557
NotI & EcoRI	1807, 6481
EcoRI & ApaI	428, 7860
NotI & ApaI	2235, 6053
NheI & ApaI	2966, 5322



pcDNA3-ACE2tr-eCFP
(8294 bp)

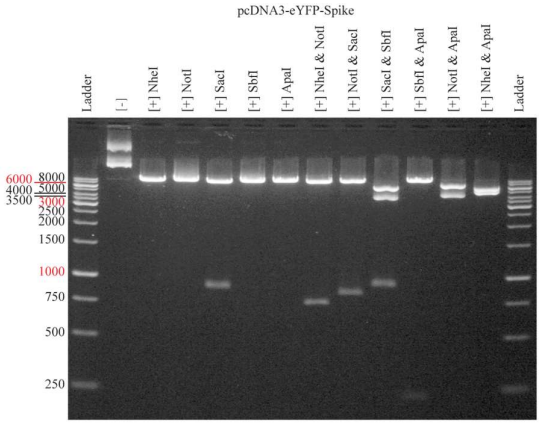
pcDNA3-ACE2tr-eCFP (8294 bp)	
Enzyme	Expected fragment lengths (bp)
NheI	8294
HindIII	8294
EcoRI	8294
NotI	8294
ApaI	8294
NheI & HindIII	292, 8002
HindIII & EcoRI	1520, 6774
EcoRI & NotI	1942, 6352
NheI & NotI	2234, 6060
NotI & ApaI	738, 7556
NheI & ApaI	2972, 5322

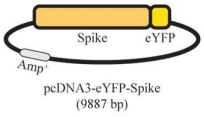




pcDNA3-eYFP-Spike
(9887 bp)

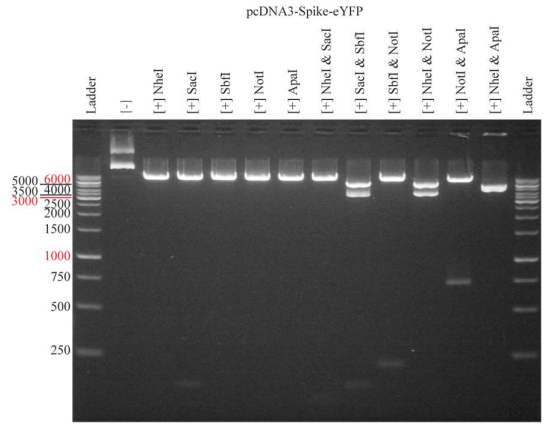
pcDNA3-eYFP-Spike (9887 bp)	
Enzyme	Expected fragment lengths (bp)
NheI	9887
NotI	9887
SacI	852, 9035
SbfI	9887
ApaI	9887
NheI & NotI	731, 9156
NotI & SacI	44, 808, 9035
SacI & SbfI	852, 3565, 5470
SbfI & ApaI	225, 9662
NotI & ApaI	3834, 6053
NheI & ApaI	4565, 5322

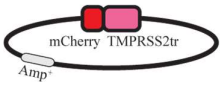




pcDNA3-eYFP-Spike
(9887 bp)

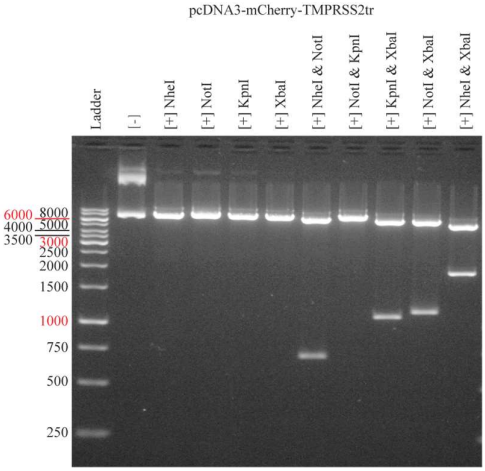
pcDNA3-Spike-eYFP (9887 bp)	
Enzyme	Expected fragment lengths (bp)
NheI	9887
SacI	126, 9761
SbfI	9887
NotI	9887
ApaI	9887
NheI & SacI	44, 79, 9761
SacI & SbfI	126, 3565, 6196
SbfI & NotI	219, 9668
NheI & NotI	3833, 6054
NotI & ApaI	732, 9155
NheI & ApaI	4565, 5322





pcDNA3.-mCherry-TMPRSS2tr
(7220 bp)

pcDNA3-mCherry-TMPRSS2tr (7226 bp)	
Enzyme	Expected fragment lengths (bp)
NheI	7226
NotI	7226
KpnI	7226
XbaI	7226
NheI & NotI	722, 6504
NotI & KpnI	66, 7160
KpnI & XbaI	1106, 6120
NotI & XbaI	1172, 6054
NheI & XbaI	1894, 5332



Length
Ruler

1000 bp

Appendix 29. Verification of fusion protein plasmids containing only both coding sequences of genes of interest and fluorescence reporter via restriction enzyme digestions.

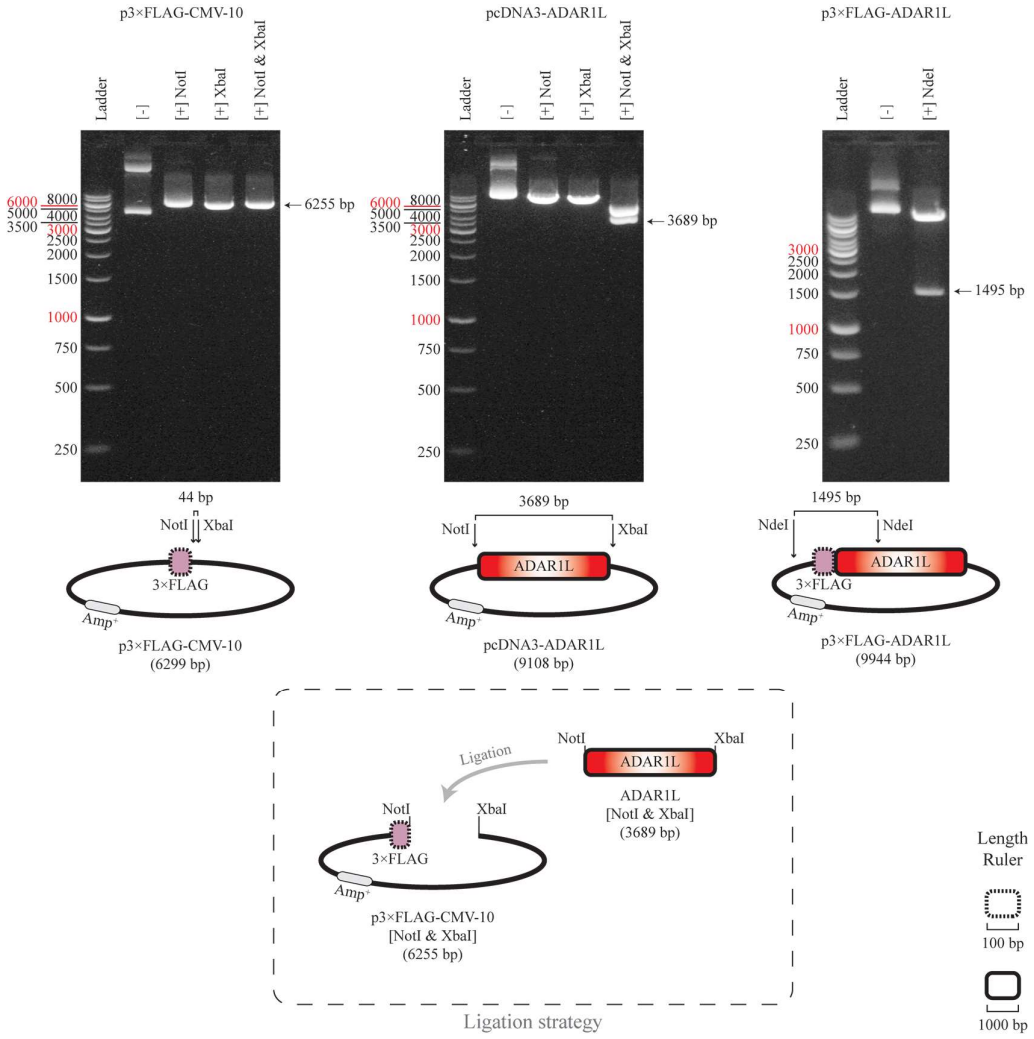
In Appendix 29, all the constructed fusion proteins plasmids including pcDNA3-eCFP-ACE2tr, pcDNA3-ACE2tr-eCFP, pcDNA3-eYFP-Spike, pcDNA3-Spike-eYFP, and

pcDNA3-mCherry-TMPRSS2tr are verified with enzyme digestions, although the data of pcDNA3-eCFP-ACE2tr is missed due to mis-handling of data. The sequences of all the constructed fusion proteins plasmids are verified with Sanger sequencing.

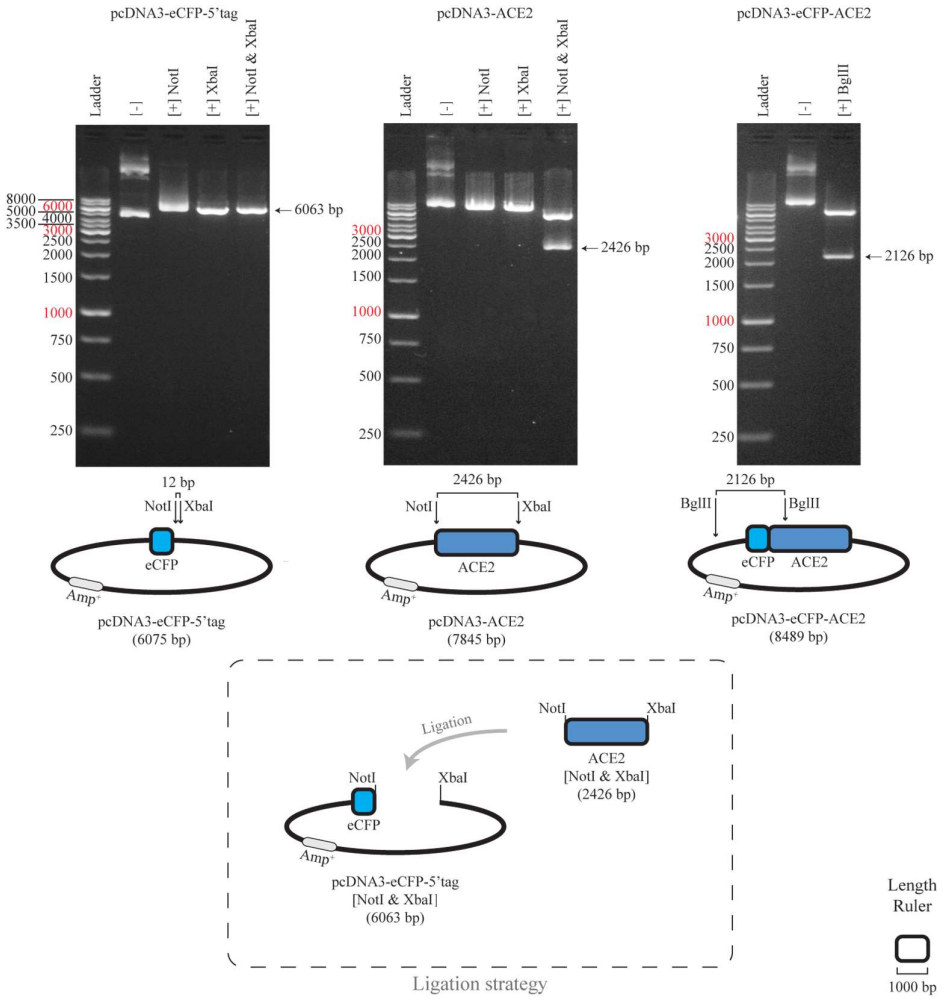
Appendix 30. Step-by-step detailed description of molecular cloning constructions of plasmids used in RNA editing events detection.

A.4.1.4. Molecular cloning in constructions of plasmids for RNA editing events detections

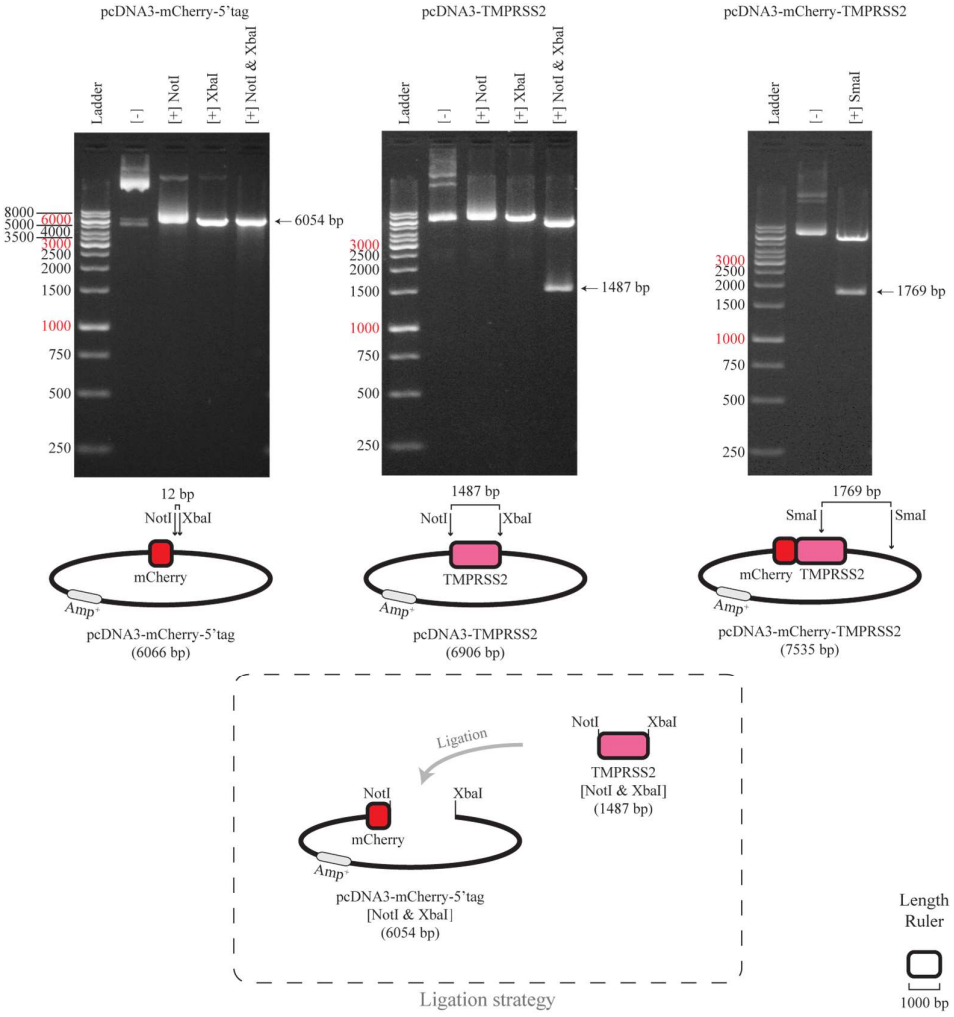
Plasmid of ADAR1L 5'-fused with 3×FLAG tags p3×FLAG-ADAR1L was constructed by ligation with *NotI-XbaI* digested ADAR1L wild-type CDS fragment from commercially synthesised pcDNA3-ADAR1L plasmid, and *NotI-XbaI* digested p3×FLAG-CMV-10 vector backbone (Appendix 31). ACE2 5'-fused with eCFP pcDNA3-eCFP-ACE2 was constructed by ligation with *NotI-XbaI* digested ACE2 wild-type CDS fragment from pcDNA3-ACE2 plasmid, and *NotI-XbaI* digested pcDNA3-eCFP-5'tag vector backbone (Appendix 32). Lastly, TMPRSS2 5'-fused with mCherry pcDNA3-mCherry-TMPRSS2 was constructed by ligation with *NotI-XbaI* digested TMPRSS2 wild-type CDS fragment from pcDNA3-TMPRSS2 plasmid, and *NotI-XbaI* digested pcDNA3-mCherry-5'tag vector backbone (Appendix 33). The sequences of all three plasmids were verified with Sanger sequencing, in which no mutations and other errors were found.



Appendix 31. Construction and digestion verification of p3×FLAG-ADAR1L. This plasmid is for RNA editing events detection.



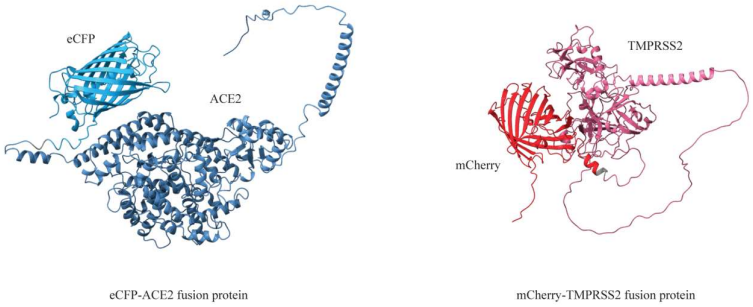
Appendix 32. Construction and digestion verification of pcDNA3-eCFP-ACE2. This plasmid is for RNA editing events detection.



Appendix 33. Construction and digestion verification of pcDNA3-mCherry-TMPRSS2. This plasmid is for RNA editing events detection.

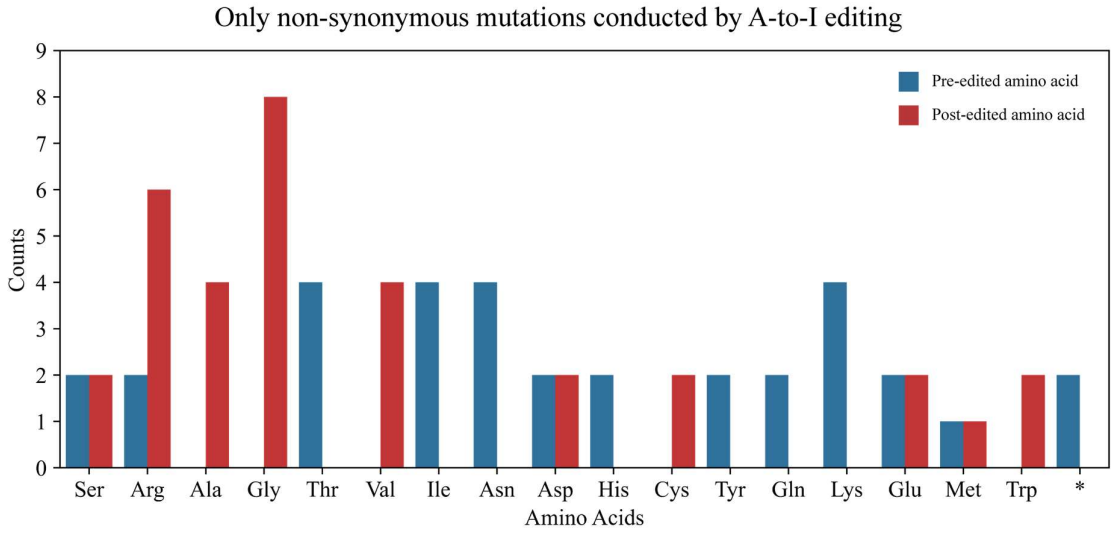
Plasmid Group	Plasmid name	Backbone	Antibiotics	Insert Clone Site	Source	Purpose	Mutation	Sequencing	Note
Empty Fluoresence Fusion Vectors	pcDNA3-eCFP-3'tag	pcDNA3.1(+)	Ampicillin	NotI-eCFP-ApaI	Commercial DNA synthesis	eCFP fusion protein (3')	None	Whole plasmid	
Empty Fluoresence Fusion Vectors	pcDNA3-eCFP-5'tag	pcDNA3.1(+)	Ampicillin	NheI-eCFP-NotI	Commercial DNA synthesis	eCFP fusion protein (5')	None	Whole plasmid	
Empty Fluoresence Fusion Vectors	pcDNA3-eYFP-3'tag	pcDNA3.1(+)	Ampicillin	NotI-eYFP-ApaI	Commercial DNA synthesis	eYFP fusion protein (3')	None	Whole plasmid	
Empty Fluoresence Fusion Vectors	pcDNA3-eYFP-5'tag	pcDNA3.1(+)	Ampicillin	NheI-eYFP-NotI	Commercial DNA synthesis	eYFP fusion protein (5')	None	Whole plasmid	
Empty Fluoresence Fusion Vectors	pcDNA3-mCherry-3'tag	pcDNA3.1(+)	Ampicillin	NotI-mCherry-ApaI	Subcloning	mCherry fusion protein (3')	None	Insert	
Empty Fluoresence Fusion Vectors	pcDNA3-mCherry-5'tag	pcDNA3.1(+)	Ampicillin	NheI-mCherry-NotI	Subcloning	mCherry fusion protein (5')	None	Insert	
Empty Fluoresence Fusion Vectors	pcDNA3-mRFP-3'tag	pcDNA3.1(+)	Ampicillin	NotI-mRFP-ApaI	Subcloning	mRFP fusion protein (3')	None	Insert	
Empty Fluoresence Fusion Vectors	pcDNA3-mRFP-5'tag	pcDNA3.1(+)	Ampicillin	NheI-mRFP-NotI	Subcloning	mRFP fusion protein (5')	None	Insert	
Fluoresence Fusion - Human Genes	pcDNA3-ACE2tr-eCFP	pcDNA3-eCFP-3'tag	Ampicillin	NheI-ACE2tr-NotI	Subcloning	ACE2tr-eCFP fusion overexpression	None	Insert	
Fluoresence Fusion - Human Genes	pcDNA3-eCFP-ACE2	pcDNA3-eCFP-5'tag	Ampicillin	NotI-ACE2-XbaI	Subcloning	eCFP-ACE2 fusion overexpression	None	Insert	NM_021804.3
Fluoresence Fusion - Human Genes	pcDNA3-eCFP-ACE2tr	pcDNA3-eCFP-5'tag	Ampicillin	NotI-ACE2tr-ApaI	Subcloning	eCFP-ACE2tr fusion overexpression	None	Insert	
Fluoresence Fusion - Human Genes	pcDNA3-mCherry-TMPRSS2	pcDNA3-mCherry-5'tag	Ampicillin	NotI-TMPRSS2-XbaI	Subcloning	mCherry-TMPRSS2 fusion overexpression	None	Insert	NM_005656.4
Fluoresence Fusion - Human Genes	pcDNA3-mCherry-TMPRSS2tr	pcDNA3-mCherry-5'tag	Ampicillin	NotI-TMPRSS2tr-XbaI	Subcloning	mCherry-TMPRSS2tr fusion overexpression	None	Insert	
Fluoresence Fusion - SARS-CoV-2	pcDNA3-eYFP-Spike	pcDNA3-eYFP-5'tag	Ampicillin	NotI-Spike-ApaI	Subcloning	eYFP-Spike fusion overexpression	None	Insert	NC_045512.2
Fluoresence Fusion - SARS-CoV-2	pcDNA3-Spike-eYFP	pcDNA3-eYFP-3'tag	Ampicillin	NheI-Spike-NotI	Subcloning	Spike-eYFP fusion overexpression	None	Insert	NC_045512.2
Tag Fusion - Human Genes	p3xFLAG-ADAR1L	p3xFLAG-CMV-10	Ampicillin	NotI-ADAR1L-XbaI	Subcloning	3xFLAG-ADAR1L fusion overexpression	None	Insert	NM_001111.5
T-A cloning	pCRII-ACE2tr-3'tag-5'end	pCRII	Ampicillin		Subcloning	Store ACE2tr-3'tag-5'end PCR product	None	Insert	
T-A cloning	pCRII-Spike-5'tag-3'end	pCRII	Ampicillin		Subcloning	Store Spike-5'tag-3'end PCR product	None	Insert	
T-A cloning	pCRII-Spike-3'tag-3'end	pCRII	Ampicillin		Subcloning	Store Spike-3'tag-3'end PCR product	None	Insert	
Fragment storing	pcDNA3-ACE2tr-5'tag-3'end	pcDNA3.1(+)	Ampicillin		Commercial DNA synthesis	Store ACE2tr-5'tag-3'end	None	Whole plasmid	
Fragment storing	pcDNA3-ACE2tr-3'tag-3'end	pcDNA3.1(+)	Ampicillin		Commercial DNA synthesis	Store ACE2tr-3'tag-3'end	None	Whole plasmid	

Appendix 34. Summary list of all constructed plasmid.



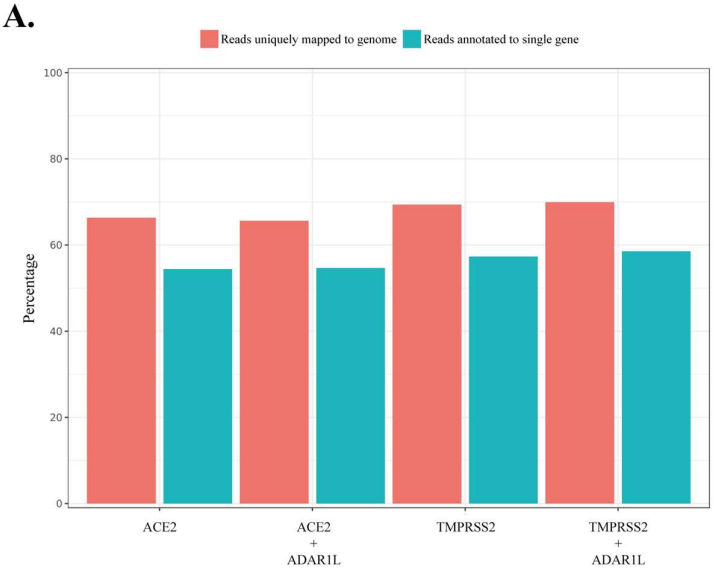
eCFP-ACE2 fusion protein　　mCherry-TMPRSS2 fusion protein

Appendix 35. Demonstrating predicted folding structures of fusion proteins through AlphaFold2.



Appendix 36. Counts of pre-edited amino acids and post-edited amino acids are changed after applying A-to-I editing, which non-synonymous mutations are only included.

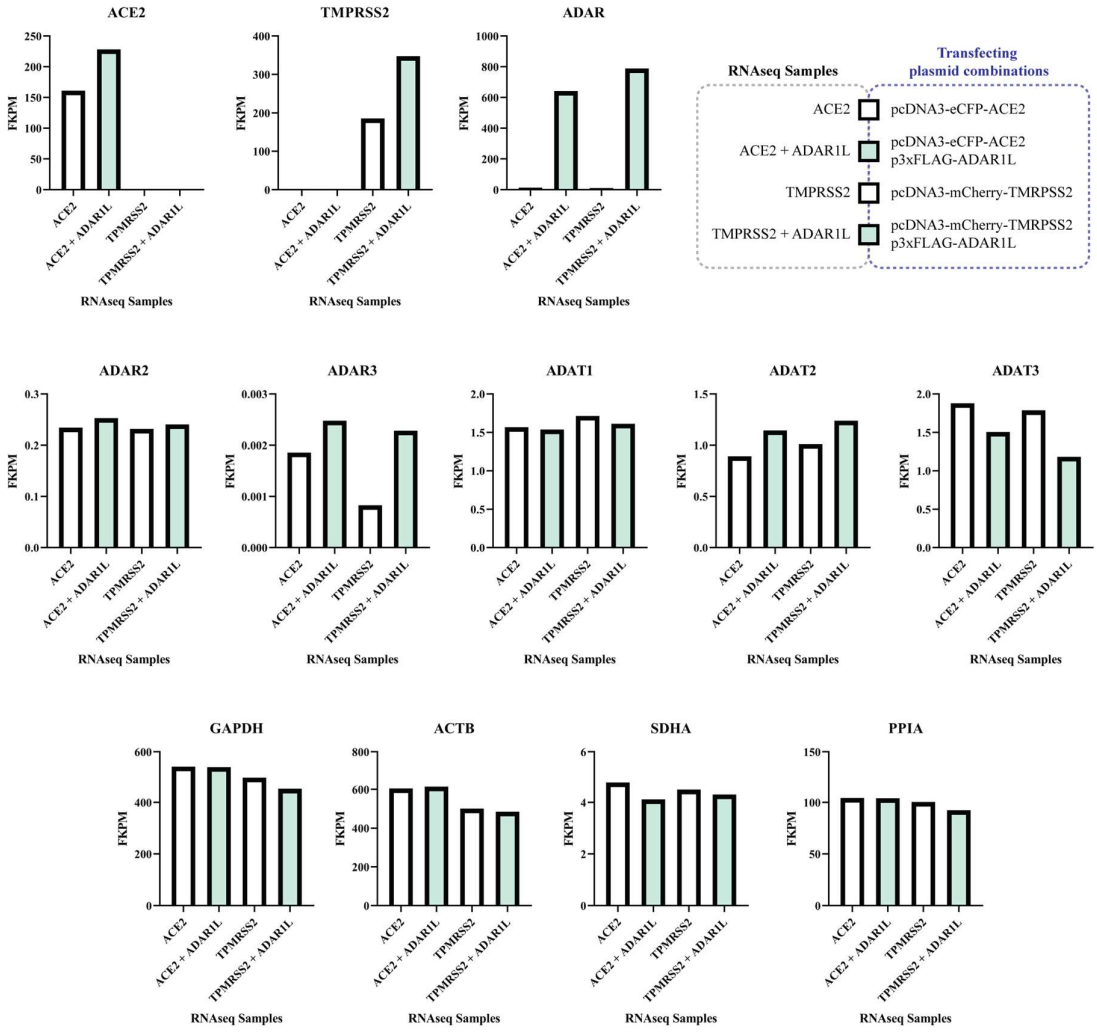
A.



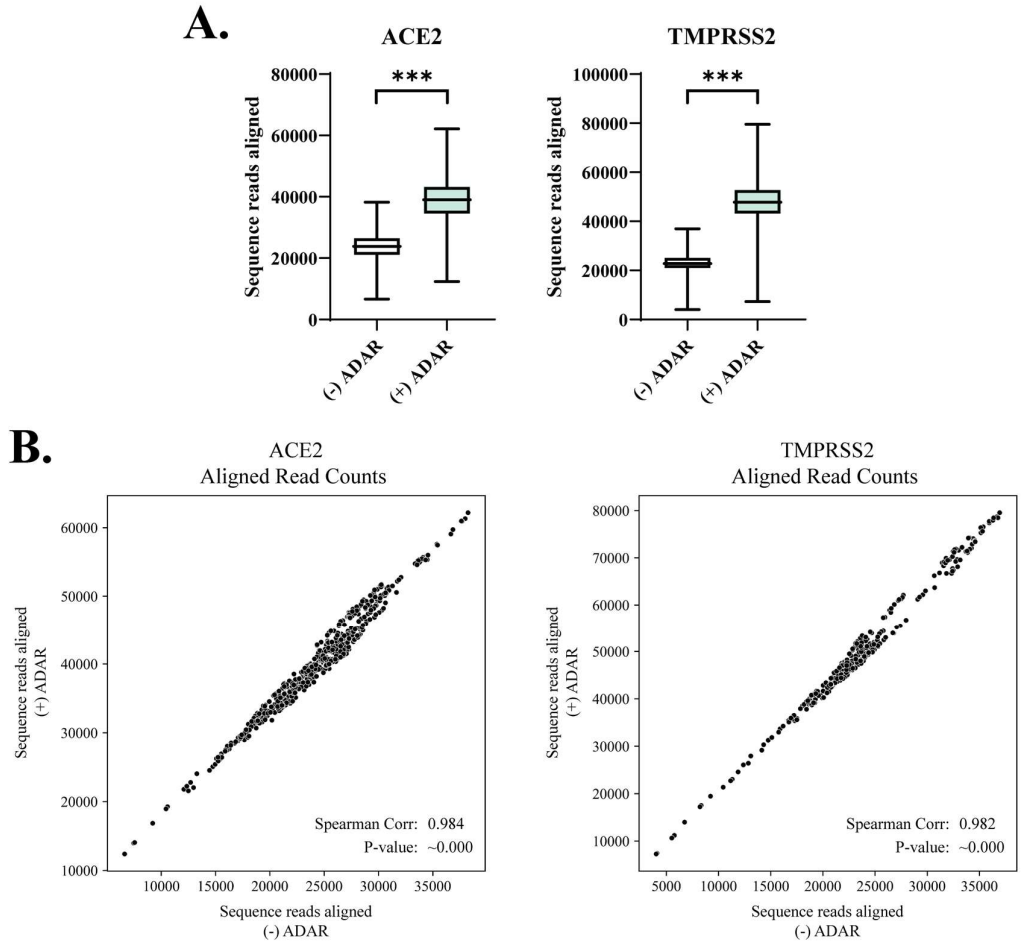
B.

Sample name	Total reads	Paired reads mapped to genome	Percentage of paired reads mapped to genome	Paired reads mapped to one feature	Percentage of paired reads mapped to one feature	Paired mapped reads not mapped to any known feature	Percentage of paired reads that mapped to the genome but not to any known feature
ACE2	30025209	19915862	66.33%	16347112	54.44%	3273579	10.90%
ACE2+ADAR1L	32930513	21614279	65.64%	17997906	54.65%	3325442	10.10%
TMPRSS2	30083014	20880114	69.41%	17249916	57.34%	3326587	11.06%
TMPRSS2+ADAR1L	33403978	23354497	69.92%	19550451	58.53%	3482630	10.43%

Appendix 37. Quality control data of RNAseq results with samples ACE2, ACE2+ADAR1L, TMPRSS2, and TMPRSS2+ADAR1L. (A) Percentages of sequence reads mapped to the genome and to single genes. (B) Other important quality control metrics of the RNAseq experiment.



Appendix 38. Expression levels (FKPM) of different genes from RNAseq data of HEK293T cells transfected with different plasmid combinations. No obviously varied expression levels observed in housekeeping genes (GAPDH, ACTB, SDHA and PPIA), and other RNA editing related genes (ADAR2, ADAR3, ADAT1, ADAT2, ADAT3). Although ADAR3 seems to have slightly lower expression level in samples with ADAR expressions, ADAR3 does not have RNA editing capabilities anyway.



Appendix 39. Sequence Depth of RNAseq data aligned to ACE2 and TMPRSS2 genes. (A) Sequence reads of ACE2 and TMPRSS2 are significantly enriched in the samples with or without ADAR1L overexpression; (B) The sequence reads aligned to ACE2 and TMPRSS2 are approximately equal in counts for every locations between samples with or without ADAR1L overexpression.

References

1. Stelzer, G., et al., *The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses.* Curr Protoc Bioinformatics, 2016. **54**: p. 1 30 1-1 30 33.
2. Matthews, M.M., et al., *Structures of human ADAR2 bound to dsRNA reveal base-flipping mechanism and basis for site selectivity.* Nat Struct Mol Biol, 2016. **23**(5): p. 426-33.
3. Su, S., *Defining Roles of Adenosine Deaminase Acting on RNA (ADAR) in Virus Infection*, in *Faculty of Science*. 2019, University of Technology Sydney.
4. Thomas, J.M. and P.A. Beal, *How do ADARs bind RNA? New protein-RNA structures illuminate substrate recognition by the RNA editing ADARs.* Bioessays, 2017. **39**(4).
5. Crick, F.H., *Codon--anticodon pairing: the wobble hypothesis.* J Mol Biol, 1966. **19**(2): p. 548-55.
6. Bass, B.L. and H. Weintraub, *A developmentally regulated activity that unwinds RNA duplexes.* Cell, 1987. **48**(4): p. 607-13.
7. Wagner, R.W., et al., *A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and Xenopus eggs.* Proc Natl Acad Sci U S A, 1989. **86**(8): p. 2647-51.
8. Bass, B.L. and H. Weintraub, *An unwinding activity that covalently modifies its double-stranded RNA substrate.* Cell, 1988. **55**(6): p. 1089-98.
9. Eisenberg, E. and E.Y. Levanon, *A-to-I RNA editing - immune protector and transcriptome diversifier.* Nat Rev Genet, 2018. **19**(8): p. 473-490.
10. Nishikura, K., *Functions and regulation of RNA editing by ADAR deaminases.* Annu Rev Biochem, 2010. **79**: p. 321-49.
11. Kawahara, Y., et al., *Frequency and fate of microRNA editing in human brain.* Nucleic Acids Res, 2008. **36**(16): p. 5270-80.
12. Jin, Y., W. Zhang, and Q. Li, *Origins and evolution of ADAR-mediated RNA editing.* IUBMB Life, 2009. **61**(6): p. 572-8.
13. Dahabieh, M.S., et al., *Sequence-dependent structural dynamics of primate adenosine-to-inosine editing substrates.* Chembiochem, 2012. **13**(18): p. 2714-21.
14. Melcher, T., et al., *RED2, a brain-specific member of the RNA-specific adenosine deaminase family.* J Biol Chem, 1996. **271**(50): p. 31795-8.
15. Maas, S., A.P. Gerber, and A. Rich, *Identification and characterization of a human tRNA-specific adenosine deaminase related to the ADAR family of pre-mRNA editing enzymes.* Proc Natl Acad Sci U S A, 1999. **96**(16): p. 8895-900.
16. Gerber, A.P. and W. Keller, *RNA editing by base deamination: more enzymes, more targets, new mysteries.* Trends Biochem Sci, 2001. **26**(6): p. 376-84.
17. Lykke-Andersen, S., S. Pinol-Roma, and J. Kjems, *Alternative splicing of the ADAR1 transcript in a region that functions either as a 5'-UTR or an ORF.* RNA, 2007. **13**(10): p. 1732-44.
18. Liu, Y., et al., *Functionally distinct double-stranded RNA-binding domains associated with alternative splice site variants of the interferon-inducible double-stranded RNA-specific adenosine deaminase.* J Biol Chem, 1997. **272**(7): p. 4419-28.

19. Melcher, T., et al., *A mammalian RNA editing enzyme.* Nature, 1996. **379**(6564): p. 460-4.
20. Patterson, J.B. 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.* Mol Cell Biol, 1995. **15**(10): p. 5376-88.
21. Keegan, L.P., et al., *Functional conservation in human and Drosophila of Metazoan ADAR2 involved in RNA editing: loss of ADAR1 in insects.* Nucleic Acids Res, 2011. **39**(16): p. 7249-62.
22. George, C.X. and C.E. Samuel, *Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible.* Proc Natl Acad Sci U S A, 1999. **96**(8): p. 4621-6.
23. Kawakubo, K. and C.E. Samuel, *Human RNA-specific adenosine deaminase (ADAR1) gene specifies transcripts that initiate from a constitutively active alternative promoter.* Gene, 2000. **258**(1-2): p. 165-72.
24. Patterson, J.B., et al., *Mechanism of interferon action: double-stranded RNA-specific adenosine deaminase from human cells is inducible by alpha and gamma interferons.* Virology, 1995. **210**(2): p. 508-11.
25. Der, S.D., et al., *Identification of genes differentially regulated by interferon alpha, beta, or gamma using oligonucleotide arrays.* Proc Natl Acad Sci U S A, 1998. **95**(26): p. 15623-8.
26. George, C.X. and C.E. Samuel, *Characterization of the 5'-flanking region of the human RNA-specific adenosine deaminase ADAR1 gene and identification of an interferon-inducible ADAR1 promoter.* Gene, 1999. **229**(1-2): p. 203-13.
27. Sung, Y.J., et al., *Gene-smoking interactions identify several novel blood pressure loci in the Framingham Heart Study.* Am J Hypertens, 2015. **28**(3): p. 343-54.
28. Filippini, A., et al., *Differential Enzymatic Activity of Rat ADAR2 Splicing Variants Is Due to Altered Capability to Interact with RNA in the Deaminase Domain.* Genes (Basel), 2018. **9**(2).
29. Tan, M.H., et al., *Dynamic landscape and regulation of RNA editing in mammals.* Nature, 2017. **550**(7675): p. 249-254.
30. Wang, Y., et al., *RNA binding candidates for human ADAR3 from substrates of a gain of function mutant expressed in neuronal cells.* Nucleic Acids Res, 2019. **47**(20): p. 10801-10814.
31. Hurst, S.R., et al., *Deamination of mammalian glutamate receptor RNA by Xenopus dsRNA adenosine deaminase: similarities to in vivo RNA editing.* RNA, 1995. **1**(10): p. 1051-60.
32. Dabiri, G.A., et al., *Editing of the GLuR-B ion channel RNA in vitro by recombinant double-stranded RNA adenosine deaminase.* EMBO J, 1996. **15**(1): p. 34-45.
33. Polson, A.G., B.L. Bass, and J.L. Casey, *RNA editing of hepatitis delta virus antigenome by dsRNA-adenosine deaminase.* Nature, 1996. **380**(6573): p. 454-6.
34. Maraia, R.J. and A.G. Arimbasseri, *Factors That Shape Eukaryotic tRNAomes: Processing, Modification and Anticodon-Codon Use.* Biomolecules, 2017. **7**(1).
35. Torres, A.G., et al., *A-to-I editing on tRNAs: biochemical, biological and evolutionary implications.* FEBS Lett, 2014. **588**(23): p. 4279-86.
36. Agris, P.F., et al., *Celebrating wobble decoding: Half a century and still much is new.* RNA Biol, 2018. **15**(4-5): p. 537-553.

37. Eggington, J.M., T. Greene, and B.L. Bass, *Predicting sites of ADAR editing in double-stranded RNA.* Nat Commun, 2011. **2**: p. 319.
38. Stephens, O.M., B.L. Haudenschild, and P.A. Beal, *The binding selectivity of ADAR2's dsRBMs contributes to RNA-editing selectivity.* Chem Biol, 2004. **11**(9): p. 1239-50.
39. Stefl, R., et al., *Structure and specific RNA binding of ADAR2 double-stranded RNA binding motifs.* Structure, 2006. **14**(2): p. 345-55.
40. Ohman, M., A.M. Kallman, and B.L. Bass, *In vitro analysis of the binding of ADAR2 to the pre-mRNA encoding the GluR-B R/G site.* RNA, 2000. **6**(5): p. 687-97.
41. Kallman, A.M., M. Sahlin, and M. Ohman, *ADAR2 A-->I editing: site selectivity and editing efficiency are separate events.* Nucleic Acids Res, 2003. **31**(16): p. 4874-81.
42. Liu, Y., et al., *Double-stranded RNA-specific adenosine deaminase: nucleic acid binding properties.* Methods, 1998. **15**(3): p. 199-205.
43. Tolbert, B.S., et al., *Major groove width variations in RNA structures determined by NMR and impact of 13C residual chemical shift anisotropy and 1H-13C residual dipolar coupling on refinement.* J Biomol NMR, 2010. **47**(3): p. 205-19.
44. Stefl, R., et al., *The solution structure of the ADAR2 dsRBM-RNA complex reveals a sequence-specific readout of the minor groove.* Cell, 2010. **143**(2): p. 225-37.
45. Jayachandran, U., H. Grey, and A.G. Cook, *Nuclear factor 90 uses an ADAR2-like binding mode to recognize specific bases in dsRNA.* Nucleic Acids Res, 2016. **44**(4): p. 1924-36.
46. Blaszczyk, J., et al., *Noncatalytic assembly of ribonuclease III with double-stranded RNA.* Structure, 2004. **12**(3): p. 457-66.
47. Wang, Y. and P.A. Beal, *Probing RNA recognition by human ADAR2 using a high-throughput mutagenesis method.* Nucleic Acids Res, 2016. **44**(20): p. 9872-9880.
48. Herbert, A. and A. Rich, *The role of binding domains for dsRNA and Z-DNA in the in vivo editing of minimal substrates by ADAR1.* Proc Natl Acad Sci U S A, 2001. **98**(21): p. 12132-7.
49. Wang, Y., S. Park, and P.A. Beal, *Selective Recognition of RNA Substrates by ADAR Deaminase Domains.* Biochemistry, 2018. **57**(10): p. 1640-1651.
50. Schwartz, T., et al., *Structure of the DLM-1-Z-DNA complex reveals a conserved family of Z-DNA-binding proteins.* Nat Struct Biol, 2001. **8**(9): p. 761-5.
51. Herbert, A., et al., *A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase.* Proc Natl Acad Sci U S A, 1997. **94**(16): p. 8421-6.
52. Wang, G. and K.M. Vasquez, *Dynamic alternative DNA structures in biology and disease.* Nat Rev Genet, 2023. **24**(4): p. 211-234.
53. Krall, J.B., et al., *Structure and Formation of Z-DNA and Z-RNA.* Molecules, 2023. **28**(2).
54. Athanasiadis, A., et al., *The crystal structure of the Zbeta domain of the RNA-editing enzyme ADAR1 reveals distinct conserved surfaces among Z-domains.* J Mol Biol, 2005. **351**(3): p. 496-507.
55. Ha, S.C., et al., *The crystal structure of the second Z-DNA binding domain of human DAI (ZBP1) in complex with Z-DNA reveals an unusual binding mode to Z-DNA.* Proc Natl Acad Sci U S A, 2008. **105**(52): p. 20671-6.
56. Schwartz, T., et al., *Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA.* Science, 1999. **284**(5421): p. 1841-5.

57. Kolimi, N., Y. Ajjugal, and T. Rathinavelan, *A B-Z junction induced by an A ... A mismatch in GAC repeats in the gene for cartilage oligomeric matrix protein promotes binding with the hZalphaADAR1 protein.* J Biol Chem, 2017. **292**(46): p. 18732-18746.
58. Bothe, J.R., K. Lowenhaupt, and H.M. Al-Hashimi, *Incorporation of CC steps into Z-DNA: interplay between B-Z junction and Z-DNA helical formation.* Biochemistry, 2012. **51**(34): p. 6871-9.
59. Herbert, A.G., et al., *Z-DNA binding protein from chicken blood nuclei.* Proc Natl Acad Sci U S A, 1993. **90**(8): p. 3339-42.
60. Herbert, A., *Z-DNA and Z-RNA in human disease.* Commun Biol, 2019. **2**: p. 7.
61. Ditlevson, J.V., et al., *Inhibitory effect of a short Z-DNA forming sequence on transcription elongation by T7 RNA polymerase.* Nucleic Acids Res, 2008. **36**(10): p. 3163-70.
62. Bae, S., et al., *Energetics of Z-DNA binding protein-mediated helicity reversals in DNA, RNA, and DNA-RNA duplexes.* J Phys Chem B, 2013. **117**(44): p. 13866-71.
63. Zheng, Y., C. Lorenzo, and P.A. Beal, *DNA editing in DNA/RNA hybrids by adenosine deaminases that act on RNA.* Nucleic Acids Res, 2017. **45**(6): p. 3369-3377.
64. Mannion, N.M., et al., *The RNA-editing enzyme ADAR1 controls innate immune responses to RNA.* Cell Rep, 2014. **9**(4): p. 1482-94.
65. Haudenschild, B.L., et al., *A transition state analogue for an RNA-editing reaction.* J Am Chem Soc, 2004. **126**(36): p. 11213-9.
66. Jimeno, S., et al., *ADAR-mediated RNA editing of DNA:RNA hybrids is required for DNA double strand break repair.* Nat Commun, 2021. **12**(1): p. 5512.
67. Phelps, K.J., et al., *Recognition of duplex RNA by the deaminase domain of the RNA editing enzyme ADAR2.* Nucleic Acids Res, 2015. **43**(2): p. 1123-32.
68. Pokharel, S., et al., *Matching active site and substrate structures for an RNA editing reaction.* J Am Chem Soc, 2009. **131**(33): p. 11882-91.
69. Kuttan, A. and B.L. Bass, *Mechanistic insights into editing-site specificity of ADARs.* Proc Natl Acad Sci U S A, 2012. **109**(48): p. E3295-304.
70. Macbeth, M.R., et al., *Inositol hexakisphosphate is bound in the ADAR2 core and required for RNA editing.* Science, 2005. **309**(5740): p. 1534-9.
71. Wong, S.K., S. Sato, and D.W. Lazinski, *Substrate recognition by ADAR1 and ADAR2.* RNA, 2001. **7**(6): p. 846-58.
72. Jayalath, P., et al., *Synthesis and evaluation of an RNA editing substrate bearing 2'-deoxy-2'-mercaptoadenosine.* Nucleosides Nucleotides Nucleic Acids, 2009. **28**(2): p. 78-88.
73. Li, J.B., et al., *Genome-wide identification of human RNA editing sites by parallel DNA capturing and sequencing.* Science, 2009. **324**(5931): p. 1210-3.
74. Harjes, E., et al., *Structure of the RNA claw of the DNA packaging motor of bacteriophage Phi29.* Nucleic Acids Res, 2012. **40**(19): p. 9953-63.
75. Zhang, Q., et al., *Structurally conserved five nucleotide bulge determines the overall topology of the core domain of human telomerase RNA.* Proc Natl Acad Sci U S A, 2010. **107**(44): p. 18761-8.
76. Wang, Y., J. Havel, and P.A. Beal, *A Phenotypic Screen for Functional Mutants of Human Adenosine Deaminase Acting on RNA 1.* ACS Chem Biol, 2015. **10**(11): p. 2512-9.
77. Rice, G.I., et al., *Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature.* Nat Genet, 2012. **44**(11): p. 1243-8.

78. Tojo, K., et al., *Dystonia, mental deterioration, and dyschromatosis symmetrica hereditaria in a family with ADAR1 mutation.* Mov Disord, 2006. **21**(9): p. 1510-3.
79. Wong, S.K. and D.W. Lazinski, *Replicating hepatitis delta virus RNA is edited in the nucleus by the small form of ADAR1.* Proc Natl Acad Sci U S A, 2002. **99**(23): p. 15118-23.
80. Azad, M.T.A., U. Qulsum, and T. Tsukahara, *Comparative Activity of Adenosine Deaminase Acting on RNA (ADARs) Isoforms for Correction of Genetic Code in Gene Therapy.* Curr Gene Ther, 2019. **19**(1): p. 31-39.
81. Lai, F., et al., *Dramatic increase of the RNA editing for glutamate receptor subunits during terminal differentiation of clonal human neurons.* J Neurochem, 1997. **69**(1): p. 43-52.
82. Maas, S., et al., *Underediting of glutamate receptor GluR-B mRNA in malignant gliomas.* Proc Natl Acad Sci U S A, 2001. **98**(25): p. 14687-92.
83. Wahlstedt, H., et al., *Large-scale mRNA sequencing determines global regulation of RNA editing during brain development.* Genome Res, 2009. **19**(6): p. 978-86.
84. Chen, C.X., et al., *A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains.* RNA, 2000. **6**(5): p. 755-67.
85. Maas, S. and W.M. Gommans, *Novel exon of mammalian ADAR2 extends open reading frame.* PLoS One, 2009. **4**(1): p. e4225.
86. Elias, Y. and R.H. Huang, *Biochemical and structural studies of A-to-I editing by tRNA:A34 deaminases at the wobble position of transfer RNA.* Biochemistry, 2005. **44**(36): p. 12057-65.
87. Roura Frigole, H., et al., *tRNA deamination by ADAT requires substrate-specific recognition mechanisms and can be inhibited by tRFs.* RNA, 2019. **25**(5): p. 607-619.
88. Auxilien, S., et al., *Mechanism, specificity and general properties of the yeast enzyme catalysing the formation of inosine 34 in the anticodon of transfer RNA.* J Mol Biol, 1996. **262**(4): p. 437-58.
89. Saint-Leger, A., et al., *Saturation of recognition elements blocks evolution of new tRNA identities.* Sci Adv, 2016. **2**(4): p. e1501860.
90. Balasubramaian, R. and P. Seetharamulu, *A conformational rationale for the wobble behaviour of the first base of the anticodon triplet in tRNA.* J Theor Biol, 1983. **101**(1): p. 77-86.
91. Schaub, M. and W. Keller, *RNA editing by adenosine deaminases generates RNA and protein diversity.* Biochimie, 2002. **84**(8): p. 791-803.
92. Torres, A.G., et al., *Inosine modifications in human tRNAs are incorporated at the precursor tRNA level.* Nucleic Acids Res, 2015. **43**(10): p. 5145-57.
93. Rubio Gomez, M.A. and M. Ibba, *Aminoacyl-tRNA synthetases.* RNA, 2020. **26**(8): p. 910-936.
94. Li, J., et al., *APOBEC3 multimerization correlates with HIV-1 packaging and restriction activity in living cells.* J Mol Biol, 2014. **426**(6): p. 1296-307.
95. Brar, S.S., et al., *Activation-induced deaminase, AID, is catalytically active as a monomer on single-stranded DNA.* DNA Repair (Amst), 2008. **7**(1): p. 77-87.
96. Byeon, I.J., et al., *NMR structure of human restriction factor APOBEC3A reveals substrate binding and enzyme specificity.* Nat Commun, 2013. **4**: p. 1890.
97. Opi, S., et al., *Monomeric APOBEC3G is catalytically active and has antiviral activity.* J Virol, 2006. **80**(10): p. 4673-82.

98. Kouno, T., et al., *Crystal structure of APOBEC3A bound to single-stranded DNA reveals structural basis for cytidine deamination and specificity.* Nat Commun, 2017. **8**: p. 15024.
99. Logue, E.C., et al., *A DNA sequence recognition loop on APOBEC3A controls substrate specificity.* PLoS One, 2014. **9**(5): p. e97062.
100. Krzysiak, T.C., et al., *APOBEC2 is a monomer in solution: implications for APOBEC3G models.* Biochemistry, 2012. **51**(9): p. 2008-17.
101. Thuy-Boun, A.S., et al., *Asymmetric dimerization of adenosine deaminase acting on RNA facilitates substrate recognition.* Nucleic Acids Res, 2020. **48**(14): p. 7958-7972.
102. Jaikaran, D.C., C.H. Collins, and A.M. MacMillan, *Adenosine to inosine editing by ADAR2 requires formation of a ternary complex on the GluR-B R/G site.* J Biol Chem, 2002. **277**(40): p. 37624-9.
103. Chilibeck, K.A., et al., *FRET analysis of in vivo dimerization by RNA-editing enzymes.* J Biol Chem, 2006. **281**(24): p. 16530-5.
104. Gallo, A., et al., *An ADAR that edits transcripts encoding ion channel subunits functions as a dimer.* EMBO J, 2003. **22**(13): p. 3421-30.
105. Cho, D.S., et al., *Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA.* J Biol Chem, 2003. **278**(19): p. 17093-102.
106. Sansam, C.L., K.S. Wells, and R.B. Emeson, *Modulation of RNA editing by functional nucleolar sequestration of ADAR2.* Proc Natl Acad Sci U S A, 2003. **100**(24): p. 14018-23.
107. Valente, L. and K. Nishikura, *RNA binding-independent dimerization of adenosine deaminases acting on RNA and dominant negative effects of nonfunctional subunits on dimer functions.* J Biol Chem, 2007. **282**(22): p. 16054-61.
108. Ramos, A., et al., *RNA recognition by a Staufen double-stranded RNA-binding domain.* EMBO J, 2000. **19**(5): p. 997-1009.
109. Ryter, J.M. and S.C. Schultz, *Molecular basis of double-stranded RNA-protein interactions: structure of a dsRNA-binding domain complexed with dsRNA.* EMBO J, 1998. **17**(24): p. 7505-13.
110. Gerber, A.P. and W. Keller, *An adenosine deaminase that generates inosine at the wobble position of tRNAs.* Science, 1999. **286**(5442): p. 1146-9.
111. Binder, J.X., et al., *COMPARTMENTS: unification and visualization of protein subcellular localization evidence.* Database (Oxford), 2014. **2014**: p. bau012.
112. Desterro, J.M., et al., *Dynamic association of RNA-editing enzymes with the nucleolus.* J Cell Sci, 2003. **116**(Pt 9): p. 1805-18.
113. Poulsen, H., et al., *CRM1 mediates the export of ADAR1 through a nuclear export signal within the Z-DNA binding domain.* Mol Cell Biol, 2001. **21**(22): p. 7862-71.
114. Strehblow, A., M. Hallegger, and M.F. Jantsch, *Nucleocytoplasmic distribution of human RNA-editing enzyme ADAR1 is modulated by double-stranded RNA-binding domains, a leucine-rich export signal, and a putative dimerization domain.* Mol Biol Cell, 2002. **13**(11): p. 3822-35.
115. Fritz, J., et al., *RNA-regulated interaction of transportin-1 and exportin-5 with the double-stranded RNA-binding domain regulates nucleocytoplasmic shuttling of ADAR1.* Mol Cell Biol, 2009. **29**(6): p. 1487-97.
116. Pinol-Roma, S. and G. Dreyfuss, *Shuttling of pre-mRNA binding proteins between nucleus and cytoplasm.* Nature, 1992. **355**(6362): p. 730-2.

117. Michael, W.M., M. Choi, and G. Dreyfuss, *A nuclear export signal in hnRNP A1: a signal-mediated, temperature-dependent nuclear protein export pathway.* Cell, 1995. **83**(3): p. 415-22.
118. Eckmann, C.R., et al., *The human but not the Xenopus RNA-editing enzyme ADAR1 has an atypical nuclear localization signal and displays the characteristics of a shuttling protein.* Mol Biol Cell, 2001. **12**(7): p. 1911-24.
119. Maas, S. and W.M. Gommans, *Identification of a selective nuclear import signal in adenosine deaminases acting on RNA.* Nucleic Acids Res, 2009. **37**(17): p. 5822-9.
120. Enstero, M., et al., *Recognition and coupling of A-to-I edited sites are determined by the tertiary structure of the RNA.* Nucleic Acids Res, 2009. **37**(20): p. 6916-26.
121. Gommans, W.M., et al., *A mammalian reporter system for fast and quantitative detection of intracellular A-to-I RNA editing levels.* Anal Biochem, 2010. **399**(2): p. 230-6.
122. Lehmann, K.A. and B.L. Bass, *The importance of internal loops within RNA substrates of ADAR1.* J Mol Biol, 1999. **291**(1): p. 1-13.
123. Daniel, C., et al., *Editing inducer elements increases A-to-I editing efficiency in the mammalian transcriptome.* Genome Biol, 2017. **18**(1): p. 195.
124. Bhalla, T., et al., *Control of human potassium channel inactivation by editing of a small mRNA hairpin.* Nat Struct Mol Biol, 2004. **11**(10): p. 950-6.
125. Galipon, J., et al., *Differential Binding of Three Major Human ADAR Isoforms to Coding and Long Non-Coding Transcripts.* Genes (Basel), 2017. **8**(2).
126. Ramaswami, G. and J.B. Li, *RADAR: a rigorously annotated database of A-to-I RNA editing.* Nucleic Acids Res, 2014. **42**(Database issue): p. D109-13.
127. Hundley, H.A., A.A. Krauchuk, and B.L. Bass, *C. elegans and H. sapiens mRNAs with edited 3' UTRs are present on polysomes.* RNA, 2008. **14**(10): p. 2050-60.
128. Dawson, T.R., C.L. Sansam, and R.B. Emeson, *Structure and sequence determinants required for the RNA editing of ADAR2 substrates.* J Biol Chem, 2004. **279**(6): p. 4941-51.
129. Brummer, A., et al., *Structure-mediated modulation of mRNA abundance by A-to-I editing.* Nat Commun, 2017. **8**(1): p. 1255.
130. Scadden, A.D., *The RISC subunit Tudor-SN binds to hyper-edited double-stranded RNA and promotes its cleavage.* Nat Struct Mol Biol, 2005. **12**(6): p. 489-96.
131. Xiang, J.F., et al., *N(6)-Methyladenosines Modulate A-to-I RNA Editing.* Mol Cell, 2018. **69**(1): p. 126-135 e6.
132. Yang, C., et al., *The role of m(6)A modification in physiology and disease.* Cell Death Dis, 2020. **11**(11): p. 960.
133. Veliz, E.A., L.M. Easterwood, and P.A. Beal, *Substrate analogues for an RNA-editing adenosine deaminase: mechanistic investigation and inhibitor design.* J Am Chem Soc, 2003. **125**(36): p. 10867-76.
134. Hoang, H.D., et al., *Emerging translation strategies during virus-host interaction.* Wiley Interdiscip Rev RNA, 2020: p. e1619.
135. Mizrahi, R.A., et al., *A Fluorescent Adenosine Analogue as a Substrate for an A-to-I RNA Editing Enzyme.* Angew Chem Int Ed Engl, 2015. **54**(30): p. 8713-6.
136. Yi-Brunozzi, H.Y., et al., *Synthetic substrate analogs for the RNA-editing adenosine deaminase ADAR-2.* Nucleic Acids Res, 1999. **27**(14): p. 2912-7.
137. Goto-Ito, S., T. Ito, and S. Yokoyama, *Trm5 and TrmD: Two Enzymes from Distinct Origins Catalyze the Identical tRNA Modification, m(1)G37.* Biomolecules, 2017. **7**(1).

138. Chan, P.P. and T.M. Lowe, *GtRNAdb 2.0: an expanded database of transfer RNA genes identified in complete and draft genomes.* Nucleic Acids Res, 2016. **44**(D1): p. D184-9.
139. Evans, M.E., et al., *Determination of tRNA aminoacylation levels by high-throughput sequencing.* Nucleic Acids Res, 2017. **45**(14): p. e133.
140. Rafels-Ybern, A., C.S. Attolini, and L. Ribas de Pouplana, *Distribution of ADAT-Dependent Codons in the Human Transcriptome.* Int J Mol Sci, 2015. **16**(8): p. 17303-14.
141. Rafels-Ybern, A., et al., *Codon adaptation to tRNAs with Inosine modification at position 34 is widespread among Eukaryotes and present in two Bacterial phyla.* RNA Biol, 2018. **15**(4-5): p. 500-507.
142. Novoa, E.M., et al., *A role for tRNA modifications in genome structure and codon usage.* Cell, 2012. **149**(1): p. 202-13.
143. Rafels-Ybern, A., et al., *The Expansion of Inosine at the Wobble Position of tRNAs, and Its Role in the Evolution of Proteomes.* Mol Biol Evol, 2019. **36**(4): p. 650-662.
144. Ou, X., et al., *Errors in translational decoding: tRNA wobbling or misincorporation?* PLoS Genet, 2019. **15**(3): p. e1008017.
145. Lei, L. and Z.F. Burton, *Evolution of Life on Earth: tRNA, Aminoacyl-tRNA Synthetases and the Genetic Code.* Life (Basel), 2020. **10**(3).
146. Athey, J., et al., *A new and updated resource for codon usage tables.* BMC Bioinformatics, 2017. **18**(1): p. 391.
147. Perche-Letuvee, P., et al., *Wybutosine biosynthesis: structural and mechanistic overview.* RNA Biol, 2014. **11**(12): p. 1508-18.
148. Ohira, T. and T. Suzuki, *Retrograde nuclear import of tRNA precursors is required for modified base biogenesis in yeast.* Proc Natl Acad Sci U S A, 2011. **108**(26): p. 10502-7.
149. Droogmans, L. and H. Grosjean, *Enzymatic conversion of guanosine 3' adjacent to the anticodon of yeast tRNAPhe to N1-methylguanosine and the wye nucleoside: dependence on the anticodon sequence.* EMBO J, 1987. **6**(2): p. 477-83.
150. Haumont, E., et al., *Enzymatic conversion of adenosine to inosine in the wobble position of yeast tRNAAsp: the dependence on the anticodon sequence.* Nucleic Acids Res, 1984. **12**(6): p. 2705-15.
151. Asselah, T. and M. Rizzetto, *Hepatitis D Virus Infection.* N Engl J Med, 2023. **389**(1): p. 58-70.
152. Caviglia, G.P., A. Ciancio, and M. Rizzetto, *A Review of HDV Infection.* Viruses, 2022. **14**(8).
153. Negro, F. and A.S. Lok, *Hepatitis D: A Review.* JAMA, 2023. **330**(24): p. 2376-2387.
154. Mentha, N., et al., *A review on hepatitis D: From virology to new therapies.* J Adv Res, 2019. **17**: p. 3-15.
155. Burwitz, B.J., Z. Zhou, and W. Li, *Animal models for the study of human hepatitis B and D virus infection: New insights and progress.* Antiviral Res, 2020. **182**: p. 104898.
156. Sato, S., S.K. Wong, and D.W. Lazinski, *Hepatitis delta virus minimal substrates competent for editing by ADAR1 and ADAR2.* J Virol, 2001. **75**(18): p. 8547-55.
157. Weiner, A.J., et al., *A single antigenomic open reading frame of the hepatitis delta virus encodes the epitope(s) of both hepatitis delta antigen polypeptides p24 delta and p27 delta.* J Virol, 1988. **62**(2): p. 594-9.
158. Wu, T.T., et al., *Hepatitis delta virus mutant: effect on RNA editing.* J Virol, 1995. **69**(11): p. 7226-31.

159. Casey, J.L. and J.L. Gerin, *Hepatitis D virus RNA editing: specific modification of adenosine in the antigenomic RNA.* J Virol, 1995. **69**(12): p. 7593-600.
160. Polson, A.G., et al., *Hepatitis delta virus RNA editing is highly specific for the amber/W site and is suppressed by hepatitis delta antigen.* Mol Cell Biol, 1998. **18**(4): p. 1919-26.
161. Bergmann, K.F. and J.L. Gerin, *Antigens of hepatitis delta virus in the liver and serum of humans and animals.* J Infect Dis, 1986. **154**(4): p. 702-6.
162. Casey, J.L., et al., *Structural requirements for RNA editing in hepatitis delta virus: evidence for a uridine-to-cytidine editing mechanism.* Proc Natl Acad Sci U S A, 1992. **89**(15): p. 7149-53.
163. Luo, G.X., et al., *A specific base transition occurs on replicating hepatitis delta virus RNA.* J Virol, 1990. **64**(3): p. 1021-7.
164. Kuo, M.Y., et al., *Molecular cloning of hepatitis delta virus RNA from an infected woodchuck liver: sequence, structure, and applications.* J Virol, 1988. **62**(6): p. 1855-61.
165. Wang, K.S., et al., *Structure, sequence and expression of the hepatitis delta (delta) viral genome.* Nature, 1986. **323**(6088): p. 508-14.
166. Wang, J.G., J. Cullen, and S.M. Lemon, *Immunoblot analysis demonstrates that the large and small forms of hepatitis delta virus antigen have different C-terminal amino acid sequences.* J Gen Virol, 1992. **73 (Pt 1)**: p. 183-8.
167. Chao, M., S.Y. Hsieh, and J. Taylor, *Role of two forms of hepatitis delta virus antigen: evidence for a mechanism of self-limiting genome replication.* J Virol, 1990. **64**(10): p. 5066-9.
168. Ryu, W.S., M. Bayer, and J. Taylor, *Assembly of hepatitis delta virus particles.* J Virol, 1992. **66**(4): p. 2310-5.
169. Lai, M.M., *The molecular biology of hepatitis delta virus.* Annu Rev Biochem, 1995. **64**: p. 259-86.
170. Kuo, M.Y., M. Chao, and J. Taylor, *Initiation of replication of the human hepatitis delta virus genome from cloned DNA: role of delta antigen.* J Virol, 1989. **63**(5): p. 1945-50.
171. Chang, F.L., et al., *The large form of hepatitis delta antigen is crucial for assembly of hepatitis delta virus.* Proc Natl Acad Sci U S A, 1991. **88**(19): p. 8490-4.
172. Hsieh, S.Y., et al., *Hepatitis delta virus genome replication: a polyadenylated mRNA for delta antigen.* J Virol, 1990. **64**(7): p. 3192-8.
173. Ryu, W.S., et al., *Ribonucleoprotein complexes of hepatitis delta virus.* J Virol, 1993. **67**(6): p. 3281-7.
174. Taylor, J.M., *Hepatitis delta virus: cis and trans functions required for replication.* Cell, 1990. **61**(3): p. 371-3.
175. Tavanez, J.P., et al., *Hepatitis delta virus ribonucleoproteins shuttle between the nucleus and the cytoplasm.* RNA, 2002. **8**(5): p. 637-46.
176. Gudima, S., et al., *Parameters of human hepatitis delta virus genome replication: the quantity, quality, and intracellular distribution of viral proteins and RNA.* J Virol, 2002. **76**(8): p. 3709-19.
177. Macnaughton, T.B. and M.M. Lai, *Genomic but not antigenomic hepatitis delta virus RNA is preferentially exported from the nucleus immediately after synthesis and processing.* J Virol, 2002. **76**(8): p. 3928-35.

178. Jayan, G.C. and J.L. Casey, *Increased RNA editing and inhibition of hepatitis delta virus replication by high-level expression of ADAR1 and ADAR2.* J Virol, 2002. **76**(8): p. 3819-27.
179. Branche, A.R. and A.R. Falsey, *Parainfluenza Virus Infection.* Semin Respir Crit Care Med, 2016. **37**(4): p. 538-54.
180. Vidal, S., J. Curran, and D. Kolakofsky, *Editing of the Sendai virus P/C mRNA by G insertion occurs during mRNA synthesis via a virus-encoded activity.* J Virol, 1990. **64**(1): p. 239-46.
181. Galinski, M.S., R.M. Troy, and A.K. Banerjee, *RNA editing in the phosphoprotein gene of the human parainfluenza virus type 3.* Virology, 1992. **186**(2): p. 543-50.
182. Thomas, S.M., R.A. Lamb, and R.G. Paterson, *Two mRNAs that differ by two nontemplated nucleotides encode the amino coterminal proteins P and V of the paramyxovirus SV5.* Cell, 1988. **54**(6): p. 891-902.
183. Rima, B.K., et al., *Stability of the parainfluenza virus 5 genome revealed by deep sequencing of strains isolated from different hosts and following passage in cell culture.* J Virol, 2014. **88**(7): p. 3826-36.
184. Lee, Y.N. and C. Lee, *Complete genome sequence of a novel porcine parainfluenza virus 5 isolate in Korea.* Arch Virol, 2013. **158**(8): p. 1765-72.
185. Atoynatan, T. and G.D. Hsiung, *Epidemiologic studies of latent virus infections in captive monkeys and baboons. II. Serologic evidence of myxovirus infections with special reference to SV5.* Am J Epidemiol, 1969. **89**(4): p. 472-9.
186. Hsiung, G.D., *Parainfluenza-5 virus. Infection of man and animal.* Prog Med Virol, 1972. **14**: p. 241-74.
187. Tribe, G.W., *An investigation of the incidence, epidemiology and control of Simian virus 5.* Br J Exp Pathol, 1966. **47**(5): p. 472-9.
188. Goswami, K.K., et al., *Does simian virus 5 infect humans?* J Gen Virol, 1984. **65 (Pt 8)**: p. 1295-303.
189. Chatziandreou, N., et al., *Relationships and host range of human, canine, simian and porcine isolates of simian virus 5 (parainfluenza virus 5).* J Gen Virol, 2004. **85**(Pt 10): p. 3007-3016.
190. Murray, M.J., *Ebola Virus Disease: A Review of Its Past and Present.* Anesth Analg, 2015. **121**(3): p. 798-809.
191. Baseler, L., et al., *The Pathogenesis of Ebola Virus Disease.* Annu Rev Pathol, 2017. **12**: p. 387-418.
192. Sanchez, A., et al., *The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing.* Proc Natl Acad Sci U S A, 1996. **93**(8): p. 3602-7.
193. Feldmann, H., H.D. Klenk, and A. Sanchez, *Molecular biology and evolution of filoviruses.* Arch Virol Suppl, 1993. **7**: p. 81-100.
194. Whitfield, Z.J., et al., *Species-Specific Evolution of Ebola Virus during Replication in Human and Bat Cells.* Cell Rep, 2020. **32**(7): p. 108028.
195. Khrustalev, V.V., E.V. Barkovsky, and T.A. Khrustaleva, *Local Mutational Pressures in Genomes of Zaire Ebolavirus and Marburg Virus.* Adv Bioinformatics, 2015. **2015**: p. 678587.
196. Moss, W.J., *Measles.* Lancet, 2017. **390**(10111): p. 2490-2502.
197. Amurri, L., et al., *Measles Virus-Induced Host Immunity and Mechanisms of Viral Evasion.* Viruses, 2022. **14**(12).

198. Cattaneo, R., et al., *Measles virus editing provides an additional cysteine-rich protein.* Cell, 1989. **56**(5): p. 759-64.
199. Patterson, J.B., et al., *Evidence that the hypermutated M protein of a subacute sclerosing panencephalitis measles virus actively contributes to the chronic progressive CNS disease.* Virology, 2001. **291**(2): p. 215-25.
200. Wong, T.C., et al., *Generalized and localized biased hypermutation affecting the matrix gene of a measles virus strain that causes subacute sclerosing panencephalitis.* J Virol, 1989. **63**(12): p. 5464-8.
201. Wong, T.C., et al., *Role of biased hypermutation in evolution of subacute sclerosing panencephalitis virus from progenitor acute measles virus.* J Virol, 1991. **65**(5): p. 2191-9.
202. Cattaneo, R., et al., *Biased hypermutation and other genetic changes in defective measles viruses in human brain infections.* Cell, 1988. **55**(2): p. 255-65.
203. Suspene, R., et al., *Double-stranded RNA adenosine deaminase ADAR-1-induced hypermutated genomes among inactivated seasonal influenza and live attenuated measles virus vaccines.* J Virol, 2011. **85**(5): p. 2458-62.
204. Fehrholz, M., et al., *The innate antiviral factor APOBEC3G targets replication of measles, mumps and respiratory syncytial viruses.* J Gen Virol, 2012. **93**(Pt 3): p. 565-576.
205. Melikyan, G.B., *HIV entry: a game of hide-and-fuse?* Curr Opin Virol, 2014. **4**: p. 1-7.
206. Fanales-Belasio, E., et al., *HIV virology and pathogenetic mechanisms of infection: a brief overview.* Ann Ist Super Sanita, 2010. **46**(1): p. 5-14.
207. Kingsman, S.M. and A.J. Kingsman, *The regulation of human immunodeficiency virus type-1 gene expression.* Eur J Biochem, 1996. **240**(3): p. 491-507.
208. Doria, M., et al., *Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection.* Nucleic Acids Res, 2009. **37**(17): p. 5848-58.
209. Phuphuakrat, A., et al., *Double-stranded RNA adenosine deaminases enhance expression of human immunodeficiency virus type 1 proteins.* J Virol, 2008. **82**(21): p. 10864-72.
210. Bannwarth, S. and A. Gatignol, *HIV-1 TAR RNA: the target of molecular interactions between the virus and its host.* Curr HIV Res, 2005. **3**(1): p. 61-71.
211. Laughrea, M. and L. Jette, *A 19-nucleotide sequence upstream of the 5' major splice donor is part of the dimerization domain of human immunodeficiency virus 1 genomic RNA.* Biochemistry, 1994. **33**(45): p. 13464-74.
212. Jakobovits, A., et al., *A discrete element 3' of human immunodeficiency virus 1 (HIV-1) and HIV-2 mRNA initiation sites mediates transcriptional activation by an HIV trans activator.* Mol Cell Biol, 1988. **8**(6): p. 2555-61.
213. Laspia, M.F., A.P. Rice, and M.B. Mathews, *HIV-1 Tat protein increases transcriptional initiation and stabilizes elongation.* Cell, 1989. **59**(2): p. 283-92.
214. Rosen, C.A., et al., *Post-transcriptional regulation accounts for the trans-activation of the human T-lymphotropic virus type III.* Nature, 1986. **319**(6054): p. 555-9.
215. Kao, S.Y., et al., *Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product.* Nature, 1987. **330**(6147): p. 489-93.
216. Feng, S. and E.C. Holland, *HIV-1 tat trans-activation requires the loop sequence within tar.* Nature, 1988. **334**(6178): p. 165-7.
217. Hauber, J. and B.R. Cullen, *Mutational analysis of the trans-activation-responsive region of the human immunodeficiency virus type I long terminal repeat.* J Virol, 1988. **62**(3): p. 673-9.

218. Muesing, M.A., D.H. Smith, and D.J. Capon, *Regulation of mRNA accumulation by a human immunodeficiency virus trans-activator protein.* Cell, 1987. **48**(4): p. 691-701.
219. Sharmeen, L., et al., *Tat-dependent adenosine-to-inosine modification of wild-type transactivation response RNA.* Proc Natl Acad Sci U S A, 1991. **88**(18): p. 8096-100.
220. Hartwig, D., et al., *The large form of ADAR 1 is responsible for enhanced hepatitis delta virus RNA editing in interferon-alpha-stimulated host cells.* J Viral Hepat, 2006. **13**(3): p. 150-7.
221. Lazinski, D.W. and J.M. Taylor, *Recent developments in hepatitis delta virus research.* Adv Virus Res, 1994. **43**: p. 187-231.
222. Casey, J.L. and J.L. Gerin, *Genotype-specific complementation of hepatitis delta virus RNA replication by hepatitis delta antigen.* J Virol, 1998. **72**(4): p. 2806-14.
223. Casey, J.L., et al., *A genotype of hepatitis D virus that occurs in northern South America.* Proc Natl Acad Sci U S A, 1993. **90**(19): p. 9016-20.
224. Casey, J.L., et al., *Hepatitis B virus (HBV)/hepatitis D virus (HDV) coinfection in outbreaks of acute hepatitis in the Peruvian Amazon basin: the roles of HDV genotype III and HBV genotype F.* J Infect Dis, 1996. **174**(5): p. 920-6.
225. Casey, J.L., *RNA editing in hepatitis delta virus genotype III requires a branched double-hairpin RNA structure.* J Virol, 2002. **76**(15): p. 7385-97.
226. Long, J.S., et al., *Host and viral determinants of influenza A virus species specificity.* Nat Rev Microbiol, 2019. **17**(2): p. 67-81.
227. Spackman, E., *A Brief Introduction to Avian Influenza Virus.* Methods Mol Biol, 2020. **2123**: p. 83-92.
228. Cao, Y., et al., *A comprehensive study on cellular RNA editing activity in response to infections with different subtypes of influenza a viruses.* BMC Genomics, 2018. **19**(Suppl 1): p. 925.
229. Yin, J.K., et al., *The threat of human influenza: the viruses, disease impacts, and vaccine solutions.* Infect Disord Drug Targets, 2014. **14**(3): p. 150-4.
230. Vincent, A., et al., *Review of influenza A virus in swine worldwide: a call for increased surveillance and research.* Zoonoses Public Health, 2014. **61**(1): p. 4-17.
231. Poovorawan, Y., et al., *Global alert to avian influenza virus infection: from H5N1 to H7N9.* Pathog Glob Health, 2013. **107**(5): p. 217-23.
232. Tong, S., et al., *New world bats harbor diverse influenza A viruses.* PLoS Pathog, 2013. **9**(10): p. e1003657.
233. Bouvier, N.M. and P. Palese, *The biology of influenza viruses.* Vaccine, 2008. **26 Suppl 4**: p. D49-53.
234. Sueoka, N., *Directional mutation pressure, mutator mutations, and dynamics of molecular evolution.* J Mol Evol, 1993. **37**(2): p. 137-53.
235. Cuevas, J.M., et al., *Human norovirus hyper-mutation revealed by ultra-deep sequencing.* Infect Genet Evol, 2016. **41**: p. 233-239.
236. Khrustalev, V.V., et al., *Mutational Pressure in Zika Virus: Local ADAR-Editing Areas Associated with Pauses in Translation and Replication.* Front Cell Infect Microbiol, 2017. **7**: p. 44.
237. Albin, J.S. and R.S. Harris, *Interactions of host APOBEC3 restriction factors with HIV-1 in vivo: implications for therapeutics.* Expert Rev Mol Med, 2010. **12**: p. e4.
238. Aruscavage, P.J. and B.L. Bass, *A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing.* RNA, 2000. **6**(2): p. 257-69.
239. Herbert, A., et al., *Chicken double-stranded RNA adenosine deaminase has apparent specificity for Z-DNA.* Proc Natl Acad Sci U S A, 1995. **92**(16): p. 7550-4.

240. Gutierrez, R.A., et al., *Biased mutational pattern and quasispecies hypothesis in H5N1 virus.* Infect Genet Evol, 2013. **15**: p. 69-76.
241. Emmott, E., et al., *Quantitative proteomics using SILAC coupled to LC-MS/MS reveals changes in the nucleolar proteome in influenza A virus-infected cells.* J Proteome Res, 2010. **9**(10): p. 5335-45.
242. Ngamurulert, S., T. Limjindaporn, and P. Auewaraku, *Identification of cellular partners of Influenza A virus (H5N1) non-structural protein NS1 by yeast two-hybrid system.* Acta Virol, 2009. **53**(3): p. 153-9.
243. Wang, X., et al., *Influenza A virus NS1 protein prevents activation of NF-kappaB and induction of alpha/beta interferon.* J Virol, 2000. **74**(24): p. 11566-73.
244. Labudova, M., J. Pastorek, and S. Pastorekova, *Lymphocytic choriomeningitis virus: ways to establish and maintain non-cytolytic persistent infection.* Acta Virol, 2016. **60**(1): p. 15-26.
245. Zahn, R.C., et al., *A-to-G hypermutation in the genome of lymphocytic choriomeningitis virus.* J Virol, 2007. **81**(2): p. 457-64.
246. Langland, J.O., et al., *Inhibition of PKR by RNA and DNA viruses.* Virus Res, 2006. **119**(1): p. 100-10.
247. Griffiths, C., S.J. Drews, and D.J. Marchant, *Respiratory Syncytial Virus: Infection, Detection, and New Options for Prevention and Treatment.* Clin Microbiol Rev, 2017. **30**(1): p. 277-319.
248. Borchers, A.T., et al., *Respiratory syncytial virus--a comprehensive review.* Clin Rev Allergy Immunol, 2013. **45**(3): p. 331-79.
249. Levine, S., R. Klaiber-Franco, and P.R. Paradiso, *Demonstration that glycoprotein G is the attachment protein of respiratory syncytial virus.* J Gen Virol, 1987. **68 (Pt 9)**: p. 2521-4.
250. Anderson, L.J., et al., *Antigenic characterization of respiratory syncytial virus strains with monoclonal antibodies.* J Infect Dis, 1985. **151**(4): p. 626-33.
251. Mufson, M.A., et al., *Two distinct subtypes of human respiratory syncytial virus.* J Gen Virol, 1985. **66 (Pt 10)**: p. 2111-24.
252. Garcia-Barreno, B., et al., *Marked differences in the antigenic structure of human respiratory syncytial virus F and G glycoproteins.* J Virol, 1989. **63**(2): p. 925-32.
253. Cristina, J., et al., *Analysis of genetic variability in human respiratory syncytial virus by the RNase A mismatch cleavage method: subtype divergence and heterogeneity.* Virology, 1990. **174**(1): p. 126-34.
254. Cane, P.A., D.A. Matthews, and C.R. Pringle, *Identification of variable domains of the attachment (G) protein of subgroup A respiratory syncytial viruses.* J Gen Virol, 1991. **72 (Pt 9)**: p. 2091-6.
255. Sullender, W.M., et al., *Genetic diversity of the attachment protein of subgroup B respiratory syncytial viruses.* J Virol, 1991. **65**(10): p. 5425-34.
256. Garcia, O., et al., *Evolutionary pattern of human respiratory syncytial virus (subgroup A): cocirculating lineages and correlation of genetic and antigenic changes in the G glycoprotein.* J Virol, 1994. **68**(9): p. 5448-59.
257. Cane, P.A. and C.R. Pringle, *Evolution of subgroup A respiratory syncytial virus: evidence for progressive accumulation of amino acid changes in the attachment protein.* J Virol, 1995. **69**(5): p. 2918-25.
258. Mart Nez, I., et al., *Evolutionary pattern of the G glycoprotein of human respiratory syncytial viruses from antigenic group B: the use of alternative termination codons and lineage diversification.* J Gen Virol, 1999. **80 (Pt 1)**: p. 125-130.

259. Martinez, I. and J.A. Melero, *A model for the generation of multiple A to G transitions in the human respiratory syncytial virus genome: predicted RNA secondary structures as substrates for adenosine deaminases that act on RNA.* J Gen Virol, 2002. **83**(Pt 6): p. 1445-1455.
260. Vos, L.M., et al., *High epidemic burden of RSV disease coinciding with genetic alterations causing amino acid substitutions in the RSV G-protein during the 2016/2017 season in The Netherlands.* J Clin Virol, 2019. **112**: p. 20-26.
261. Lee, J., et al., *Protective antigenic sites in respiratory syncytial virus G attachment protein outside the central conserved and cysteine noose domains.* PLoS Pathog, 2018. **14**(8): p. e1007262.
262. de Graaf, M., J. van Beek, and M.P. Koopmans, *Human norovirus transmission and evolution in a changing world.* Nat Rev Microbiol, 2016. **14**(7): p. 421-33.
263. Smith, E.C. and M.R. Denison, *Coronaviruses as DNA wannabes: a new model for the regulation of RNA virus replication fidelity.* PLoS Pathog, 2013. **9**(12): p. e1003760.
264. Ulferts, R. and J. Ziebuhr, *Nidovirus ribonucleases: Structures and functions in viral replication.* RNA Biol, 2011. **8**(2): p. 295-304.
265. Debbink, K., et al., *Norovirus immunity and the great escape.* PLoS Pathog, 2012. **8**(10): p. e1002921.
266. White, P.A., *Evolution of norovirus.* Clin Microbiol Infect, 2014. **20**(8): p. 741-5.
267. Lindesmith, L.C., et al., *Mechanisms of GII.4 norovirus persistence in human populations.* PLoS Med, 2008. **5**(2): p. e31.
268. Martinez, M.P., J. Al-Saleem, and P.L. Green, *Comparative virology of HTLV-1 and HTLV-2.* Retrovirology, 2019. **16**(1): p. 21.
269. Stufano, A., et al., *Work-Related Human T-lymphotropic Virus 1 and 2 (HTLV-1/2) Infection: A Systematic Review.* Viruses, 2021. **13**(9).
270. Zhang, L.L., et al., *Human T-cell lymphotropic virus type 1 and its oncogenesis.* Acta Pharmacol Sin, 2017. **38**(8): p. 1093-1103.
271. Ko, N.L., et al., *Hyperediting of human T-cell leukemia virus type 2 and simian T-cell leukemia virus type 3 by the dsRNA adenosine deaminase ADAR-1.* J Gen Virol, 2012. **93**(Pt 12): p. 2646-2651.
272. Mahieux, R. and A. Gessain, *HTLV-3/STLV-3 and HTLV-4 viruses: discovery, epidemiology, serology and molecular aspects.* Viruses, 2011. **3**(7): p. 1074-90.
273. Slattery, J.P., G. Franchini, and A. Gessain, *Genomic evolution, patterns of global dissemination, and interspecies transmission of human and simian T-cell leukemia/lymphotropic viruses.* Genome Res, 1999. **9**(6): p. 525-40.
274. Cox, R.M. and R.K. Plemper, *Structure and organization of paramyxovirus particles.* Curr Opin Virol, 2017. **24**: p. 105-114.
275. Rubin, S., et al., *Molecular biology, pathogenesis and pathology of mumps virus.* J Pathol, 2015. **235**(2): p. 242-52.
276. Paterson, R.G. and R.A. Lamb, *RNA editing by G-nucleotide insertion in mumps virus P-gene mRNA transcripts.* J Virol, 1990. **64**(9): p. 4137-45.
277. Baud, D., et al., *An update on Zika virus infection.* Lancet, 2017. **390**(10107): p. 2099-2109.
278. Musso, D. and D.J. Gubler, *Zika Virus.* Clin Microbiol Rev, 2016. **29**(3): p. 487-524.
279. Plourde, A.R. and E.M. Bloch, *A Literature Review of Zika Virus.* Emerg Infect Dis, 2016. **22**(7): p. 1185-92.
280. Piontkivska, H., et al., *RNA editing by the host ADAR system affects the molecular evolution of the Zika virus.* Ecol Evol, 2017. **7**(12): p. 4475-4485.

281. Dana, A. and T. Tuller, *The effect of tRNA levels on decoding times of mRNA codons.* Nucleic Acids Res, 2014. **42**(14): p. 9171-81.
282. Wolin, S.L. and P. Walter, *Ribosome pausing and stacking during translation of a eukaryotic mRNA.* Embo j, 1988. **7**(11): p. 3559-69.
283. Cea, V., L. Cipolla, and S. Sabbioneda, *Replication of Structured DNA and its implication in epigenetic stability.* Front Genet, 2015. **6**: p. 209.
284. Vilfan, I.D., et al., *Reinitiated viral RNA-dependent RNA polymerase resumes replication at a reduced rate.* Nucleic Acids Res, 2008. **36**(22): p. 7059-67.
285. Damania, B., S.C. Kenney, and N. Raab-Traub, *Epstein-Barr virus: Biology and clinical disease.* Cell, 2022. **185**(20): p. 3652-3670.
286. Yu, H. and E.S. Robertson, *Epstein-Barr Virus History and Pathogenesis.* Viruses, 2023. **15**(3).
287. Cao, S., et al., *New Noncoding Lytic Transcripts Derived from the Epstein-Barr Virus Latency Origin of Replication, oriP, Are Hyperedited, Bind the Paraspeckle Protein, NONO/p54nrb, and Support Viral Lytic Transcription.* J Virol, 2015. **89**(14): p. 7120-32.
288. Touitou, R., C. Cochet, and I. Joab, *Transcriptional analysis of the Epstein-Barr virus interleukin-10 homologue during the lytic cycle.* J Gen Virol, 1996. **77 (Pt 6)**: p. 1163-8.
289. Lau, R., J. Middeldorp, and P.J. Farrell, *Epstein-Barr virus gene expression in oral hairy leukoplakia.* Virology, 1993. **195**(2): p. 463-74.
290. Hudson, G.S., et al., *The short unique region of the B95-8 Epstein-Barr virus genome.* Virology, 1985. **147**(1): p. 81-98.
291. Moss, W.N. and J.A. Steitz, *Genome-wide analyses of Epstein-Barr virus reveal conserved RNA structures and a novel stable intronic sequence RNA.* BMC Genomics, 2013. **14**: p. 543.
292. Hundley, H.A. and B.L. Bass, *ADAR editing in double-stranded UTRs and other noncoding RNA sequences.* Trends Biochem Sci, 2010. **35**(7): p. 377-83.
293. Lin, Z., et al., *Whole-genome sequencing of the Akata and Mutu Epstein-Barr virus strains.* J Virol, 2013. **87**(2): p. 1172-82.
294. Iizasa, H., et al., *Editing of Epstein-Barr virus-encoded BART6 microRNAs controls their dicer targeting and consequently affects viral latency.* J Biol Chem, 2010. **285**(43): p. 33358-70.
295. Lei, T., et al., *Perturbation of biogenesis and targeting of Epstein-Barr virus-encoded miR-BART3 microRNA by adenosine-to-inosine editing.* J Gen Virol, 2013. **94**(Pt 12): p. 2739-2744.
296. Schuster, J.E. and J.V. Williams, *Human Metapneumovirus.* Microbiol Spectr, 2014. **2**(5).
297. Divarathna, M.V.M., R.A.M. Rafeek, and F. Noordeen, *A review on epidemiology and impact of human metapneumovirus infections in children using TIAB search strategy on PubMed and PubMed Central articles.* Rev Med Virol, 2020. **30**(1): p. e2090.
298. Huang, A.S., *Defective interfering viruses.* Annu Rev Microbiol, 1973. **27**: p. 101-17.
299. van den Hoogen, B.G., et al., *Excessive production and extreme editing of human metapneumovirus defective interfering RNA is associated with type I IFN induction.* J Gen Virol, 2014. **95**(Pt 8): p. 1625-1633.
300. Banos-Lara Mdel, R., A. Ghosh, and A. Guerrero-Plata, *Critical role of MDA5 in the interferon response induced by human metapneumovirus infection in dendritic cells and in vivo.* J Virol, 2013. **87**(2): p. 1242-51.

301. Bao, X., et al., *Airway epithelial cell response to human metapneumovirus infection.* Virology, 2007. **368**(1): p. 91-101.
302. Liao, S., et al., *Role of retinoic acid inducible gene-I in human metapneumovirus-induced cellular signalling.* J Gen Virol, 2008. **89**(Pt 8): p. 1978-1986.
303. Bao, X., et al., *Human metapneumovirus glycoprotein G inhibits innate immune responses.* PLoS Pathog, 2008. **4**(5): p. e1000077.
304. van Doorn, L.J., *Review: molecular biology of the hepatitis C virus.* J Med Virol, 1994. **43**(4): p. 345-56.
305. Tsukiyama-Kohara, K. and M. Kohara, *Hepatitis C Virus: Viral Quasispecies and Genotypes.* Int J Mol Sci, 2017. **19**(1).
306. Taylor, D.R., et al., *New antiviral pathway that mediates hepatitis C virus replicon interferon sensitivity through ADAR1.* J Virol, 2005. **79**(10): p. 6291-8.
307. Smith, G.L., C. Talbot-Cooper, and Y. Lu, *How Does Vaccinia Virus Interfere With Interferon?* Adv Virus Res, 2018. **100**: p. 355-378.
308. Liu, Y., et al., *Vaccinia virus E3L interferon resistance protein inhibits the interferon-induced adenosine deaminase A-to-I editing activity.* Virology, 2001. **289**(2): p. 378-87.
309. Beattie, E., et al., *Reversal of the interferon-sensitive phenotype of a vaccinia virus lacking E3L by expression of the reovirus S4 gene.* J Virol, 1995. **69**(1): p. 499-505.
310. Beattie, E., J. Tartaglia, and E. Paoletti, *Vaccinia virus-encoded eIF-2 alpha homolog abrogates the antiviral effect of interferon.* Virology, 1991. **183**(1): p. 419-22.
311. Chang, H.W., L.H. Uribe, and B.L. Jacobs, *Rescue of vaccinia virus lacking the E3L gene by mutants of E3L.* J Virol, 1995. **69**(10): p. 6605-8.
312. Davies, M.V., et al., *The E3L and K3L vaccinia virus gene products stimulate translation through inhibition of the double-stranded RNA-dependent protein kinase by different mechanisms.* J Virol, 1993. **67**(3): p. 1688-92.
313. Watson, J.C., H.W. Chang, and B.L. Jacobs, *Characterization of a vaccinia virus-encoded double-stranded RNA-binding protein that may be involved in inhibition of the double-stranded RNA-dependent protein kinase.* Virology, 1991. **185**(1): p. 206-16.
314. Chang, H.W., J.C. Watson, and B.L. Jacobs, *The E3L gene of vaccinia virus encodes an inhibitor of the interferon-induced, double-stranded RNA-dependent protein kinase.* Proc Natl Acad Sci U S A, 1992. **89**(11): p. 4825-9.
315. Jagus, R. and M.M. Gray, *Proteins that interact with PKR.* Biochimie, 1994. **76**(8): p. 779-91.
316. Watanabe, M., et al., *Adenovirus Biology, Recombinant Adenovirus, and Adenovirus Usage in Gene Therapy.* Viruses, 2021. **13**(12).
317. Greber, U.F. and J.W. Flatt, *Adenovirus Entry: From Infection to Immunity.* Annu Rev Virol, 2019. **6**(1): p. 177-197.
318. Maran, A. and M.B. Mathews, *Characterization of the double-stranded RNA implicated in the inhibition of protein synthesis in cells infected with a mutant adenovirus defective for VA RNA.* Virology, 1988. **164**(1): p. 106-13.
319. Lei, M., Y. Liu, and C.E. Samuel, *Adenovirus VAI RNA antagonizes the RNA-editing activity of the ADAR adenosine deaminase.* Virology, 1998. **245**(2): p. 188-96.
320. Atkin, S.J., B.E. Griffin, and S.M. Dilworth, *Polyoma virus and simian virus 40 as cancer models: history and perspectives.* Semin Cancer Biol, 2009. **19**(4): p. 211-7.
321. Kumar, M. and G.G. Carmichael, *Nuclear antisense RNA induces extensive adenosine modifications and nuclear retention of target transcripts.* Proc Natl Acad Sci U S A, 1997. **94**(8): p. 3542-7.

322. Schoggins, J.W., *Interferon-Stimulated Genes: What Do They All Do?* Annu Rev Virol, 2019. **6**(1): p. 567-584.
323. Walter, M.R., *The Role of Structure in the Biology of Interferon Signaling.* Front Immunol, 2020. **11**: p. 606489.
324. Yu, C., et al., *Hepatitis B virus (HBV) codon adapts well to the gene expression profile of liver cancer: an evolutionary explanation for HBV's oncogenic role.* J Microbiol, 2022. **60**(11): p. 1106-1112.
325. Arella, D., M. Dilucca, and A. Giansanti, *Codon usage bias and environmental adaptation in microbial organisms.* Mol Genet Genomics, 2021. **296**(3): p. 751-762.
326. Yang, S., et al., *Synonymous Codon Pattern of Cowpea Mild Mottle Virus Sheds Light on Its Host Adaptation and Genome Evolution.* Pathogens, 2022. **11**(4).
327. Raftery, N. and N.J. Stevenson, *Advances in anti-viral immune defence: revealing the importance of the IFN JAK/STAT pathway.* Cell Mol Life Sci, 2017. **74**(14): p. 2525-2535.
328. Sharp, P.M., T.M. Tuohy, and K.R. Mosurski, *Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes.* Nucleic Acids Res, 1986. **14**(13): p. 5125-43.
329. Puigbo, P., I.G. Bravo, and S. Garcia-Vallve, *CAIcal: a combined set of tools to assess codon usage adaptation.* Biol Direct, 2008. **3**: p. 38.
330. dos Reis, M., R. Savva, and L. Wernisch, *Solving the riddle of codon usage preferences: a test for translational selection.* Nucleic Acids Res, 2004. **32**(17): p. 5036-44.
331. Ji, W., et al., *Cross-species transmission of the newly identified coronavirus 2019-nCoV.* J Med Virol, 2020. **92**(4): p. 433-440.
332. Yao, H., M. Chen, and Z. Tang, *Analysis of Synonymous Codon Usage Bias in Flaviviridae Virus.* Biomed Res Int, 2019. **2019**: p. 5857285.
333. Tao, J. and H. Yao, *Comprehensive analysis of the codon usage patterns of polyprotein of Zika virus.* Prog Biophys Mol Biol, 2020. **150**: p. 43-49.
334. Cheng, S., H. Wu, and Z. Chen, *Evolution of Transmissible Gastroenteritis Virus (TGEV): A Codon Usage Perspective.* Int J Mol Sci, 2020. **21**(21).
335. Brister, J.R., et al., *NCBI viral genomes resource.* Nucleic Acids Res, 2015. **43**(Database issue): p. D571-7.
336. Cock, P.J., et al., *Biopython: freely available Python tools for computational molecular biology and bioinformatics.* Bioinformatics, 2009. **25**(11): p. 1422-3.
337. Chan, P.P., et al., *tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes.* Nucleic Acids Res, 2021. **49**(16): p. 9077-9096.
338. McInnes, L., J. Healy, and J. Melville *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2018. arXiv:1802.03426.
339. Yang, S. and Z. Gui, *An introduction on the multivariate normal-ratio distribution.* arXiv preprint arXiv:2310.14306, 2023.
340. Karthikeyan, S., et al., *Wastewater sequencing reveals early cryptic SARS-CoV-2 variant transmission.* Nature, 2022. **609**(7925): p. 101-108.
341. Kia, P., et al., *Genomic characterization of SARS-CoV-2 from Uganda using MinION nanopore sequencing.* Sci Rep, 2023. **13**(1): p. 20507.
342. Barbe, L., et al., *SARS-CoV-2 Whole-Genome Sequencing Using Oxford Nanopore Technology for Variant Monitoring in Wastewaters.* Front Microbiol, 2022. **13**: p. 889811.

343. Lu, R., et al., *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding.* Lancet, 2020. **395**(10224): p. 565-574.
344. Hadfield, J., et al., *Nextstrain: real-time tracking of pathogen evolution.* Bioinformatics, 2018. **34**(23): p. 4121-4123.
345. Irving, A.T., et al., *Lessons from the host defences of bats, a unique viral reservoir.* Nature, 2021. **589**(7842): p. 363-370.
346. de Wit, E., et al., *SARS and MERS: recent insights into emerging coronaviruses.* Nat Rev Microbiol, 2016. **14**(8): p. 523-34.
347. Zell, R., et al., *Cocirculation of Swine H1N1 Influenza A Virus Lineages in Germany.* Viruses, 2020. **12**(7).
348. Starick, E., et al., *Reassorted pandemic (H1N1) 2009 influenza A virus discovered from pigs in Germany.* J Gen Virol, 2011. **92**(Pt 5): p. 1184-1188.
349. Wang, L.F. and B.T. Eaton, *Bats, civets and the emergence of SARS.* Curr Top Microbiol Immunol, 2007. **315**: p. 325-44.
350. Graham, R.L. and R.S. Baric, *Recombination, reservoirs, and the modular spike: mechanisms of coronavirus cross-species transmission.* J Virol, 2010. **84**(7): p. 3134-46.
351. Guan, Y., et al., *Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China.* Science, 2003. **302**(5643): p. 276-8.
352. Martinez, M.A., et al., *Synonymous Virus Genome Recoding as a Tool to Impact Viral Fitness.* Trends Microbiol, 2016. **24**(2): p. 134-147.
353. Battles, M.B. and J.S. McLellan, *Respiratory syncytial virus entry and how to block it.* Nat Rev Microbiol, 2019. **17**(4): p. 233-245.
354. Jackson, C.B., et al., *Mechanisms of SARS-CoV-2 entry into cells.* Nat Rev Mol Cell Biol, 2022. **23**(1): p. 3-20.
355. Minkoff, J.M. and B. tenOever, *Innate immune evasion strategies of SARS-CoV-2.* Nat Rev Microbiol, 2023. **21**(3): p. 178-194.
356. Chen, F. and J.R. Yang, *Distinct codon usage bias evolutionary patterns between weakly and strongly virulent respiratory viruses.* iScience, 2022. **25**(1): p. 103682.
357. Chen, F., et al., *Dissimilation of synonymous codon usage bias in virus-host coevolution due to translational selection.* Nat Ecol Evol, 2020. **4**(4): p. 589-600.
358. Hernandez-Alias, X., et al., *Translational adaptation of human viruses to the tissues they infect.* Cell Rep, 2021. **34**(11): p. 108872.
359. Gale, M., Jr., S.L. Tan, and M.G. Katze, *Translational control of viral gene expression in eukaryotes.* Microbiol Mol Biol Rev, 2000. **64**(2): p. 239-80.
360. Balvay, L., et al., *Translational control of retroviruses.* Nat Rev Microbiol, 2007. **5**(2): p. 128-40.
361. Pinto, R.M., et al., *Hepatitis A Virus Codon Usage: Implications for Translation Kinetics and Capsid Folding.* Cold Spring Harb Perspect Med, 2018. **8**(10).
362. Deb, B., A. Uddin, and S. Chakraborty, *Analysis of codon usage of Horseshoe Bat Hepatitis B virus and its host.* Virology, 2021. **561**: p. 69-79.
363. Hou, W., *Characterization of codon usage pattern in SARS-CoV-2.* Virol J, 2020. **17**(1): p. 138.
364. Gu, H., et al., *Multivariate analyses of codon usage of SARS-CoV-2 and other betacoronaviruses.* Virus Evol, 2020. **6**(1): p. veaa032.
365. Suomalainen, M. and U.F. Greber, *Virus Infection Variability by Single-Cell Profiling.* Viruses, 2021. **13**(8).

366. Smatti, M.K., et al., *Viruses and Autoimmunity: A Review on the Potential Interaction and Molecular Mechanisms.* Viruses, 2019. **11**(8).
367. Novoa, E.M., et al., *Elucidation of Codon Usage Signatures across the Domains of Life.* Mol Biol Evol, 2019. **36**(10): p. 2328-2339.
368. Hatcher, E.L., et al., *Virus Variation Resource - improved response to emergent viral outbreaks.* Nucleic Acids Res, 2017. **45**(D1): p. D482-D490.
369. Rambaut, A., et al., *A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology.* Nat Microbiol, 2020. **5**(11): p. 1403-1407.
370. Alexaki, A., et al., *Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design.* J Mol Biol, 2019. **431**(13): p. 2434-2441.
371. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python.* the Journal of machine Learning research, 2011. **12**: p. 2825-2830.
372. Lemaître, G., F. Nogueira, and C.K. Aridas, *Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning.* The Journal of Machine Learning Research, 2017. **18**(1): p. 559-563.
373. Akiba, T., et al. *Optuna: A next-generation hyperparameter optimization framework*. in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2019.
374. Wu, F., et al., *A new coronavirus associated with human respiratory disease in China.* Nature, 2020. **579**(7798): p. 265-269.
375. Xu, C., et al., *Conformational dynamics of SARS-CoV-2 trimeric spike glycoprotein in complex with receptor ACE2 revealed by cryo-EM.* Sci Adv, 2021. **7**(1).
376. Walls, A.C., et al., *Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein.* Cell, 2020. **181**(2): p. 281-292 e6.
377. Harrison, A.G., T. Lin, and P. Wang, *Mechanisms of SARS-CoV-2 Transmission and Pathogenesis.* Trends Immunol, 2020. **41**(12): p. 1100-1115.
378. He, Y., et al., *A survey on deep learning in DNA/RNA motif mining.* Brief Bioinform, 2021. **22**(4).
379. Collaborators, C.-E.M., *Estimating excess mortality due to the COVID-19 pandemic: a systematic analysis of COVID-19-related mortality, 2020-21.* Lancet, 2022. **399**(10334): p. 1513-1536.
380. Crook, H., et al., *Long covid-mechanisms, risk factors, and management.* BMJ, 2021. **374**: p. n1648.
381. Peng, R., et al., *Cell entry by SARS-CoV-2.* Trends Biochem Sci, 2021. **46**(10): p. 848-860.
382. Bestle, D., et al., *TMPRSS2 and furin are both essential for proteolytic activation of SARS-CoV-2 in human airway cells.* Life Sci Alliance, 2020. **3**(9).
383. Flati, T., et al., *HPC-REDItools: a novel HPC-aware tool for improved large scale RNA-editing analysis.* BMC Bioinformatics, 2020. **21**(Suppl 10): p. 353.
384. Lo Giudice, C., et al., *Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal.* Nat Protoc, 2020. **15**(3): p. 1098-1131.
385. Maveyraud, L. and L. Mourey, *Protein X-ray Crystallography and Drug Discovery.* Molecules, 2020. **25**(5).
386. Milne, J.L., et al., *Cryo-electron microscopy--a primer for the non-microscopist.* FEBS J, 2013. **280**(1): p. 28-45.
387. Herzik, M.A., Jr., *Cryo-electron microscopy reaches atomic resolution.* Nature, 2020. **587**(7832): p. 39-40.

388. Algar, W.R., et al., *FRET as a biomolecular research tool - understanding its potential while avoiding pitfalls.* Nat Methods, 2019. **16**(9): p. 815-829.
389. Fang, C., Y. Huang, and Y. Zhao, *Review of FRET biosensing and its application in biomolecular detection.* Am J Transl Res, 2023. **15**(2): p. 694-709.
390. Ni, Z., *Optimising Multi-Protein Interaction Screening through Machine Learning Algorithm Development*, in *Faculty of Science*. 2024, University of Technology Sydney: University of Technology Sydney.
391. Inc., A.B. *Peptide and Protein Molecular Weight Calculator*. 2024 2024-01-13; Available from: https://www.aatbio.com/tools/calculate-peptide-and-protein-molecular-weight-mw.
392. Shaw, G., et al., *Preferential transformation of human neuronal cells by human adenoviruses and the origin of HEK 293 cells.* FASEB J, 2002. **16**(8): p. 869-71.
393. ATCC. *293 [HEK-293] (ATCC® CRL-1573™)*. 2019; Available from: https://www.atcc.org/products/all/CRL-1573.aspx.
394. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nat Methods, 2022. **19**(6): p. 679-682.
395. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
396. Shaner, N.C., et al., *Improved monomeric red, orange and yellow fluorescent proteins derived from Discosoma sp. red fluorescent protein.* Nat Biotechnol, 2004. **22**(12): p. 1567-72.
397. McCullock, T.W., D.M. MacLean, and P.J. Kammermeier, *Comparing the performance of mScarlet-I, mRuby3, and mCherry as FRET acceptors for mNeonGreen.* PLoS One, 2020. **15**(2): p. e0219886.
398. Picardi, E., et al., *REDIportal: a comprehensive database of A-to-I RNA editing events in humans.* Nucleic Acids Res, 2017. **45**(D1): p. D750-D757.
399. Ogawa, J., et al., *The D614G mutation in the SARS-CoV2 Spike protein increases infectivity in an ACE2 receptor dependent manner.* bioRxiv, 2020.
400. Cecon, E., et al., *SARS-COV-2 spike binding to ACE2 in living cells monitored by TR-FRET.* Cell Chem Biol, 2022. **29**(1): p. 74-83 e4.
401. Cheng, F.J., et al., *Umbelliferone and eriodictyol suppress the cellular entry of SARS-CoV-2.* Cell Biosci, 2023. **13**(1): p. 118.
402. Chen, X., J.L. Zaro, and W.C. Shen, *Fusion protein linkers: property, design and functionality.* Adv Drug Deliv Rev, 2013. **65**(10): p. 1357-69.
403. Choi, S.R. and M. Lee, *Transformer Architecture and Attention Mechanisms in Genome Data Analysis: A Comprehensive Review.* Biology (Basel), 2023. **12**(7).
404. Ji, Y., et al., *DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome.* Bioinformatics, 2021. **37**(15): p. 2112-2120.
405. Baer, A. and K. Kehn-Hall, *Viral concentration determination through plaque assays: using traditional and novel overlay systems.* J Vis Exp, 2014(93): p. e52065.
406. Dominguez, J., M.M. Lorenzo, and R. Blasco, *Green fluorescent protein expressed by a recombinant vaccinia virus permits early detection of infected cells by flow cytometry.* J Immunol Methods, 1998. **220**(1-2): p. 115-21.
407. Lan, T., et al., *Generating mutants of monotone affinity towards stronger protein complexes through adversarial learning.* Nature Machine Intelligence, 2024.
408. Chavez, M., et al., *Stable expression of large transgenes via the knock-in of an integrase-deficient lentivirus.* Nat Biomed Eng, 2023. **7**(5): p. 661-671.