Dissertation

submitted to the
Combined Faculty of Natural Sciences and Mathematics
of the Ruperto Carola University Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

Presented by
Rafail Nikolaos Tasakis, BSc
born in Giannitsa, Greece

Oral examination: November 16th, 2021

Collateral genomic damage due to aberrant RNA editing activity in cancer

Referees

Prof. Dr. Frank Lyko Prof. Dr. Nina Papavasiliou

Acknowledgements

Throughout my PhD I am deeply grateful for all the people I met and the number of fruitful experiences I had, none of which I will ever forget. First and foremost, I wish to thank my PhD supervisor, **Prof. Dr. Nina Papavasiliou**, for trusting me with a wide range of groundbreaking projects that challenged me to become a better scientist, for the amazing opportunities she gave me (inside and outside of Germany), as well as for her amazing support in my professional development. My lab mentor (and the best conference co-traveler), **Ricca**, for his support, advice, the opportunities he gave me to participate in his cool projects and for all the answers to my (not so few) questions in and out of the lab. And of course, all my colleagues (and friends) in the lab for their help, feedback, advice and for simply making it a true pleasure to be in the lab every day: **Anastasia, Anna, Annette, Alex, Bea, Erec, Evi, Gianna, Taga, Joey, Konstantina, Laura, Lena, Monica, Paulo, Sandra, Salvo, Sonal, Tim, Xico**. Especially the master students, **Dimitra, George and Jasmin**, whom I had the pleasure of working with, and whose important contributions for sure made a difference to my PhD. I am moreover very thankful to my faculty supervisor, **Prof. Dr. Frank Lyko**, as well as all my TAC members and examiners, **Dr. Judith Zaugg, Prof. Dr. Michaela Frye**, **Dr. Apostolos Zaravinos,** for their constructive feedback which undoubtedly helped me fulfil my PhD projects and for their time reviewing my dissertation.

I am, furthermore, very grateful to all the collaborators I had the pleasure of working with. Especially, **Dr. Samir Parekh and Dr. Alessandro Laganà**, from the Icahn School of Medicine at Mount Sinai in New York, for introducing me to Multiple Myeloma research, their mentorship in computational analyses, their overall feedback and guidance throughout my PhD and for hosting me in their lab(s) in New York. **Prof. Dr. Marilyn Diaz and Prof. Dr. Laurent Verkoczy**, from the San Diego Biomedical Research Institute, for introducing me to the world of evolutionary biology and virology, which has been key in addressing my big question from an evolutionary perspective. Moreover, **Sandrine, Meiqi and Eirini; Anna, Gimi and Thorsten; Mitch** for the nice collaborations we had. Finally, thanks to the collaboration between **Dr. Naomi Kohen** from the Nightingale-Bamford School, New York and Prof. Papavasiliou's lab, I had the chance to mentor high school summer students in the lab, **Alexandra, Anna, Ellie, Michelle, Sophia**, whose young talent and commitment has been truly inspiring!

Naturally, this would not be a good acknowledgments page without thanking family and friends. I wish to thank my parents, **Stefanos and Georgia**, as well as my brothers, **Dimitris and Tasos**, for their unprecedented support ever since my early student years, for believing in me even when I don't, for being my number one fans and for being my inspiration to start all this. To my parents especially: thank you for teaching me to work harder and harder, to always be honest (even when it is not easy), to never give up, and to never forget where I started from. My nephews, **Stefanos Jr. and Stavros**, whom I am extremely proud of, and I hope they are proud of me too. But also, to my "Belgian family", **Gerda,**

Andreas, Charis, Greg, Guido and Veerle for the warmest welcomes, the new experiences and their amazing support. All my friends out of the lab; **Loukia, Mika, Sylvain, Angelos, Nancy, Vassilis, Christina, Anna Maria, Eleftheria, Nikoleta, Isidora, Alkisti, Anestis and… my therapist**. And last, but definitely not least, I wholeheartedly wish to thank my partner in life and husband, as of recently, **Leander**, for his unconditional support, understanding and patience. Leander, I owe you so many weekends without going to the cell culture (although by now you must have become very good at splitting cells yourself)!

Last, I wish to dedicate this thesis to my father, and my late grandfather, whom I dearly love and admire for a number of reasons. At the time of writing my dissertation, my father was diagnosed with the main topic of my PhD, cancer. But he's gonna beat it, just like my brother did. It appears to be relatively easy to draw correlations and survival rates for cancer patients in a publication or a thesis, but I consistently fail every time to draw any kind of "correlations" when it's happening around me. I can only hope that my work, in research or elsewhere, today and from now on will benefit people and will not let them down. *Dad, thank you for showing me that quitting is never an option.* On the other hand, my grandfather lived a life full of amazing experiences and achievements; and he did that first for others and then for himself. Although he was discriminated for his identity and beliefs, he was a fighter; no matter how "wrong" he was. He actually fought against a system of oppression, while risking his life a number of times. But he made it. He was seeing opportunities and never obstacles. But what I admire most of him is not his honors, but his determination to make a difference for 'his people' every single day of his life at any cost or place; at the construction site, at the vineyard, at home, in the labor union, in politics, in Greece, in France, in Germany. Both of them have been *my* motivation to never quit, to not be afraid to shoot for the impossible, to dream big and work hard, but also to accept myself no matter how "wrong" I am; for myself and 'my people'. *Thank you!*

Table of contents

Abbreviations

Ab – **A**nti**b**ody

ACE2 – Angiotensin**-C**onverting **E**nzyme **2**

ADAR – **A**denosine **D**eaminase **A**cting on **R**NA

AEI – Alu **E**diting **I**ndex

AID – **A**denosine Cytidine **I**nduced **D**eaminase

APC – **A**llo**p**hy**c**ocyanin

APOBEC – **Apo**lipoprotein **B** mRNA **e**diting enzyme, **c**atalytic polypeptide-like

arRNA – ADAR**-r**ecruiting **R**NA

ASOs – **a**nti**s**ense **o**ligonucleotide**s**

BER – **B**ase **E**xcision **R**epair

cDNA – complementary **DNA**

CDS – Co**d**ing **S**equence

COVID-19 – Corona**vi**rus **D**isease 20**19**

CSR – Class **S**witch **R**ecombination

DD – Deaminase **D**omain

DSB – **D**ouble-**s**trand DNA **B**reak

FACS – **F**luorescence-**A**ctivated **C**ell **S**orting

gDNA – genomic **DNA**

GFP – **G**reen **F**luorescent **P**rotein

gRNA – guide **RNA**

IFN – Inter**f**ero**n**

IVT – In **V**itro **T**ranscription

LEAPER – Leveraging **E**ndogenous **A**DAR for **P**rogrammable **E**diting of **R**NA

LFSM – Low **F**requency **S**pike **M**utations

MM – **M**ultiple **M**yeloma

NES – **N**uclear **E**xport **S**ignal

NGS – **N**ext **G**eneration **S**equencing

NLS – **N**uclear **L**ocalization **S**ignal

Nsp – **N**on-**s**tructural **p**rotein

ORF – **O**pen **R**eading **F**rame

PCR – Polymerase **C**hain **R**eaction

PI – Propidium **I**odine

PKR – **P**rotein **K**inase **R**

RBM – **R**NA **B**inding **M**otif

RBP – **R**NA **B**inding **P**rotein

RdRp – **R**NA**-d**ependent **R**NA **p**olymerase

RESTORE - **R**ecruiting **e**ndogenous ADAR to **s**pecific **t**ranscripts for **o**ligonucleotide-mediated **R**NA **e**diting

RLP – **R**IG-I-**L**ike Receptor **p**athway

RTC – **R**everse **T**ranscription **C**omplex

SARS-CoV-2 – **S**evere **A**cute **R**espiratory **S**yndrome **Co**rona**v**irus **2**

SHM – **S**omatic **H**yper**m**utation

SNP – **S**ingle **N**ucleotide **P**olymorphism

SNV – **S**ingle **N**ucleotide **V**ariation

ssRNA/ssDNA – **s**ingle-**s**tranded **RNA/DNA**

TP – **T**ime**p**oint

UTR – **U**n**t**ranslated **R**egion

VOC – **V**ariant **o**f **C**oncern

WES – **W**hole **E**xome **S**equencing

WT – **W**ild**t**ype

Abstract

RNA editing is an epitranscriptomic modification of rising prominence in health and disease. It is catalyzed by enzymes from the families of 'Adenosine Deaminases Acting on RNA’ (ADAR) or ‘Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like’ (APOBEC). Multiple RNA editing deaminases, however, not only can they edit RNA, but also mutate DNA. ADARs particularly, are naturally capable of editing dsRNA co-transcriptionally, as well as mutating DNA in DNA/RNA hybrids. Although, the mutagenic role of ADARs is well-studied *in vitro*, its relevance with *in vivo* models has yet to be explored. DNA/RNA hybrids (or R-loops) form co-transcriptionally in the human genome between the nascent RNA and the template DNA strand, and I hypothesized that ADARs can access them to mutate the DNA strand in the hybrid, after losing touch with the nascent RNA-target. Here, I focus on ADAR1, which is overexpressed in Multiple Myeloma (MM) leading to aberrant editing activity and poor disease outcomes. RNA-seq and Whole-Exome Sequencing (WES) matched datasets from 23 MM patients pre- and post-relapse revealed acquisition of unique mutations post-relapse, enriched in the vicinity of RNA editing events pre-relapse. For proof-of-concept experiments in cell lines, I employed site-directed mRNA editing tools to target ADARs to specific transcripts, and evaluated whether ADAR-mediated DNA mutation was generated in their cognate genes. I found that ADARs *may* mutate genomic DNA in a rate of 1 in 25 000. Last, I explored the evolutionary impact of mutagenesis mediated by RNA editing enzymes (ADARs and APOBECs) in single-stranded RNA viral genomes from SARS-CoV-2 and showed that RNA editing enzymes may drive genome evolution by gradually accumulating co-occurring mutations, which similarly in cancer biology would translate to clonal expansion for tumor adaptation. Overall, my findings, suggest that DNA mutations may arise as collateral genomic damage by RNA editing deaminases, the initial job of which was to edit the cognate transcript *in situ*.

Zusammenfassung

Die RNA-Edierung ist eine epitranskriptomische Modifikation von aufkommender Bedeutung für Gesundheit und Krankheit. Es wird durch Enzyme aus den Familien der "Doppelsträngige RNA-spezifische Adenosin-Desaminase" (DRADA oder ADAR) oder "apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like" (APOBEC) katalysiert. Multiple RNA-Editing-Deaminasen können jedoch nicht nur RNA editieren, sondern auch DNA mutieren. Insbesondere ADARs sind von Natur aus in der Lage, dsRNA kotranskriptionell zu edieren sowie DNA in DNA/RNA-Hybriden zu mutieren. Obwohl die mutagene Rolle von ADARs *in vitro* schon gut untersucht ist, ist ihre Relevanz für *in vivo* Modelle größtenteils noch zu erforschen. DNA/RNA-Hybriden (oder R-Loops) bilden sich im menschlichen Genom zwischen dem entstehenden RNA und der Matrize, und ich stellte die Hypothese auf, dass ADARs auf sie zugreifen können, um den DNA-Strang im Hybrid zu mutieren, nachdem sie den Kontakt mit dem entstehenden RNA-Ziel verloren haben. Hier konzentriere ich mich auf ADAR1, das beim Multiplen Myelom (MM) überexprimiert wird, was zu einer aberranten Edieraktivität und schlechten Krankheitsverläufe führt. RNA-seq und Sequenzierung des gesamten Exoms (WES) übereinstimmende Datensätze von 23 MM-Patienten vor und nach dem Rückfall zeigten den Erwerb einzigartiger Mutationen nach dem Rückfall, angereichert in der Nähe von RNA-Ediereereignissen vor dem Rückfall. Für Proof-of-Concept-Experimente in Zelllinien habe ich ortsgerichtete mRNA-Edierungs-Tools verwendet, um ADARs auf bestimmte Transkriptionen abzuzielen, und ich habe bewertet, ob ADAR-vermittelte DNA-Mutation in den verwandten Genen erzeugt wurde. Ich fand heraus, dass ADARs genomische DNA mit einer Rate von 1 zu 25 000 mutieren könnten. Zuletzt habe ich den evolutionären Einfluss der Mutagenese vermittelt durch RNA-edierende Enzyme (ADARs und APOBECs) in einzelsträngigen RNA-Virusgenomen aus SARS-CoV-2 untersucht, und ich zeigte, dass RNA-edierende Enzyme die Genomentwicklung vorantreiben könnten, indem sie schrittweise gleichzeitig auftretende Mutationen akkumulieren, was in der Krebsbiologie analog wäre zu einer klonalen Expansion zur Tumoranpassung. Insgesamt deuten meine Ergebnisse darauf hin, dass DNA-Mutationen als kollaterale genomische Schäden durch RNA-edierende Desaminasen auftreten könnten, deren ursprüngliche Aufgabe darin bestand, das verwandte Transkript *in situ* zu edieren.

List of figures

1. Introduction

1.1 The significance of RNA modifications

RNA modifications play a prominent role in cellular homeostasis, in tissue development, but also in health and disease, by regulating gene expression through type- and site-specific modifications on the transcript level (Frye et al., 2018; Livneh et al., 2020). Recent advances in epitranscriptomics, the field that studies RNA modifications, have unveiled a great deal of diversity in 1) the types of RNA modifications, 2) the types of modified RNAs and 3) the topology within the modified transcript, which subsequently impact translation (Hoernes and Erlacher, 2017). Up to date, there are over 170 different kinds of RNA modifications described in the mammalian epitranscriptome, most of which are found in the tRNA and rRNA, and a few have also been found in the mRNA (Delaunay and Frye, 2019).

tRNA modifications are amongst the very well-studied and are crucial for the correct tRNA structure and function, including the proper anticodon loop formation, aminoacylation and tRNA metabolism (El Yacoubi et al., 2012). rRNA modifications generally contribute to structural ribosomal stability, while they cluster at functionally crucial ribosomal sites, such as the peptidyltransferase center and the decoding site, guaranteeing the efficacy and accuracy of translation (Sloan et al., 2017). For mRNA modifications, on the other hand, there is not an utterly clear rule for their effect overall on the mRNA; there are examples of RNA modifications involved in mRNA decay, transcript stability or translation (Hoernes and Erlacher, 2017; Frye et al., 2018). Modifications most frequently found on the mRNA are summarized in Figure 1.1. In brief, among the modifications found on the mRNA, grouped by the originally modified base are:

- **Methylation of Adenosines**: N^6-methyladenosine (m^{6}A) appears to be one of the most abundant modifications, written cooperatively by a complex of 7 proteins (purple-highlighted in Figure 1.1), but methylation is primarily catalyzed by methyltransferase-like 3 (METTL3) upon substrate recognition of METTL14 (Roundtree et al., 2017). Furthermore, N^6,2′-O-dimethyladenosine (m^{6}Am) is catalyzed by the Phosphorylated CTD Interacting Factor 1 (PCIF1), later termed as Cap Adenosine N6-Methyltransferase (CAPAM), as it methylates the first Adenosine transcribed to create the mRNA cap structure (Cowling, 2019; Sendinc et al., 2019). N^1-methyladenosine (m^{1}A), is one of the most rare modifications found on the mRNA (Schwartz, 2018) and it is written by a complex of tRNA (adenine-N(1)-)-methyltransferases TRMT6 (substrate recognition) and TRMT61A (catalytic subunit) (Safra et al., 2017).
- **Methylation of Cytosines**: 5-methylcytosine (m^{5}C) is written on the mRNA by the RNA methyltransferase NSUN2 in various sites throughout the transcript (Bohnsack et al., 2019). Hydroxylation of m^{5}C, catalyzed by the Tet methylcytosine dioxygenases (TETs) activity, leads to the formation of 5-hydroxymethylcytosine (hm^{5}C), which is thought to play an important role in differentiation, as it has been found in mouse embryonic stem cells (ESCs) in abundance in

transcripts of key pluripotency-related factors and its levels decrease during differentiation (Lan et al., 2020).

- **Pseudouridylation of Uridine**: isomerization of uridine to pseudouridine (ψ) on the mRNA in human cells appears to be generally accumulated in response to environmental stress (Carlile et al., 2014). ψ is catalyzed by pseudouridine synthetases (PUSs) or a small ribonucleoprotein complex with DKC1 (Dyskerin Pseudouridine Synthase 1) as its catalytic subunit (Balogh et al., 2020).
- **Deamination of Adenosines and Cytosines**: this modification is generally known as RNA editing and it results in base change: Adenosines are deaminated to Inosines, which are recognized as Guanosines, (hereafter A-to-I(G) editing) by enzymes of the ADAR family (Adenosine Deaminases Acting on RNA), while Cytosines are deaminated to Uracils (hereafter C-to-U editing) by enzymes of the APOBEC family (apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like) (Keegan et al., 2001). RNA editing is the main focus of the present dissertation and their physiological roles, as well as the shift of their function in cancer are reviewed in detail in the subchapter 1.2.

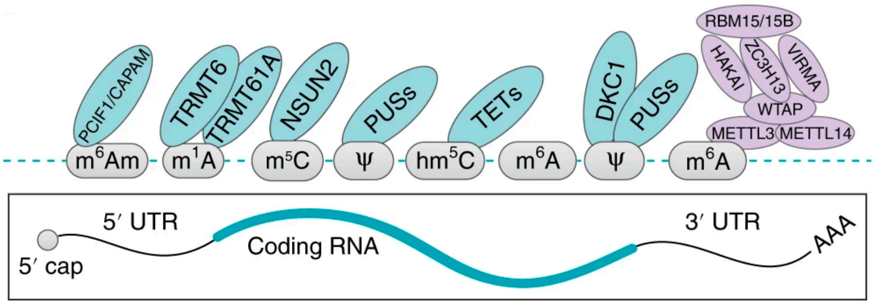


Figure 1.1 Scheme representing the mRNA modifications and their respective transcript topology. Namely the modifications shown are: N6-methyladenosine (m6A), N6,2′-O-dimethyladenosine (m6Am), N1-methyladenosine (m1A), 5-methylcytosine (m5C), 5-hydroxymethylcytosine (hm5C), pseudouridine (ψ). The writer(s) of each modification are shown above the modifications in light-blue and purple shapes. For further information, see main text above. Adapted figure from Delaunay and Frye, 2019. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Nature Cell Biology*, "RNA modifications regulating cell fate in cancer", Delaunay and Frye, Copyright: Springer Nature Limited (2019).

Overall, RNA modifications are widespread and diverse. They comprise an extensive repertoire of chemical modifications, and they occur on most of the types of RNA found in a cell, often with specific transcript topology. Though the functional consequence of most remains unclear, emerging evidence suggests that in aggregate they provide an additional layer of regulation on the transcript level. In this dissertation, I will focus on A-to-I and C-to-U (on- and off-target) RNA editing and in particular how they impact cancer development and progression.

1.2 The polynucleotide deaminases and their known functions

1.2.1 Adenosine Deaminases Acting on RNA (ADARs)

The most abundant type of RNA editing by deamination is the Adenosine-to-Inosine (A-to-I) conversion (Figure 1.2a) catalyzed by enzymes of the family of Adenosine Deaminases Acting on RNA (ADAR) (Zinshteyn and Nishikura, 2009). There are three ADARs in humans, only two of which (ADAR1 and ADAR2) demonstrate A-to-I deaminase activity (Slotkin and Nishikura, 2013). As summarized in Figure 1.3b, all ADARs have a deaminase domain and at least two double-stranded RNA (dsRNA) binding domains (Slotkin and Nishikura, 2013). ADAR1 is expressed ubiquitously in two isoforms, ADAR1-p110 (~110 kDa) and ADAR1-p150 (~150 kDa), which are being generated through alternative promoters (George and Samuel, 1999). ADAR1-p110 is constitutively expressed, while ADAR1-p150 is expressed through an interferon inducible promoter upstream the one responsible for the ADAR1-p110 expression (George et al., 2005). Through this alternative promoter usage, the protein sequence of ADAR1-p150 is longer than the one of ADAR1-p110 by an additional exon at the N-terminus, encoding for an extra Z-DNA binding domain. This additional Z-DNA binding domain provides ADAR1-p150 with a nuclear export signal (NES), allowing its presence into the cytoplasm where it is mostly found, while ADAR1-p110 resides in the nucleus through nuclear localization signal (NLS) present in the dsRNA binding domain closer to the C-terminus, also present in ADAR1-p150 (Poulsen et al., 2001).

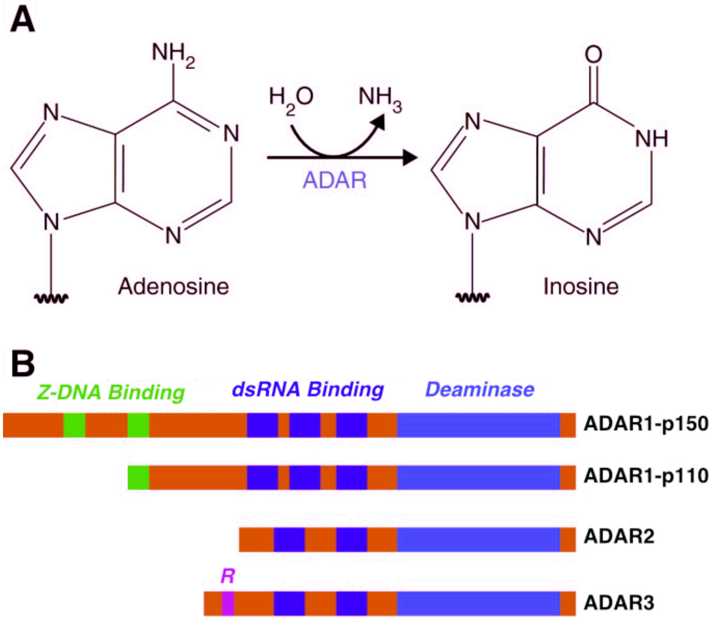


Figure 1.2 Adenosine-to-Inosine deamination and the enzymes of the ADAR family. (**A**) Chemical reaction of A-to-I deamination, in which an Adenosine loses an amino-group and is converted to inosine. (**B**) The ADAR family consists of three main enzymes: ADAR1, which has two isoforms (ADAR1-p150 and ADAR1-p110), ADAR2 and ADAR3. All ADARs contain deaminase and dsRNA binding domains, while Z-DNA biding domains are present only in the isoforms of ADAR1, and an arginine-rich (R) domain in ADAR3. Adapted figure from Slotkin and Nishikura, 2013. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Genome Medicine*, “Adenosine-to-inosine RNA editing and human disease”, Slotkin and Nishikura, Copyright: BioMed Central Ltd (2013).

As mentioned above, A-to-I RNA editing is widespread in the human transcriptome; it is mainly driven by ADAR1 and it is particularly prevalent in *Alu* element sequences of transcripts, majorly present in introns and untranslated regions (UTRs) (Athanasiadis et al., 2004; Kim et al., 2004; Levanon et al., 2004). Although mRNAs can be post-transcriptionally edited by the mainly cytoplasmic ADAR1-p150 isoform, its abundance in intronic regions of pre-mRNAs highlighted that this modification can also occur co-transcriptionally and it is, in fact, tightly coordinated with pre-mRNA splicing (Laurencikiene et al., 2006; Licht et al., 2019). A number of transcript-targets have been identified for ADAR1, highlighting its prominent role in targeting components of a number of molecular pathways and cellular functions (Lamers et al., 2019): its primary function is to edit endogenous dsRNAs, in order to inhibit response to them as non-self dsRNAs, which would normally activate the RIG-I-Like Receptor pathway (RLP) and stimulate interferon (IFN) type-I response (Liddicoat et al., 2015). Although the last highlights the importance of ADAR1 in immunity, allowing the host to trigger antiviral response, pro-viral roles have been suggested as well for this enzyme, as it can block RLP or PKR (Protein Kinase R), which also recognizes dsRNA (George et al., 2009). Despite its role in immunity, A-to-I editing also plays a prominent role in cancer primarily attributed to ADAR1 activity (discussed in subchapter 1.2.3), which is overexpressed in several cancer types (Han et al., 2015) and, interestingly, it is moreover involved in diversifying sequence of RNA viral genomes, such as in HIV-1 (Doria et al., 2009) or the novel coronavirus SARS-CoV-2 (Giorgio et al., 2020). The last, will be discussed in subchapter 1.2.4.

The other catalytically active deaminase of the family, ADAR2, is also nuclear and it is suggested to be responsible for the few editing events described in the coding sequence (CDS) of the transcripts (Nishikura, 2010). For instance, thanks to ADAR2 editing targeting the CDS of the transcript of *GRIA2* (encoding for the glutamate receptor subunit B; GluR-B in mice), a glutamine-to-arginine (Q-to-R) amino acid change makes the AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazole propionate) receptor impermeable to calcium, which is crucial for the brain function (Brusa et al., 1995; Higuchi et al., 2000). *In vitro* experiments have demonstrated that ADAR1 and ADAR2 homodimerize for editing dsRNA, while ADAR3 remains a monomer (Cho et al., 2003). ADAR3 is not a catalytically active deaminase; it is expressed in brain and contains an arginine-rich (R) domain, which allows the enzyme to bind to single-stranded RNA (ssRNA), as well as to dsRNA through the respective domains (Chen et al., 2000). Recent findings demonstrate that ADAR3 may be an inhibitor for ADAR2 to edit RNA, as for example it binds to the *GRIA2* transcript in human glioblastoma tumors and hinders the crucial Q-to-R editing by ADAR2 (Oakes et al., 2017).

Overall, ADAR-mediated RNA editing not only is it crucial for a number of functions as discussed thus far, but also it is essential for life. ADAR1-deficient mice die during embryogenesis for a number of reasons, including high IFN type I levels, liver failure, dysfunctional hematopoiesis and extended apoptosis (Hartner et al., 2004; Wang et al., 2004; Hartner et al., 2009; Mannion et al., 2014). In human, mutations in the *ADAR1* gene and in particular in key domains (i.e. deaminase, Z-DNA binding, dsRNA binding domains) are problematic and causal to diseases. For instance, 9 mutations (7 of which in the deaminase domain) are causal to Aicardi-Goutières syndrome, a rare neurological disorder with high inflammation levels of the brain and skin (Rice et al., 2012). About 130 mutations throughout the *ADAR1* CDS are further associated with another rare genetic disorder, dyschromatosis symmetrica hereditaria, which is characterized by hyper- and hypopigmented macules on the extremities (Kono et al., 2014). Furthermore, downregulation of ADAR2, which leads to minimal editing levels of *GRIA2* at the Q-to-R site is associated with deterioration of motor neurons in amyotrophic lateral sclerosis, a fatal neurodegenerative disease, but also with glioblastoma, an advanced-stage brain malignancy of astrocytes (Maas et al., 2001; Kawahara et al., 2003). Conclusively, ADAR-mediated RNA editing is required for systemic homeostasis and its deregulation my lead to severe cases in human disease.

1.2.2 The AID/APOBEC family

The AID/APOBEC (**A**ctivated **I**nduced Cytidine **D**eaminase / **Apo**lipoprotein-**B**-mRNA **E**diting enzyme **C**atalytic polypeptide-1) is a family of cytidine deaminases, which according to Salter et al., 2016, they are "united by structure and divergent in function". In human, the AID/APOBEC family consists of 11 enzymes, APOBEC1, APOBEC2, APOBEC3A-D, APOBEC3F-H (collectively referred to as APOBEC3s), APOBEC4 and AID (Conticello et al., 2007). Not all the members of the AID/APOBEC family are deaminases; Cytidine deamination leads to Uridine, generally known as C-to-U editing and it is catalyzed by APOBEC1, some APOBEC3s and AID (Conticello et al., 2005). The first cytidine deaminase to be discovered was APOBEC1, which deaminates the Cytidine in a CAA codon (amino acid Q2180) of the pre-mRNA of apolipoprotein B (ApoB), leading to a stop codon (UAA) and therefore a truncated protein, known as ApoB-48 (Teng et al., 1993). ApoB-48 is expressed primarily in the small intestine, while ApoB-100, deriving from the unedited transcript of ApoB in liver (Blanc and Davidson, 2003). The two proteins present distinct functions relevant to their tissue-specific expression; ApoB-100 is an LDL-R (Low Density Lipoprotein Receptor) ligand, but ApoB-48 misses the ligand domain and it is instead involved in chylomicrons metabolism (Zheng et al., 2006).

The aforementioned recoding effect of RNA editing in ApoB expression is an example of the striking biological impact RNA editing may have. However, RNA editing events in the coding regions are rather rare; comparative transcriptome-wide screens between *Apobec1*$^{-/-}$ and wild-type (WT) mice from jejunal epithelial cells of their small intestine revealed that APOBEC1 edits Cytidines within AU-rich regions in transcript 3'UTRs, a finding that indicates that RNA editing may also be involved in transcript regulation or processing (Rosenberg et al., 2011). APOBEC1 edits single-stranded RNA (ssRNA) normally with the help of other proteins, known as cofactors, which recognize ribonucleotide motifs (mooring sequences) and "tether" the editing enzyme to the on-target Cytidine (Keegan et al., 2001). The most well-known cofactors of APOBEC1 are RBM47 (RNA Binding Motif 47) and A1CF (APOBEC1 Complementation Factor), which each recruits APOBEC1 to certain transcripts (Blanc et al., 2019). Indeed, Rbm47-deficient mice fail to edit specific transcripts, different from those that A1CFdeficient mice fail to edit, while mice with double deficiency (of both RBM47 and A1CF) present a global decrease of RNA editing but not utter loss (Fossat et al., 2014; Snyder et al., 2017). This suggests that APOBEC1 may have occasionally additional cofactors than the aforementioned ones, considering the fact that RBPs (RNA Binding Proteins) present similarities in sequence, the encoded protein domains, and their preferential binding motifs (Dominguez et al., 2018). It is not unlikely, that the cofactors of APOBEC1 may also function as "molecular switches", determining the function of the deaminase, not only by enhancing target-transcript specificity and versatility, but also by allowing APOBEC1 to mutate DNA instead of RNA. And this is because, it has been shown in *E. coli* that when APOBEC1 is ectopically expressed is capable of mutating DNA without its cofactors (Harris et al., 2002), while mutational signatures compatible with APOBEC1 activity where also detected in advanced esophageal adenocarcinoma cells (Saraconi et al., 2014). Therefore, APOBEC1 residing in the nucleus to edit pre-mRNA with the help of cofactors (Chester et al., 2003), may also mutate DNA without them; a likely scenario is that this dual role of APOBEC1 may be coordinated co-transcriptionally (as summarized in Figure 1.3), in which APOBEC1 may lose touch with its cofactor and the "on-hold" nascent RNA and access the ssDNA coding strand of the transcribed locus (Tasakis et al., 2019).

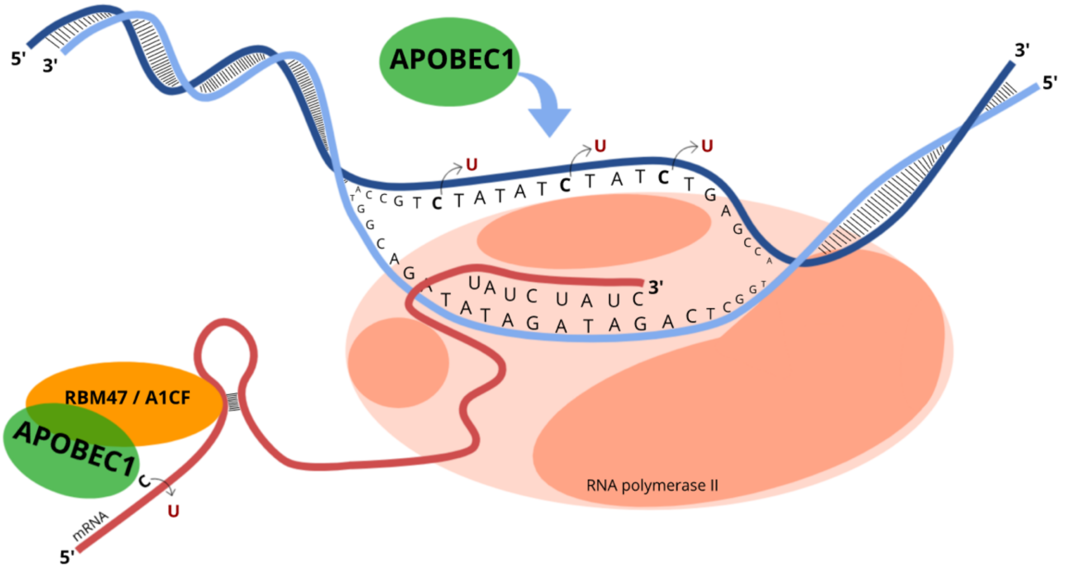


Figure 1.3 RNA editing and DNA mutation by APOBEC1 may be co-transcriptionally coordinated. RNA editing enzymes, known to edit RNA co-transcriptionally, here showing APOBEC1 and its two predominant cofactors (RBM47 and A1CF). It is, however, also known that APOBEC1 can gain access and mutate DNA without a co-factor (Harris et al., 2002). It is possible that APOBEC1 may lose its touch with the nascent RNA and mutate DNA *in situ*. A DNA molecule in close proximity to the nascent RNA will by necessity be the cognate gene of the transcribed locus. Therefore, RNA editing and DNA mutation by the same enzyme (here APOBEC1, but not limited to it) may be temporally linked during transcription. Figure from Tasakis et al., 2019. It is reused under the Creative Commons License 4.0. This illustration was created by myself.

The second deaminase from the AID/APOBECs to be discovered was AID (Activated Induced Cytidine Deaminase, encoded by the *Aicda* locus), which is expressed in germinal center B cells and plays a prominent role in adaptive immunity, as it is involved in antibody diversification through Class Switch Recombination (CSR) and Somatic Hypermutation (SHM) of the Immunoglobulin (Ig) locus (Muramatsu et al., 1999, 2000). AID was originally thought to be an RNA editing enzyme due to similarities with APOBEC1 in structure and deamination activity (Muramatsu et al., 1999), and also because their genetic loci were generated in the mammalian genome through gene duplication (Muto et al., 2000; Conticello et al., 2007). It was, however, demonstrated *in vitro* that AID deaminates ssDNA (Dickerson et al., 2003), which at Ig loci is accessible by AID during transcription thanks to R-loop formation between the nascent RNA and the template strand (Ramiro et al., 2003; Sohail et al., 2003). Therefore, there is a strong strand bias for C-to-U AID-mediated editing at Ig loci, which may be incorporated in the DNA sequence as an A:T base pair through DNA replication, or lead to DNA break because of Base Excision Repair (BER) mechanisms resulting in CSR or translocations (Longerich et al., 2006).

Other enzymes that are catalytically active deaminases from the AID/APOBEC family are the subfamily of APOBEC3s; there are seven different APOBEC3s (3A-D, 3F-H) expressed in human, while only one in mouse (Conticello et al., 2005). APOBEC3s are predominantly expressed in immune system cells and have key functions in anti-retroviral immune response, but also in regulating the innate retrotransposon activity (Chiu and Greene, 2008; Stavrou and Ross, 2015). Most of APOBEC3s are primarily cytoplasmic (besides APOBEC3B being mostly nuclear and APOBEC3A and -3C being both nuclear and cytoplasmic), where they are catalytically active against a number of RNA viruses, targeting their cDNA intermediates (Salter et al., 2016). One of the most well-studied members of this subfamily is APOBEC3G, which is an antiviral against HIV-1 targeting its ssDNA reverse-transcribed intermediates (Sheehy et al., 2002). Interestingly, APOBEC3G is “hijacked” into the HIV virions, in which it actually mutates the cDNA of HIV-1, but because of the Vif (Viral infectivity factor) of HIV enhancing the proteasome-mediated proteolysis of APOBEC3G, enhancing the viral infectivity (Lecossier et al., 2003; Mangeat et al., 2003; Marin et al., 2003). APOBEC3s are potent ssDNA mutators and APOBEC3A and -3B, as they can access and reside in the nucleus, are involved in aberrant hypermutation of cancer genomes, a phenomenon termed “kataegis”, as first discovered in breast cancers (Nik-Zainal et al., 2012). It was later shown that APOBEC3A can also edit RNA in monocytes and macrophages (Sharma et al., 2015) and it has been recently demonstrated that, RNA editing and DNA mutation by APOBEC3A are in fact co-dependent, because mutations in the DNA can be monitored through “hotspots” of RNA editing activity (Jalili et al., 2020). Given the fact that APOBEC3s can bind to ssRNA as well (Salter et al., 2016), it is not impossible that more APOBEC3s can present this versatile role of RNA/DNA targeting, perhaps with the right cofactor, which has yet to be discovered (Tasakis et al., 2019).

1.2.3 Implications of ADARs and AID/APOBECs in cancer mutagenesis

Cancer is a disease characterized by unprecedented heterogeneity in genetics, tissue pathology, leading to a diverse clinical presentation and progression (Fisher et al., 2013; Janku, 2014). The first attempt to interpret cancer development through mutagenesis was the "two-hit" hypothesis as proposed in 1971 by Alfred Knudson. According to this hypothesis, there are at least two mutations required for cancer development: one in a proto-oncogene to turn to an oncogene, to enhance cell proliferation, and one in a tumor-suppressor gene, so as forfeit the cellular capabilities control cell division so as to escape death (Knudson, 1971). Ever since that model was proposed, has become indisputably evident that cancer mutagenesis is well beyond two mutations; in most cases the mutational load is high to allow tumor evolution and adaptation (Martincorena and Campbell, 2015; Chalmers et al., 2017). Tumor development follows a clonal evolution model, according to which they originate from progenitor cells (also known as cancer stem cells), progressively anchoring mutations in the genome through selection and expansion (Beck and Blanpain, 2013). The set of mutations arising during tumorigenesis, is considered to be a load of errors that occur primarily during DNA replication, which repair mechanisms fail to correct, but it becomes more and more evident that deaminases, such as APOBECs, are actively involved in cancer mutagenesis by introducing non-canonical bases in the genome, which may be corrected by mismatch repair mechanisms (Chen et al., 2014). Such a mechanism, is for instance in case of C-to-U DNA editing, the recognition of Uracil (U) in the DNA by UDG (Uracil-DNA Glycosylase), which removes Us from the DNA and activates the BER pathway (Petljak and Maciejowski, 2020).

APOBECs often drive mutagenesis in a number of cancers throughout tumorigenesis with specific mutational signatures associated with their activity (Petljak et al., 2019). As also mentioned in subchapter 1.2.2, one of the first examples of APOBEC-mediated mutagenesis in cancer was the discovery that APOBEC3A and APOBEC3B correlate with clusters of C-to-T mutations, termed as "kataegis" mutations, originally inferred from mutation data from breast cancer genomes (Nik-Zainal et al., 2012; Starrett et al., 2016), which shaped a predictive signature with prognostic value (D'Antonio et al., 2016). APOBEC-mediated mutagenesis, however, is a pan-cancer phenomenon and has been detected in a number of cancer cell lines (Jarvis et al., 2018; Maura et al., 2018; Petljak et al., 2019). APOBEC1-mediated mutational signatures have also been detected in cancer cell line genomes, as introduced in subchapter 1.2.2 (Saraconi et al., 2014), and previous *in vivo* data demonstrated that APOBEC1-deficient mice presented reduced tumor burden indicating a direct link between the deaminase activity and cancer progression (Blanc et al., 2007). However, it should be noted that as discussed in subchapter 1.2.2, APOBEC1 is both an RNA editor and a DNA mutator in the context of (Saraconi et al., 2014), which appears to be the case also for APOBEC3A, which was originally thought to be a DNA mutator, but also proven to be an RNA editor (Jalili et al., 2020). Therefore, polynucleotide deaminases, present versatility in the target substrates and their coordination editing RNA and mutating DNA co-transcriptionally (Figure 1.3), may be key to tumorigenesis.

ADAR-mediated RNA editing is also shown to be key to tumorigenesis, with ADARs being overexpressed in virtually all tumors compared to their normal tissues, leading to elevated A-to-I editing activity, except for kidney chromophobe and renal papillary tumors (Han et al., 2015). Editing tumor load is indeed exceptionally high, found also more abundant in transcript CDSs, leading to transcriptomic diversity during tumorigenesis (Paz-Yaacov et al., 2015). Certain editing events in particular transcripts have been, in fact, described as driver events; for example, the transcript of *AZIN1* hepatocellular carcinoma (Chen et al., 2013) and the transcript of *GLI1* in Multiple Myeloma (Lazzari et al., 2017). Although it is evident that there is a certain preponderance in particular transcripts at certain cancer types, the transcriptome-wide RNA editing activity (which is aberrant) should be overall considered in the context of oncogenesis, as shown to be crucial for instance in MM prognosis (Teoh et al., 2018). Furthermore, ADARs may also present a dual role of editing RNA and mutating DNA, *in vitro* data demonstrate that ADARs deaminate DNA when that is in DNA/RNA hybrids (Zheng et al., 2017), and this appears to be crucial for resolution of telomeric R-loops and genomic stability in cancer cells (Shiromoto et al., 2021). This, naturally raises the question whether in the right context ADARs can also function as DNA mutators genome-wide and especially in the context of cancer, which is not unlikely, considering their overexpression and aberrant targeting in that context. All in all, both ADARs and APOBECs are shown to be contributing to tumorigenesis (Figure 1.4), by dynamically shifting their functions in multiple ways, which may involve altering their targeting activity, aberrantly editing RNA or mutating DNA or both, therefore enhancing tumor evolution and adaptation through genomic, transcriptional and subsequent proteomic heterogeneity (Tasakis et al., 2019).

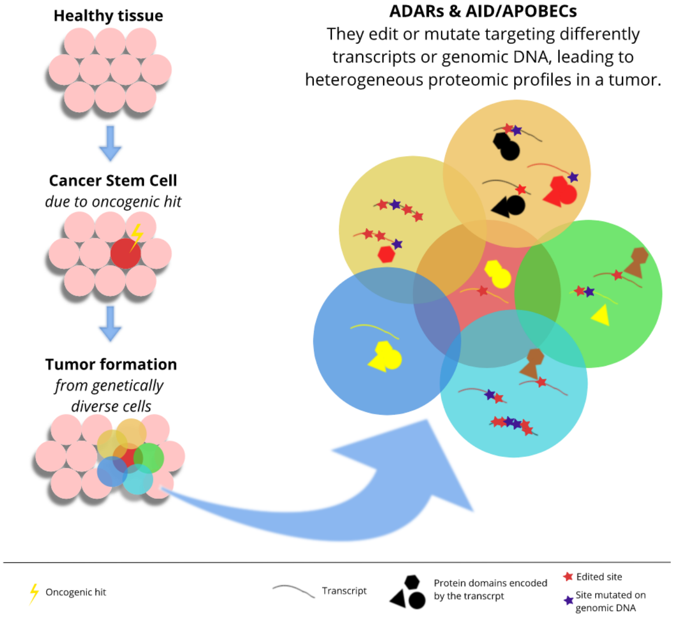


Figure 1.4 ADARs and AID/APOBECs contribute to tumor heterogeneity. According to clonal evolution models in cancer, tumorigenesis begins with a healthy cell being transformed to a cancer stem cell due to oncogenic hits forming a primary tumor. The primary tumor further evolves by receiving additional hits by deaminases, leading to altered coding information and heterogeneous proteomic profiles due to aberrant RNA editing or DNA mutation because of aberrant ADAR or AID/APOBEC functions. Figure from Tasakis et al., 2019. It is reused under the Creative Commons License 4.0. This illustration was created by myself.

1.2.4 APOBEC and ADAR mediated mutagenesis in the ssRNA viral genome of SARS-CoV-2

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the causal agent of the Coronavirus Disease 2019 (COVID-19), detected first in Wuhan, China in December 2019 (Wang et al., 2020). The rapid worldwide spread of SARS-CoV-2 led to the declaration of COVID-19 as a pandemic on March 11th 2020 by the World Health Organization (Cucinotta and Vanelli, 2020). By July 2021, more than 180 million individuals were infected, causing the death of nearly 4 million individuals ("WHO Coronavirus (COVID-19) Dashboard"). SARS-CoV-2 belongs to the *betacoronavirus* genera of the *Coronaviridae* family and is a single-stranded and positive-sense RNA (ssRNA+) virus (Machhi et al., 2020). Its genome (~29.9kb) is organized in ten ORFs (Open Reading Frames) by annotation, six of which encode for functional and structural elements, such as Nsp12 (Non-structural protein 12) encoding for the viral RNA-dependent RNA polymerase (RdRp deriving from ORF1ab, the Spike (S) glycoprotein from ORF2, the envelope protein (E) from ORF4, the membrane protein (M) from ORF5 and the nucleocapsid protein (N) from ORF9 (Chan et al., 2020; Kim et al., 2020; Hu et al., 2021). The S protein of SARS-CoV-2 is key to the viral infectivity, since it is recognized by the receptor ACE2 (Angiotensin-Converting Enzyme 2) in lungs, leading to membrane fusion with the cell membrane and subsequent internalization of the virus (Yi et al., 2020). Although it was initially thought that the proofreading activity of the RdRp is tight, thanks to Nsp14, which functions as a 3'→5' exonuclease proofreader and is of the viral Replication-Transcription Complex (RTC) (Ogando et al., 2020), a number of SARS-CoV-2 variants have been identified (Rambaut et al., 2020), several of which were characterized as variants of concern (VOCs) because of mutations in the S protein, based on which the current immunization strategies rely on (Darby and Hiscox, 2021).

Diversification of the SARS-CoV-2 genome is a challenging topic, especially when it triggers the equilibrium of population immunity, which is key to the resolution of the pandemic. As introduced in subchapter 1.2.2, ADARs and APOBEC3s are capable of deaminating the viral genomes, as part of their anti-viral properties, thus leaving their characteristic mutational fingerprint (Nishikura, 2010; Stavrou and Ross, 2015; Liu et al., 2018). 65% of the documented mutations for SARS-CoV-2, up to date, are C-to-U and A-to-G base changes (Klimczak et al., 2020; Wang et al., 2020), which likely the result of RNA deaminases (Giorgio et al., 2020). APOBEC3s deaminate single-stranded RNA or DNA (Jalili et al., 2020; Sharma et al., 2015) and the ssRNA genome of SARS-CoV-2 appears to be indeed a substrate of APOBEC3s, according to recent deamination motif analyses (Poulain et al., 2020). ADARs deaminate dsRNA (Keegan et al., 2001); and in the case of SARS-CoV-2 dsRNA instances can be formed during viral replication, which are actually recognized by MDA5 (Yin et al., 2021), known to interplay with ADAR1 in recognizing non-self dsRNAs within the cell (Liddicoat et al., 2015). Therefore, host-dependent RNA editing activity may diversify the SARS-CoV-2 ssRNA genome and that can be potentially traceable within a given population of infected individuals. The last is also a major interest of the present dissertation, which I explore under the scope of evolution.

1.3 Site-directed RNA editing technologies

Recent advances in the field of epitranscriptomics have emerged with several promising tools to perform targeted and site-directed RNA editing on the mRNA. The original idea and experiments were carried out by Woolf, Chase and Stinchcomb in 1995 (Woolf et al., 1995), for which they suggested that G-to-A DNA mutations can be transiently corrected on the mRNA by A-to-I (I recognized as G) RNA editing, for which they delivered oligoribonucleotides complementary to the target-region on the transcript. Their target-region was a premature stop codon (UAG) in the dystrophin mRNA, which upon activation through ADAR-mediated RNA editing (UAG to UGG, translated as stop codon to tryptophan) it led to expression of a downstream encoded luciferase reporter gene. Activation of the UAG stop codon through A-to-I RNA editing, is an advantageous idea that most of the recent site-directed editing technologies are still employing to test their specificity and efficacy; however now, in most cases it activates a gene of fluorescence (e.g. eGFP, as in 4.2.5 and 4.2.6), allowing immediate quantification of editing on the cell (through FACS) and transcript levels (Montiel-Gonzalez et al., 2019).

Site-directed mRNA editing tools have been developed and optimized primarily for ADAR-mediated (A-to-I) editing, while there are a few for C-to-U editing as well (Abudayyeh et al., 2019; Huang et al., 2020). For A-to-I site-directed editing, an oligoribonucleotide (hereafter guide-RNA or gRNA) antisense to the mRNA target is required, forming the dsRNA substrate ADARs prefer (Figure 1.5). This is the principle most such tools rely on and, furthermore, it has been shown that A-to-I editing is more specific and efficient for the A-targets that are mismatched to Cs (Cytidines) in the dsRNA, flanked by complementary oligomers in the dsRNA, the length of which (and thus of the gRNAs) varies between the different tools (Montiel-Gonzalez et al., 2019; Vogel and Stafforst, 2019). For the present chapter, I provide an overview of the A-to-I site-directed editing tools, which I group by whether they recruit the endogenously expressed ADAR or an exogenously introduced engineered version of the enzyme. I particularly focus on the "λN-ADAR" (Montiel-Gonzalez et al., 2013), "LEAPER" (Qu et al., 2019) and "RESTORE" (Merkle et al., 2019) tools, which I employed for my experiments, presented and discussed in this dissertation (see chapter 4.2).

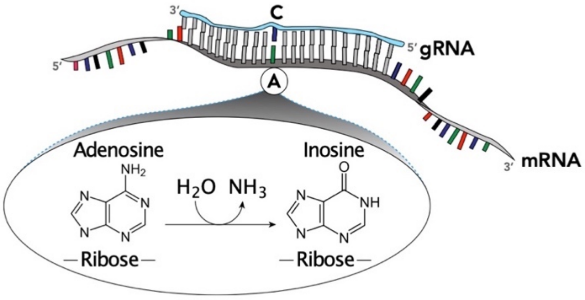


Figure 1.5 The principle of A-to-I site-directed RNA editing. An antisense guide-RNA (gRNA) is delivered against the mRNA target, to form the dsRNA substrate that ADARs require to edit. The Adenosine-target to be specifically deaminated is forming an A:C mismatch in the dsRNA substrate, while it is flanked by complementary base pairs. Adapted figure from Casati et al., 2021. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Epitranscriptomics. RNA Technologies, vol 12. Springer, Cham*, "ADAR-Mediated RNA Editing and Its Therapeutic Potentials", Casati B, Stamkopoulou D, Tasakis RN, Pecori R, Copyright: The Authors (2021).

1.3.1 Tools incorporating exogenous editing enzymes

A number of the site-directed RNA editing tools available, co-deliver gRNAs, which may be chemically modified or shape a certain structure through sequence, and engineered editing enzymes, which may derive from fusing different protein domains with specific function, one of which is typically an ADAR deaminase domain. One of the very first tools developed was the "λN-ADAR", developed by Maria Montiel-Gonzalez, Joshua Rosenthal and colleagues (Montiel-Gonzalez et al., 2013). The technology behind this tool relies on the N protein of the λ-phage (hereafter λN peptide), which binds to boxB hairpins on the RNA. Therefore, they fused the deaminase domain of ADAR2 ($ADAR_{DD}$) with a λN peptide, which is to bring the deaminase domain to edit the Adenosine-target on the mRNA, after binding to boxB hairpins on gRNA (Figure 1.6). With this set-up, they were successful in editing adenosines at about ~10% editing but not specifically to a single site. Therefore, their system underwent optimization and they achieved higher editing efficiency and specificity (up to ~70%), notably by: 1) designing guides with A:C mismatches between the mRNA and the gRNA, 2) by fusing more than one λN peptides with the $ADAR_{DD}$, 3) by evaluating the flanking nucleotides to the target (U**A**G was the most efficient, with A-target in bold) and 4) by inferring hyperactive mutants of the $ADAR_{DD}$, with the E488Q being the prominent one (Montiel-González et al., 2016). Additionally, to evaluate the aforementioned parameters on the cellular and RNA levels, they constructed a stable cell line from HEK293T cells, containing a cassette expressing mCherry followed by an inactivated eGFP gene due to a premature stop codon UAG, which they targeted for eGFP activation through U**A**G>U**G**G editing on the endogenous transcript of the cassette. This cell line, that I also use for my experiments (chapter 4.2.5 and 4.2.6), is hereafter termed as HEK293T-W58X and they employed it to evaluate on-target editing efficiency and off-target events, which were still present, but reduced later on after fusing a nuclear localization signal (NLS) with the 4λN-ADAR (Vallecillo-Viejo et al., 2017). For my experiments, I use the 4λN-ADAR E488Q mutant (chapter 4.2.5).

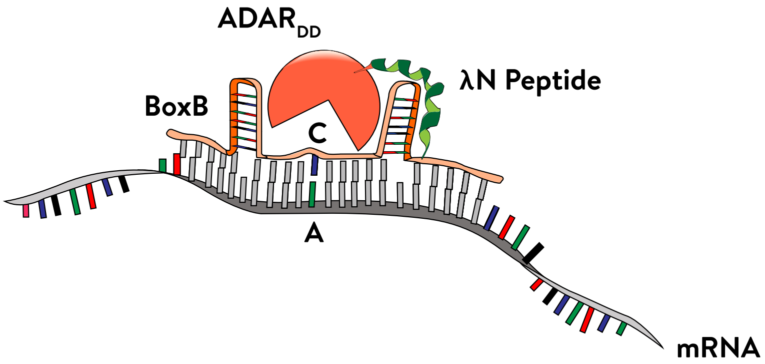


Figure 1.6 Simplified scheme recapitulating site-directed mRNA editing with the λN-ADAR tool. The deaminase domain of ADAR2 ($ADAR_{DD}$) is fused with λN peptides, which bind to BoxB hairpins of the gRNA, antisense to the mRNA target. The specific Adenosine to be edited is in a deliberate A:C mismatch within the dsRNA formed between the mRNA and the gRNA. Adapted figure from Casati et al., 2021. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Epitranscriptomics. RNA Technologies, vol 12. Springer, Cham*, "ADAR-Mediated RNA Editing and Its Therapeutic Potentials", Casati B, Stamkopoulou D, Tasakis RN, Pecori R, Copyright: The Authors (2021).

Other popular and widely-used tools that incorporate an exogenous editing enzyme are the “SNAP-ADAR” and “REPAIR”, which was evolved to “RESCUE”. Thorsten Stafforst (Stafforst and Schneider, 2012) developed a system that incorporates SNAP-ADAR, which is a fusion of the deaminase domain of the human ADAR1 and a self-labelling protein SNAP-tag, that stems from a human alkyltransferase (O^6-alkylguanine-DNA). SNAP-ADARs can covalently bind to gRNAs tagged with O^6-benzylguanine and, thus, editing the target specifically with minimal off-target effects *in vitro*. Further optimization of this tool led to higher guide specificity and editing efficiency in endogenous transcripts as well (Vogel et al., 2014, 2018). A CRISPR/Cas-based tool, named “REPAIR”, is also available in the “toolbox” of site-directed RNA editing technologies; Feng Zhang and colleagues fused the deaminase domain of the hyperactive ADAR2 (E488Q mutant) with a Cas13 enzyme (dCas13b), which is recruited to the Adenosine-target on the mRNA with a gRNA that contains a stem loop due to a repetitive sequence and specifies the target with the aforementioned A:C mismatch (Cox et al., 2017). The same group of scientists further optimized the system, primarily to efficiently edit endogenous transcripts and also further mutated residues of the ADAR2 deaminase domain, which allowed them to transform their original editing enzyme to perform C-to-U RNA editing as well; this version of was named RESCUE (Abudayyeh et al., 2019).

1.3.2 Tools recruiting endogenous and unmodified ADARs

Thanks to the rapid development of RNA technologies, such as the aforementioned ones, editing RNA molecules *in vitro* or endogenous transcripts in cell lines at specific sites, has empowered novel concepts for development of RNA therapeutics. As discussed in *Nature* (Reardon, 2020), One of the advantages of RNA engineering is that potential off-target effects of editors on the transcript are rather transient, especially when compared to genome editing technologies, such as CRISPR, where there undesired off-targets can be permanently fixed into the genome. The “transiency” of such therapeutics, may appear as a limitation at first, however many scientists in the field see it as an opportunity for flexible therapies that could alleviate pain, metabolic disorders, in addition to cancer or other genetic diseases. In fact, editing RNA can happen without “heavy machinery” (as in CRISPR for instance) and this gives hope for more efficient *in vivo* delivery of site-directed editing components in therapies. Currently, there are multiple tools available, that promise site-directed RNA editing with simply delivering gRNAs which can recruit the endogenously expressed ADARs, instead of co-delivering an engineered enzyme.

One of the first tools that achieved site-specific mRNA editing incorporating the unmodified human ADAR2 is the “GluR2-ADAR” (Wettengel et al., 2017). As also discussed in the chapter 1.2.1, ADAR2 naturally targets the pre-mRNA of the glutamate receptor B (gluR-B), which contains a specific hairpin, also known as R/G editing site (Stefl et al., 2010). Therefore, Wettengel et al engineered gRNAs that contain hairpins as encoded by the gluR-B to tether ADAR2 to edit the Adenosine-target (again, in A:C mismatch between the mRNA and the gRNA). They showed that when they co-delivered the enzyme with the aforementioned type of gRNAs they could yield up to 65% editing on-target in premature stop codons *in vitro* and about 10% editing in endogenous transcripts. But most importantly, they showed that when delivering only the gRNAs, site-directed editing was still possible by the endogenously expressed ADAR. The same group, of Thorsten Stafforst, further optimized this system and developed a method called "RESTORE" (Merkle et al., 2019), for which they show that A-to-I site-directed editing is feasible by the endogenous ADAR1, when they target mRNAs with chemosynthetic antisense oligonucleotides (ASOs), as shown in Figure 1.7. The ASOs (shown as "GluR2-adRNA" in the Figure 1.7) carry the GluR2 motif (ADAR-recruiting domain) and specific chemical modifications which are: phosphorothioate on 4 terminal residues at the 3' end and 2'-O-methylations throughout the ASO.

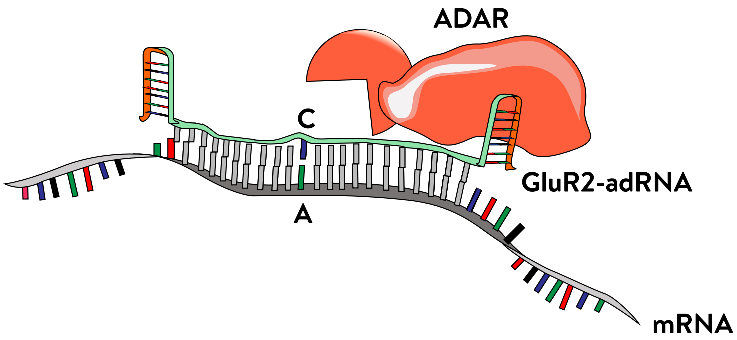


Figure 1.7 Schematic representation of the principle behind the RESTORE tool. A chemically modified antisense oligoribonucleotide to the target (mRNA) containing the GluR2 motif (shown overall as GluR2-adRNA) is employed to recruit the endogenously expressed ADAR1 for site-directed mRNA editing. The target-site is defined on the mRNA target-transcript with an A:C mismatch as previously described. Adapted figure from Casati et al., 2021. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Epitranscriptomics. RNA Technologies, vol 12. Springer, Cham*, "ADAR-Mediated RNA Editing and Its Therapeutic Potentials", Casati B, Stamkopoulou D, Tasakis RN, Pecori R, Copyright: The Authors (2021).

In a more recent tool, LEAPER (Leveraging Endogenous ADAR for Programmable Editing of RNA), Qu et al. showed that endogenous ADARs can also be recruited by non-chemically modified gRNAs, which they called ADAR-recruiting RNAs (arRNAs) and are expressed with plasmid vectors (Qu et al., 2019). According to their method (general scheme shown in Figure 1.8), the typical length of an arRNA is between 71-111 bases, in which the A-target is centered with an A:C mismatch. Qu et al, explored all the possibilities for concluding that this layout is the most efficient for their method; in particular, they tested an A:G mismatch for defining the Adenosine-target (shown in Figure 1.8 as A:G) concluding to no on-target editing and confirmed that the U**A**G (Adenosine-target in bold) is the most efficiently editable trinucleotide. Additionally, they showed that the longer the arRNA, the more on-target editing is achieved, but also it is more likely to have off-targets within the region on the mRNA the arRNA binds. Finally, they confirmed that ADAR1 is the endogenous enzyme that their system recruits. Overall, they present a single-molecule method for site-specific editing, which can be very efficient (~80% on-target editing in overexpressed target transcripts and up to 30% in endogenous transcripts).

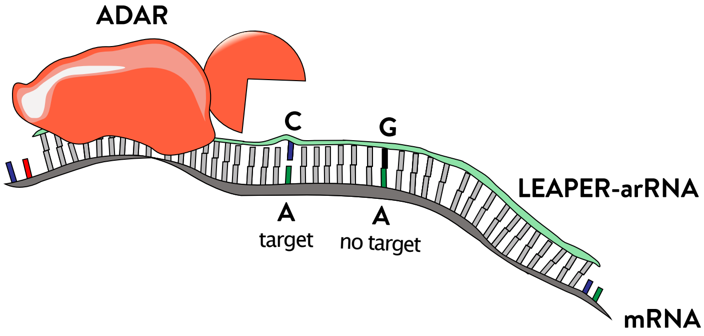


Figure 1.8 Scheme showing the principle of the LEAPER method. ADAR-recruiting oligoribonucleotides (LEAPER-arRNA) are employed to define the target-region on the mRNA. The Adenosine-target should be in an A:C mismatch, while A:G mismatches are not efficient targets. arRNAs recruit the endogenous ADAR. Adapted figure from Casati et al., 2021. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, *Epitranscriptomics. RNA Technologies, vol 12. Springer, Cham*, “ADAR-Mediated RNA Editing and Its Therapeutic Potentials”, Casati, Stamkopoulou, Tasakis, Pecori, Copyright: The Authors (2021).

All in all, there is a number of different tools available for precise and efficient site-directed mRNA editing. Such RNA editing methods are a bright hope for next-generation therapeutics and with more in the making (Katrekar et al., 2021), their application in medical care is simply a matter of time. However, before that happens, it is crucial to know all the possibilities that such applications may bring about. In particular, I am employing site-directed mRNA editing methods (4λN-ADAR and LEAPER) (chapter 4.2) to explore off-target effects on genomic DNA, which I present as proof-of-concept for ADARs being able to mutate cancer genomes.

1.4 RNA editing detection from Next-Generation Sequencing data

As discussed up to now, RNA modifications are of great significance in health and disease, with new concepts and related technologies constantly emerging, such as the aforementioned site-directed RNA editing tools (chapter 1.3). Sequencing technologies are key to detecting and validating RNA modifications and, in particular RNA editing, a mechanism that naturally leads to base changes on the transcript, easily detectable from cDNA amplicons. Traditionally, C-to-U and A-to-I(G) editing has been detected from Sanger sequencing chromatograms, in which the edited site appears as a double peak of the original and the edited bases, as for example in (Athanasiadis et al., 2004; Ohlson et al., 2007; Rosenberg et al., 2011; Ekdahl et al., 2012; Fu et al., 2017; Kluesner et al., 2021), and demonstrated in Figure 1.9. To ensure that such sites are indeed a result of RNA editing, genomic DNA (gDNA) amplicons from the same region and same lysate of cells are also necessary, in which the double peak should be absent from the respective sites and, therefore, they are not Single Nucleotide Polymorphisms (SNPs) or somatic mutations. For a certain site of the Sanger chromatogram, RNA editing is quantified as the percentage (%) of the peak height of the edited base to the sum of the peak heights of both the edited and original bases (Kluesner et al., 2018).

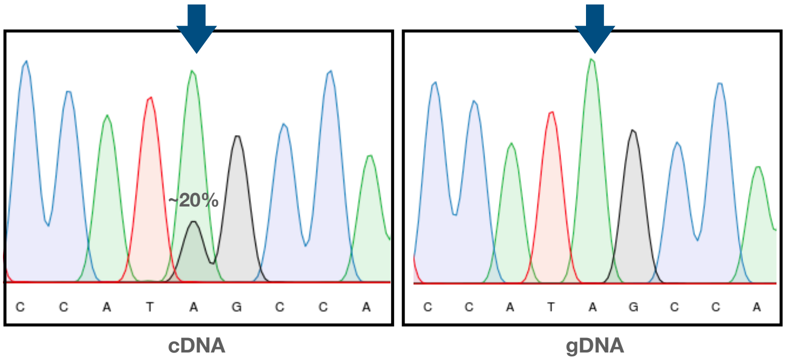


Figure 1.9 Detection of RNA editing with Sanger Sequencing from cDNA amplicons. RNA editing, in this example A-to-I(G), is detected from a cDNA amplicon as double peak of the original and the edited bases in a site from Sanger sequencing chromatograms (noted with a blue arrow; cDNA panel). A genomic DNA (gDNA) amplicon with absence of the double peak from the same site (blue arrow; gDNA panel) is necessary to validate this site as an RNA editing event. The percentage (%) of editing is the percentage of the height of the edited base to the overall sum of heights of the edited and original bases in the same site. In this case it is about 20% editing. The data and the illustration used in this figure are generated by myself from experiments presented in chapter 4.2.5.

1.4.1 *De novo* detection of RNA editing from NGS data

Although Sanger sequencing is an easy, fast, accurate and affordable method for validating RNA editing events in sites where one would expect them to, it is not powerful enough for *de novo* detection of RNA editing transcriptome-wide. The first computational analyses that employed expressed sequenced tags (ESTs) and large-scale human cDNA data yielded 12723 putative A-to-I editing sites in 1637 genes primarily found in *Alu* repeats (Levanon et al., 2004), which exponentially magnified the known ADAR editome from just 19 sites, as previously known (Morse and Bass, 1999). However, the aforementioned analysis, or other ones arriving to the same conclusions (Athanasiadis et al., 2004), due to gDNA data unavailability, did not exclude the possibility that single nucleotide variations (SNV) detected from the cDNA could in fact be SNPs. A few years later, when Next Generation Sequencing (NGS) was better established, Jin Billy Li, Erez Levanon and colleagues (Li et al., 2009), performed the first analysis for transcriptome-wide RNA editing detection from NGS data by comparing site-by-site base calls from RNA and gDNA from the same human samples. They reported overall 239 sites after stringent filtering and excluding regions of *Alu* elements, which are error-prone due to their repetitive sequence. They validated with sanger sequencing and found that 15 out of 18 sites detected were consistently edited throughout the individuals of their cohort in multiple tissues, while the rest 3 sites were tissue-specific editing events. The principle of RNA and DNA comparisons for *de novo* detection of RNA editing is still followed, while the accuracy of detection is only getting better, thanks to the constant optimization of NGS strategies and the emergence of standardized and widely used bioinformatics tools and pipelines (Auwera et al., 2013).

Although RNA/DNA comparisons provide a realistic resolution for the RNA editome, it is not always easy to figure out which enzyme is responsible for an editing event. For example, a C-to-U editing can occur due to deamination of a couple of APOBEC enzymes (eg APOBEC1, APOBEC3A) and in this case RNA/DNA comparisons cannot be conclusive about the writer of the modification. However, the writers of A-to-I and C-to-U editing are well-known and thanks to CRISPR/Cas tools one is able to knock them out (KO) when possible and look for editing sites present in transcriptome of the wildtype (WT) version of the same cells, which are absent from the KO version. Such comparisons (RNAwt/RNAko), allow the accurate detection of the relevant editing sites from NGS data deriving from a particular writer, while they eliminate the background noise. This method has been previously employed for APOBEC1 and allowed the identification of novel targets (sites and transcripts), the deamination motif and preferred transcript topology APOBEC1 has in mouse small intestine enterocytes (Rosenberg et al., 2011). Furthermore, RNAwt/RNAko comparisons of the same enzyme in mouse dendritic cells, revealed the underlying diversity of RNA editing in the transcripts of the same cell population, which underlines the transcriptomic sequence heterogeneity at a single-cell resolution (Harjanto et al., 2016).

1.4.2 RNA editing in large-scale data and databases

Thanks to advances in sequencing technologies, such as RNA-seq, and bioinformatics tools, as described above, the impact of RNA editing in fundamental biological processes, as well as in health and disease, is more and more evident and it is currently at the spotlight of public attention (Reardon, 2020). High-throughput data, often publicly available, have enhanced large-scale processing of cohorts of data, reporting loads of RNA editing in a diversity of datasets, including sets with great clinical relevance, such as human tumors (Paz-Yaacov et al., 2015). The raw counts of RNA editing sites, detected from NGS data as previously described, has till very recently been the measure of the RNA editing load in a number of studies, as for example in (Lazzari et al., 2017). However, this measure of

RNA editing activity is reliable for comparisons only between samples that derive from the same experimental set up, due to potential batch effects or other technical artifacts. Inter-sample comparisons from different experiments or even studies, has become possible with a new measure for A-to-I RNA editing quantification, the Alu Editing Index or AEI (Roth et al., 2019). AEI relies on the editability of *Alu* SINEs, which is the hotspot of A-to-I RNA editing (Bazak et al., 2014). Therefore, AEI is defined as the ratio of A-to-G mismatches to the total overage of Adenosines in predetermined regions, which for human would be the *Alu* SINEs (Roth et al., 2019). This tool, however, is not limited only for human data, it can be employed for other organisms and for any other set of predetermined regions, provided that the selected regions are highly edited so that the signal-to-noise ratio can be adequately high (Roth et al., 2019). Elevated RNA editing activity as usually seen in human tumors, is also accompanied by regional "hyper-editing" activity, which in RNA-seq data is represented with reads that would contain several mismatches compared to the reference sequence. Such reads would fail to align against the region of a reference genome with a typical RNA-seq aligner, like STAR, as they would be considered problematic, presumably with sequencing errors. These reads (also termed "hyper-edited"), when re-evaluated and re-aligned they reveal a number of new sites, which can be of great importance (Porath et al., 2014). Re-alignment of potential hyper-edited reads and AEI for inter-sample comparisons, is also taken into consideration for data related to the present dissertation.

The extend of known RNA editing targets is constantly increasing. Large-scale data analyses and extensive experimental set ups have led to compiling public databases, robustly documenting targets of RNA editing enzymes in a high resolution for several organisms. DARNED (Database of RNA editing; Kiran et al., 2013) and RADAR (Rigorously Annotated Database of A-to-I RNA editing; Ramaswami and Li, 2014) are one of the first databases for documenting primarily A-to-I RNA editing events in human, mouse and drosophila. REDIportal, developed and maintained by the developers of REDItools, started as an Atlas of the human of RNA editing in human tissues by Ernesto Picardi, Graziano Pesole and colleagues (Picardi et al., 2015b, 2017), which was further expanded with editing data from additional human cell lines (Lo Giudice et al., 2020b, 2020a; Schaffer et al., 2020) and editing data from nascent mouse pre-mRNAs (Licht et al., 2019). Moreover, editing also impacts on the sequence and function of non-coding RNAs with emerging clinical relevance in cancer (Han et al., 2015; Nishikura, 2016). MiREDiBase (miRNA Editing Database) documents putative and experimentally validated RNA editing events found in miRNAs (Marceca et al., 2021).

2. Aims of the dissertation

It is overall evident by now that RNA editing is a crucial component of transcriptome diversification and regulation, since it can naturally impact potentially all mRNA topologies (i.e. primarily introns and UTRs, less frequently exons) or non-coding RNAs, which can shift dynamically to facilitate disease development and progression, such as in cancer (Han et al., 2015; Paz-Yaacov et al., 2015; Nishikura, 2016). However, multiple deaminases present a versatile role; they can be both RNA editors and DNA mutators, often seen in cancer; APOBEC1, originally thought to be only an RNA editor, is now known to be able to mutate esophageal adenocarcinoma genomes (Saraconi et al., 2014) and when APOBEC1 is deleted in mice models, tumor burden is significantly reduced (Blanc et al., 2007). APOBEC3A and APOBEC3B are DNA mutators involved in a phenomenon termed as "kataegis" mutations in cancer genomes (Nik-Zainal et al., 2012), but as recently shown APOBEC3A can also edit RNA (Sharma et al., 2015), and this appears to be necessary for mutating DNA (Jalili et al., 2020). Furthermore, for ADARs it is known that they can mutate DNA *in vitro* when that is in DNA/RNA hybrids (Zheng et al., 2017), a role that ADAR1 presents in telomere stability in certain cancer cell lines (Shiromoto et al., 2021). ADAR1, however, is overexpressed in virtually all cancers, with very few exceptions (i.e. kidney) (Han et al., 2015), subsequently associated with elevated editing activity (Paz-Yaacov et al., 2015). A possibility that I explore with this dissertation is that ADARs - and particularly ADAR1 as a ubiquitous RNA editor - are contributing to cancer development and progression not only as RNA editors, but also as DNA mutators. With RNA editing being co-transcriptional (Laurencikiene et al., 2006), DNA mutations by ADARs must be arising in genes of highly edited transcripts, in the vicinity of RNA editing sites, because of ADARs opportunistically accessing the genome within R-loops formed during transcription. In other words, I assume that the more edited a transcript is, the more likely its gene will be mutated by ADARs, a phenomenon which may occur as a "collateral damage" of a hyper-editing ADAR. The goal of this dissertation is to address this hypothesis through the following aims:

- **Aim 1**: I explore the mutagenic role of ADAR1 in Multiple Myeloma (MM), a cancer in which it is substantially overexpressed, and I draw correlations from NGS data from MM patients, in order to detect ADAR1-dependent DNA mutations in genes whose transcripts are edited by ADAR1 on the RNA level.
- **Aim 2**: I present proof-of-concept experiments in cell lines, for which I recruit ADARs with guide-RNAs targeting specific transcripts, in order to perform site-directed mRNA editing, and I look for DNA mutations in the cognate locus.
- **Aim 3:** I address this concept from an evolutionary perspective, for which I explore the impact of host-dependent RNA editing in ssRNA viral genomes from SARS-CoV-2 isolates from a given population, considering that mutations may gradually accumulate within the population, similar to how mutations could clonally expand within a cell population from a tumor.

3. Materials and Methods

3.1 General practices for calling RNA editing sites from RNA-seq

Calling RNA editing sites from total RNA-seq (NGS data) relies on pipeline of steps that generally involves: quality control (QC) of the reads sequenced from RNA-seq libraries, mapping (also known as aligning) of the reads to the different regions of the appropriate reference genome, post-alignment proper data treatment (eg removal of PCR duplicates), base calling per genomic coordinate (aforementioned as sites), filtering of SNPs or noise to conclude to the list of RNA editing candidate sites and, last, functional their functional annotation. Here, I summarize step by step a typical pipeline, noting the appropriate software and options I used, allowing me to detect RNA editing sites from NGS data. As discussed in subchapter 1.4.1, there are two main strategies to detect RNA editing sites: (1) through RNA vs DNA comparisons (RNA/DNA), for which DNA variant sites (SNPs) are subtracted from the SNV calls from the RNA-seq, leading to the RNA editing candidates (Diroma et al., 2019) and (2) through comparing variant sites called from RNA-seq of a WT cell line vs the variant sites from RNA-seq from a writer-KO line of the same type (RNAwt/RNAko), as performed in (Rosenberg et al., 2011; Harjanto et al., 2016). RNA-seq libraries are typically prepared in replicates per condition (for example triplicates per RNAwt and RNAko). For library preparation there is a number of strategies and manufacturers. One of the most popular is the Illumina® Platforms and their HiSeq sequencing technology, as described in Lerner et al., 2021. The pipeline I used for calling RNA editing, as well as DNA mutation calls, is summarized below and also shown in Figure 3.1.

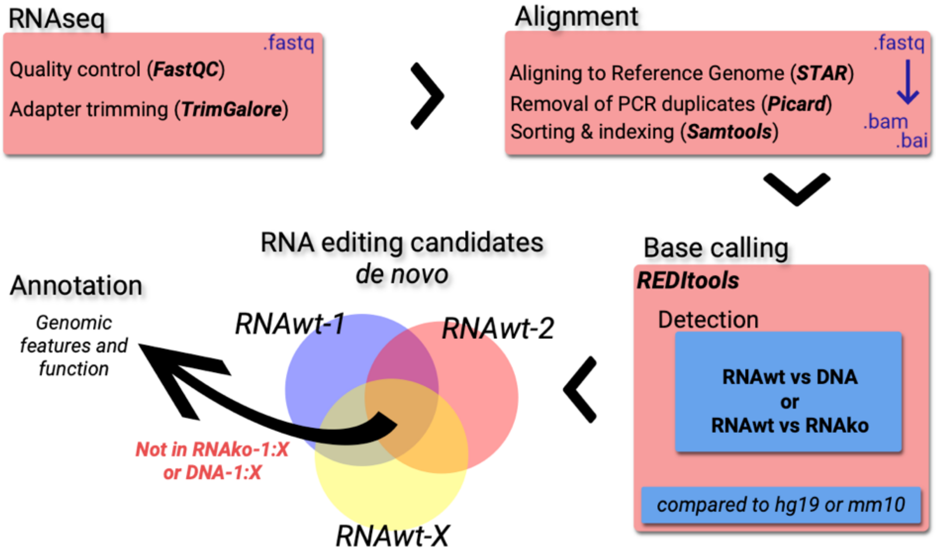


Figure 3.1 Scheme summarizing a generalized pipeline for RNA editing calling from RNA-seq. Quality control with FastQC and adapter trimming with TrimGalore takes place prior to alignment. RNA-seq data are aligned with STAR and aligned data are deduplicated with Picard, sorted and indexed with Samtools. Base calling from aligned RNA-seq is performed with REDItools, either by RNAwt vs DNA or RNAwt vs RNAko comparisons (see chapter 1.4.1), against the same reference genome (mm10 for mouse and hg19 for human). RNA editing candidates are called *de novo*

by obtaining the consensus edited sites in the RNAwt (replicates 1-X) which show no variation in the RNAko or DNA (replicates 1-X). The candidates are annotated for genomic features and function. Adapted figure from Lerner et al., 2021. Adapted by permission from Springer Nature Customer Service Centre GmbH: Springer Nature, RNA Editing. Methods in Molecular Biology, vol 2181. Humana, New York, NY, "C-to-U RNA Editing: From Computational Detection to Experimental Validation", Lerner, Kluesner, Tasakis, Moriarity, Pecori, Copyright: Springer Science Business Media, LLC, part of Springer Nature (2021).

Pipeline steps

All software and resources I used, described below, are summarized in Table 3.1.

1. **QC and adapter trimming:** I performed quality control (QC) of the raw RNA-seq data (unprocessed data upon sequencing) with the widely used QC tool for RNA-seq data "FastQC" (resource availability at Table 3.1). This allowed me to obtain metrics and statistics for important parameters which are taken into consideration for the downstream analysis, such as the overall sequencing and data quality, the GC content, the levels of PCR duplicates and the overrepresented sequences, which typically include the sequencing adapters from the library preparation. I performed trimming of the sequencing adapters using the tool TrimGalore (Table 3.1), so as to maximize the number of mapped reads and their mapping quality. At this step the file type is "fastq" and there are two fastq files per sample, if the libraries were sequenced in paired-end layout. Trimming of the pairs takes place simultaneously per sample.
2. **Alignment (mapping):** I aligned the "trimmed" reads ("fastq" files) against the appropriate reference genome with the ultrafast RNA-seq aligner "STAR" (Dobin et al., 2013) or GSNAP (Wu and Watanabe, 2005). The pairs of each sample are both employed for producing the paired-end aligned data, now represented in a single "bam" filetype. The reference genomes are publicly available from multiple resources, notably ENSEMBL and UCSC (see Table 3.1). For my analyses I employed the latest versions available (in fasta format) at the time of analysis, which for human was "hg19". Genomes were appropriately indexed prior to alignment, as prescribed by the different aligners, so as to generate genome-specific sequence and genomic feature information.
3. **Post-alignment treatment**: after alignment, the mapped data (now in bam files; one per sample) undergo removal of duplicated reads, which are typically present due to a PCR step in library preparation. I performed this step using Picard tools or Samtools (Li et al., 2009) software (see Table 3.1). This step is important because it allows me to retrieve the realistic representation of variant frequency, by excluding variation which may occur due to technical artifacts. Furthermore, I indexed and sorted the deduplicated alignments (in bam format, obtaining additional "bai" files), as required for the base calling step, described below.
4. **Base calling coordinate-wise**: at this step transcriptome-wide information for every genomic coordinate is collected, and in particular the base composition of every coordinate covered by reads. To obtain this lists of base calls per sample from the aligned and properly treated RNA-seq as mentioned above, I employed REDItools (Picardi and Pesole, 2013; Picardi et al., 2015a;

Lo Giudice et al., 2020b). REDItools provide multiple options depending on the type of comparison (RNA/DNA or RNAwt/RNAko). Using REDItools, I obtained the base calls from each sample and their replicates separately. For the coordinates to further considered for calling the RNA editing sites, I required that they are covered by at least 10 reads and there must be at least 3 bases in one coordinate that support the variation (SNV), so as to exclude potential sequencing errors. For the same reason, I excluded SNVs that are in the first 5 bases of a read or present in homopolymeric regions of more than 5 of the same nucleotides. Additionally, in paired-end experiments I required that the SNV must be supported by at least one read-pair. Last, I did not consider sites with poor mapping or sequencing quality.

5. **RNA editing candidates:** For the variant calling between the RNA/DNA or RNAwt/RNAko I considered only the coordinates with base calls that passed the aforementioned criteria. In this step, the candidates for RNA editing sites were compiled by those coordinates that gave a positive editing signal in the RNAwt (i.e. when the reference base is A, there's an A>G SNV of at least 10% variation frequency), which is absent from the same and equally well-covered coordinate in RNAko (or DNA). For *de novo* editing detection, the RNA editing candidates must be give a positive editing signal in all replicates, while this is absent from all replicates of the RNAko or DNA. For DNA-seq data, somatic variants were called following the typical GATK pipeline for best practices from the Broad Institute (DePristo et al., 2011; Auwera et al., 2013).
6. **Annotation:** the last step of the pipeline involves the profiling of those sites with regards to their genetic information and their potential functional impact. There are multiple annotation tools available, designed for several types of data. REDItools provide their own annotation scripts (Picardi et al., 2015a), which I used for every coordinate-based list of multiple features, which for example were *Alu* repeats for human (RepeatMasker database), SNP databases (dbSNP), transcript topology and isoform (i.e. from RefSeq) or other custom made lists from UCSC Table browser. Specifically for human data, I used the oncotator tool, which allows parallel coordinate-based annotation from multiple sources (Ramos et al., 2015).

Table 3.1 Software and resources availability described in the RNA editing calling pipeline, summarized by: software for RNA-seq processing, RNA editing and mutation calling tools, reference genome resources (sequence & features), annotation software and resources. Sofware citations (when available) are given in the main text (subchapter 3.1).

Software for RNA-seq processing	
FastQC	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
TrimGalore	https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
STAR	https://github.com/alexdobin/STAR
Picard	https://github.com/broadinstitute/picard
Samtools	http://www.htslib.org/
RNA editing and mutation calling	
REDItools	https://github.com/BioinfoUNIBA/REDItools
GATK Best Practices	https://gatk.broadinstitute.org/
Reference genome resources (sequence & features)	
ENSEMBL	https://www.ensembl.org/info/data/ftp/
UCSC	https://genome.ucsc.edu/
Annotation software and resources	
Oncotator	https://github.com/broadinstitute/oncotator
RepeatMasker	https://www.repeatmasker.org/
RefSeq	https://www.ncbi.nlm.nih.gov/refseq/
dbSNP	https://www.ncbi.nlm.nih.gov/snp/
UCSC Table Browser	http://genome.ucsc.edu/cgi-bin/hgTables

3.2 Methods Aim 1

3.2.1 Multiple Myeloma patient data

RNA-seq (paired-end and non-stranded) and WES data from Multiple Myeloma (MM) patients were obtained through access to the CoMMpass MMRF study (https://themmrf.org/; dbGaP accession number phs000748; http://www.ncbi.nlm.nih.gov/gap;), thanks to a collaboration with Dr. Alessandro Laganà and Dr. Samir Parekh at the Icahn School of Medicine at Mount Sinai, New York, USA. Dr. Laganà previously processed RNA-seq, WES and clinical data from overall 590 patients from the aforementioned MMRF study for expression, Copy Number Variation (CNV), and A-to-I quantification analyses (measured with the Alu Editing Index; Roth et al., 2019) to correlate with patient survival data. These methods Dr. Laganà followed, are published in our joint preprint (Tasakis et al., 2020). I focused on a subset of 23 MM patients (Patient identifiers are shown as row names in the heatmap of Figure 4.2D) from the aforementioned cohort, who each had matched RNA-seq and WES data from two timepoints of the disease: tumors at presentation (Timepoint 1 or TP1) and tumors at relapse (Timepoint 2 or TP2), so as to correlate A-to-I RNA editing events in TP1 with mutation candidates in TP2. The methods I performed for this patient data are presented and explained in the subchapter 3.2.2 and also described in Tasakis et al., 2020.

3.2.2 RNA editing, DNA mutation calling and correlation analyses in 23 Multiple Myeloma patients with matched RNA-seq and WES data.

I processed RNA-seq and WES data from 23 MM patients at two timepoints of the disease, whom I focused on for the correlation analyses of RNA editing and DNA mutation by ADAR1 (see 3.2.1). I aligned their RNA-seq against the human reference genome GRCh37 (hg19) and all annotation and gene models were based on Ensembl version 74 (see Table 3.1 for reference genome resources and availability). I aligned RNA-seq data using the aligner GSNAP v. 2017-06-20 (Wu and Watanabe, 2005) using the default parameters. I marked PCR read-duplicates with Picard, and I sorted and indexed the aligned data with Samtools v. 0.1.19 (Li et al., 2009). With the help of Dr. Laganà, I realigned RNA-seq unmapped reads to further include hyperedited reads as preciously described (Porath et al., 2014) and I processed the WES data according to the recommendations of the 'GATK Best Practices' (DePristo et al., 2011; Auwera et al., 2013).

In this case, I performed an RNA/DNA comparison to call RNA editing, following the principles explained in subchapter 3.1. I employed REDItools v1 and followed the developers' recommendations to pre-process the data as explained above and call RNA editing events in both timepoints of the MM patients (Picardi and Pesole, 2013; Picardi et al., 2015a). In bried, I employed the REDItoolDenovo.py script to Single Nucleotide Variations (SNVs) when compared to reference genome (hg19). I only further considered SNVs from well-covered sites (≥10 reads), with variation being supported by concordant read-pairs, having at least 10% variation frequency, which is also supported by at least 3 reads and, finally, reported by REDItools as statistically significant (p-value≤0.05). RNA editing candidates were those sites from the aforementioned selected set, that showed no variation in their matched mutation calls from WES data. With the help of Dr. Laganà I called mutations from WES data using the mutation calling pipelines Strelka2 (Kim et al., 2018) and VarDict (Lai et al., 2016). I considered coordinates reporting variation from the WES data from well-covered sites (≥10 reads) of good mapping and sequence quality from any variation frequency percentage. I correlated A-to-I RNA editing events from the TP1 with mutations from TP2 within ±20bp distance per patient, using the R programming language (v. 4.0.2; R Core Team, 2020) and the package 'Tidyverse' for data processing and visualization (v. 1.3.0, Wickham et al., 2019). I validated the matching "editing-to-mutation" candidates with the tool bam-readcount (https://github.com/genome/bam-readcount) and I annotated the validated sites with the tool Oncotator v.1.9.9.0 (Ramos et al., 2015). Last, I performed pathway enrichment analysis using the tool SLAPenrich (Iorio et al., 2018).

3.2.3 Experimental validation of RNA editing in Multiple Myeloma cell lines

I used two typical Multiple Myeloma cell lines, KMS-20 and KMS-11, which were a kind gift from Dr. Parekh, to experimentally validate RNA editing sites in transcripts I found edited according to the *in-silico* analysis in patients from the MMRF Cohort (see 3.2.1). I cultured cells from the aforementioned cell lines in suspension in complete growth media, which is RPMI 1640 with L-glutamine and sodium bicarbonate (from Sigma-Aldrich), supplemented with 10% Fetal Calf Serum (FCS; from PAN Biotech) and 1% Penicillin-Streptomycin (from Sigma-Aldrich). I used $2x10^6$ cells per sample to treat with growth media supplemented with 10U and 100U IFNα (from ThermoFisher Scientific) for 96h. I extracted simultaneously total RNA and gDNA from about $5x10^5$ cells per sample with the AllPrep DNA/RNA Mini Kit (Qiagen). I treated the extracted RNA with TURBO DNAse following the manufacturer's instructions (ThermoFisher Scientific). I then generated amplicons for *EIF2AK2* (see chapter 4.1.3) from gDNA with the Q5 High-Fidelity DNA polymerase (New England Biolabs), as well as cDNA amplicons from the RNA with the OneStep RT-PCR kit (Qiagen), both with primers Eif2ak2-Fw and Eif2ak2-Rv (see Appendix A). I purified the PCR products with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel), which I sent for sequencing with both forward and reverse primers at Microsynth Seqlab GmbH, in Goettingen, Germany. Last, I quantified RNA editing across the Sanger sequencing chromatograms using the tool MultiEditR (Kluesner et al., 2021). For these experiments, I received help from Dr. Violetta Leshchenko in culturing and maintaining the cell cultures and Ms. Pavithra Nedumaran helped me with the IFNα treatment experiment.

3.3 Methods Aim 2

3.3.1 Cell lines and culture

Ramos Burkitt's lymphoma human cell lines

The Ramos wildtype B-cell line (RA 1 ATCC CRL-1596) was purchased from ATCC. Ramos AID-/- clones were a kind gift by Dr. Jeroen Guikema, Academic Medical Center, Amsterdam, The Netherlands. I generated the Ramos ADAR1-/- cells by transfecting (see transection methods in 3.3.3) plasmids co-expressing Cas9 (pSpCas9), gRNAs targeting *ADAR1* in exon 3 (gRNA-01-Fw and gRNA-01-Rv, Appendix B) and exon 4 (gRNA-02-Fw and gRNA-02-Rv, Appendix B), as well as eGFP. As control I used gRNA-NT (-Fw and -Rv, Appendix B), which has no target in the genome. The relevant plasmids were previously generated by my colleague Dr. Riccardo Pecori, using as backbone the vector pSpCas9(BB)-2A-GFP (PX458, Plasmid #48138, Addgene – listed as Crispr-pl in Appendix C) originally from (Ran et al., 2013). 24h upon transfection, I sorted the GFP-positive cells (chapter 3.3.5) in 96-well plates, containing conditioned media, as in a single cell per well layout, which I further propagated. I upscaled successfully grown clones and, with the help of the student Ms. Dimitra Stamkopoulou, I tested them for absence of ADAR1 protein with Western blot (ADAR1 antibody #14175l from Cell Signaling Technology) with control protein the GAPDH (antibody #2118, Cell Signaling Technology), for which I also validated absence of A-to-I editing in the *MAVS* transcript (Figure 4.7), a known ADAR1 target (Li et al., 2021).

I cultured all Ramos cell lines in suspension, within their optimal range of growth (2x10^5-1x10^6 cells/mL), with media containing RPMI 1640 with L-glutamine and sodium bicarbonate (from Sigma-Aldrich), supplemented with 10% Fetal Calf Serum (FCS; from PAN Biotech) and 1% Penicillin-Streptomycin (from Sigma-Aldrich). For conditioned media used for all Ramos cell lines, I prepared with equal fractions of fresh sterile media and culture media after harvesting cultures within the optimal range of growth. For the last component, I centrifuged the cultures (500xg for 5min at Room Temperature) and I filtered the supernatant through a sterile 0.22um filter (Stericup and Steritop, Millipore). I incubated all Ramos cells in a pre-humidified incubator with 37^{o}C and 5% CO_2 for mammalian cell cultures.

Human Embryonic Kidney (HEK293T) cell lines

Wildtype HEK293T cell line (ATCC CRL-3216) was purchased from ATCC. A version of HEK293T cells expressing a cassette of mCherry-2A-eGFP[W58X] under a CMV promoter, hereafter named as HEK293T-W58X cell line, which was originally generated by Montiel-Gonzalez et al., 2013, and it was a kind gift by Dr. Joshua Rosenthal, Marine Biology Laboratory, The University of Chicago, IL, USA. This cell line has a premature UAG stop codon in the codon 58 of the eGFP gene and upon A-to-I(G) RNA editing (U**A**G>U**G**G) reconstitutes eGFP expression and, thus, eGFP fluorescence. I maintained HEK cells in Dulbecco's Modified Eagle Medium (DMEM; Sigma-Aldrich) with 4500mg/L glucose, supplemented with 5% Fetal Calf Serum (FCS; PAN Biotech) and 1% Penicillin-Streptomycin

(Sigma-Aldrich). I prepared conditioned media for HEK cells with the appropriate media in equal portions of fresh and culture media, which I prepared in the same way explained above for the Ramos cell lines.

All cell lines I employed for the experiments presented in this dissertation were checked for contamination at Multiplexion GmbH, Heidelberg, Germany and were found free of Mycoplasma and viruses (Squirrel Monkey Retrovirus and Epstein-Barr virus). The biosafety level of all cell lines employed was classified as S1.

3.3.2 Plasmids and gRNAs generation

The plasmids I employed in the experiments presented in this dissertation are listed in Appendix C, along with their specific characteristics. I generated plasmids expressing gRNAs for site-directed mRNA editing using the same vector (listed in Appendix C as gRNA-pl) and the gRNA sequences along with their targets and characteristics are summarized in Appendix B. In brief, gRNA expression was driven by the polymerase III promoter in the U6 RNAi Entry Vector (Invitrogen), which was a kind gift by Dr. Joshua Rosenthal. I generated plasmids expressing different gRNAs with the NEBuilder HiFi DNA Assembly Cloning Kit (New England Biolabs), incorporating the aforementioned vector in linearized dsDNA format and a 120bp ssDNA oligo coding for the gRNA and having complementary ends to the ends of the linearized vector in the appropriate orientation. I generated the linearized vector with Q5 PCR (see 3.3.4; primers *pENTR-Fw*, *pENTR-Rv*) and I purified it through agarose gel extraction (see 3.3.4).

I generated stocks for all plasmids by transforming 50uL of competent DH5 *E. coli* cells, previously prepared as in Inoue et al., 1990, with 10ng of the respective plasmid. Prior mixing the competent cells with the plasmid DNA, I thawed the competent cells on ice for 20min, and upon adding the plasmid DNA, I incubated the mix on ice for 25min. I heat-shocked the mix at 42°C for 45sec and incubated it on ice for 2min. I then added 1mL of LB medium to the mix, which I further incubated for 45min at 37°C in shake (500rpm) in an Eppendorf shaker. LB broth was purchased from Sigma-Aldrich and dissolved in demineralized water VE, following manufacturer's instructions. Upon incubation, I centrifuged the culture at 11 000xg for 1min and aspirated the supernatant. I resuspended the sediment in 20uL of LB medium, which I evenly spread on LB-agar plates with the appropriate antibiotic resistance (Appendix C) and incubated overnight (16h) at 37°C. I inoculated single colonies picked by the agar plates in 200mL of LB media with the appropriate antibiotic resistance, which I incubated overnight at 37°C in shake (500rpm). I isolated the relevant plasmid from the cultures using the HiPure Plasmid Maxiprep kit (Invitrogen), following manufacturer's instructions. I measured plasmid concentrations using nanodrop (ThermoFisher Scientific) and adjusted stock concentrations for all plasmids to 1ug/uL of the eluted product, which was stored at -20°C. Last, I validated the plasmids with Sanger Sequencing, which was performed at Microsynth Seqlab GmbH, in Goettingen, Germany.

Alternative to vector-based gRNA expression, for a few experiments (see 4.2.4) I transfected gRNAs directly upon *in vitro* transcription (IVT). For gRNA IVT, I employed the method previously published by Kellner et al., 2019. For each gRNA, I incorporated a pair of two ssDNA oligos: one positive-sense ssDNA oligo with the sequence T7-3G (5'- GAAATTAATACGACTCACTATA GGG-3') and an antisense ssDNA oligo containing the complementary sequence of T7-3G, followed by the antisense gRNA sequence. The principle behind this method is that T7 polymerase recognizes the T7-3G sequence and synthesizes the gRNA using its antisense sequence as a template. Upon an annealing reaction of the aforementioned oligos with Standard Taq buffer (New England Biolabs), I incubated the reaction at 95^{o}C for 5min and which I slowly cooled down at a PCR-thermocycler (Bio-Rad) to 4^{o}C with 0.1oC/s. I performed IVT reactions using the HiScribe T7 Quick High Yield RNA Synthesis Kit (New England Biolabs) following the manufacturer's instructions. I incubated The IVT reactions at 37^{o}C for 4h at a PCR-thermocycler (Bio-Rad). I treated the products with DNAse I (New England Biolabs) for 15min at 37^{o}C to digest the template DNA oligos. I purified the products containing the gRNAs using the Monarch RNA Cleanup Kit (New England Biolabs). The average yield of this method was about 5ug/uL of gRNAs, which I estimated with Qubit fluorometric quantification (ThermoFisher Scientific). gRNAs produced with IVT were stored at -20^{o}C until further use. For transfection of Ramos cells (see 4.2.4) 5ug of each arRNA was used per transfection of 2x10^{6} cells.

3.3.3 Transfection methods

For transfection of Ramos cell lines (see 3.2.1), which grow in suspension, I employed the Amaxa Cell Line Nucleofector Kit V (Lonza) and the protocols followed relied on the manufacturer's instructions: the day before transfection, I split Ramos cell cultures at 1:5 ratio towards being in exponential growth phase on the day of transfection. I counted 2x10^{6} viable cells per sample with a Neubauer hemocytometer after staining with trypan blue (Sigma-Aldrich). I centrifuged the appropriate culture volume at 90xg for 10min at room temperature and I resuspended the pelleted cells in 100uL of nucleofector solution V per sample. In a certified cuvette by the manufacturer for the Nucleofector 2b apparatus (Lonza), I mixed the amount of plasmid DNA (typically 2ug) with 100uL of the cell suspension (in solution V) for one sample. I inserted the cuvettes in the Nucleofector 2b for electroporation with the program O-006, appropriate for the Ramos cell lines according to the manufacturer. Upon electroporation, I added immediately 500uL of warm culture media (RPMI, 10%FCS, 1%P/S) to the sample. I transferred the samples to wells of a 12-well plate, containing 1mL of pre-incubated culture media. I incubated the transfected cells for 24h in humidified 37°C - 5% CO_2 incubator until further handling. I found that the transfection efficiency of Ramos cells for plasmid DNA was about 20% following the aforementioned procedure, which agrees with the manufacturer's guidelines (Lonza).

I transfected HEK293T adherent cells (see 3.3.1) with lipofectamine 2000 (ThermoFisher Scientific) following the manufacturer's guidelines. For most transfections with lipofectamine 2000 performed for the experiments presented in this thesis, I followed protocols for 6-well plate layouts. A day prior to transfection, I detached the cells with trypsin (Sigma-Aldrich) after washing with PBS (Dulbecco's Phosphate Buffered Saline, Sigma-Aldrich). I seeded 7x10^5 viable cells per well in 2mL culture media (DMEM, 10%FCS, 1%P/S) to achieve confluency between 80-90% on the day of transfection. On transfection day, I changed the media to 700uL OptiMEM (Sigma-Aldrich) per well. I diluted the total amount of plasmid DNA (typically between 2-4ug) in 150uL of OptiMEM per sample. I mixed 10uL of Lipofectamine 2000 with 140uL OptiMEM per sample and I incubated it for 5min at room temperature. I mixed the lipofectamine-containing solution (150uL) with the diluted plasmid DNA (150uL), which I further incubated for 20min. The final solution containing complexes of plasmid DNA and lipids was gently and evenly added to the appropriate samples. I incubated the transfected samples in humidified 37°C - 5% CO_2 incubator for 6h and then I replaced the transfection media with regular culture media. For transfections in 24-well plates or 10cm dishes, the number of cells, plasmid DNA mass and reagent volumes were down- or up-scaled respectively, following the manufacturer's instructions.

3.3.4 RNA/DNA extraction and amplicon generation

I pelleted down harvested Ramos or HEK293T cells (typically <5x10^6) with centrifugation at 500xg for 4min at room temperature. I aspirated growth media and I washed the cells by resuspension of the pellet in 1mL of PBS. Cells were recentrifuged at the same speed and time. I aspirated the supernatant and I lysed the cell pellets with 350uL of buffer RLT Plus enriched with β-mercaptoethanol (10uL per mL of RLT Plus) from the AllPrep DNA/RNA Mini Kit (Qiagen). Following the kit instructions, I simultaneously isolated total RNA and genomic DNA (gDNA) from the harvested cells. I treated the extracted RNA with TURBO DNAse (ThermoFisher Scientific) to eliminate cross-contamination of gDNA from the procedure. I generated amplicons from gDNA with the Q5 High-Fidelity DNA polymerase (New England Biolabs), and cDNA amplicons from RNA were with the OneStep RT-PCR kit (Qiagen). All the PCR primers for amplicons generation are summarized in the Appendix A. I purified the PCR products with the NucleoSpin Gel and PCR Clean-up kit (Macherey-Nagel). Purified PCR amplicons were sequenced at Microsynth Seqlab GmbH, in Goettingen, Germany. Last, I quantified base editing across the Sanger sequencing chromatograms using the tool MultiEditR (Kluesner et al., 2021).

Deep-amplicon Next-Generation Sequencing (NGS) was performed at Eurofins Genomics GmbH, NextGen Sequencing lab, Konstanz, Germany under the "NGSelect Amplicon 2nd PCR" service. I generated amplicons in the lab with the respective PCR primers for each product (Appendix A) attached to overhangs of Illumina MiSeq adapters, indicated by the manufacturer. The coverage was about 60K read pairs per amplicon in 2x300bp read mode. I processed the amplicon NGS data, for QC, adapter trimming and were aligned with STAR (see subchapter 3.1) against the V_H germline sequence (IGHV4-34*01, IMGT). I generated the reference genome according to Dobin et al., 2013, with specifications for small genomes. I screened the aligned data for site-specific base change with REDItools2 (Table 3.1, subchapter 3.1) and with command-line BLAST (*BLAST® Command Line Applications User Manual*, 2008) to count mutations and gaps per read pair.

I used a qPCR method to measure the expression of ADAR1 and ADAR2 in all Ramos cell lines (see 3.3.1 and 4.2.3). I employed DNAse-treated RNA from two biological replicates per condition (Ramos WT, AID-/- and ADAR1-/-) and performed cDNA synthesis using the ProtoScript - First Strand cDNA synthesis kit (New England Biolabs), following the manufacturer's recommendations: In brief, I used as an input 300ng of DNAse-treated RNA (measured with Qubit RNA BR Assay Kit, ThermoFisher Scientific) per sample and mixed it with the Random Primer Mix (60uM) to a final volume of 8uL. After 5min at 65^{o}C in a PCR thermocycler, I set up the reverse transcription reaction (M-MuLV), by adding on top 10uL of M-MuLV Reaction Mix and 2uL of M-MuLV Enzyme Mix, bringing the reaction to a final volume of 20uL per sample. As negative controls, I set a non-template control (no RNA) and a non-enzyme containing control (M-MuLV Enzyme Mix). I incubated the cDNA synthesis reactions for 5min at 25^{o}C and then at 42^{o}C for 1h in a PCR thermocycler. I inactivated the enzyme at 80^{o}C for 5min and proceeded to the qPCR. For the qPCR, I used the iTaq Universal SYBR Green Supermix from Bio-Rad, following the manufacturer's instructions for reactions with final volumes of 10uL with an input of 15ng per sample: 5uL of iTaq mix (2x), 0.5uL from each of the Forward and Reverse primers (10uM), 2uL nuclease-free H_2O and 2uL of the template cDNA (7.5ng/uL). The qPCR primers used for *ADAR1* were qADAR1-Fw and -Rv, for *ADAR2* were qADAR2-Fw and -Rv. As expression controls, I measured the expression of the genes *GAPDH* and *ACTB* (qGAPDH-Fw, -Rv and qActb-Fw, -Rv respectively. All primer sequences can be found in Appendix A. Non-template and non-enzyme controls were also incorporated in this step. The instrument I used was the CFX Connect Real-Time PCR System from Bio-Rad laboratories, for which I followed the manufacturer's recommendations (protocol: 1) 95oC for 3min, 2) 95oC for 5sec, 3) 60oC for 30sec, 4) Repeat steps 2-4 39X more, 5) Melt curve 75oC to 95oC, increment 0.2oC for 10sec). Expression data were processed with the relevant CFX Maestro Software (v. 4.0.2325.0418; Bio-Rad laboratories), for which *ADAR1* and *ADAR2* expression was normalized against both the *GAPDH* and *ACTB* expression.

3.3.5 Flow cytometry

IgM staining of Ramos WT and AID-/- cells

I harvested Ramos WT and AID-/- cells typically between 1-3x10^{6} cells per sample. I centrifuged the cells at 400xg for 5min at room temperature, aspirated the media and resuspended the cell pellet in 1mL PBS supplemented with 0.5% FCS per sample. I pelleted down the cells again and aspirated the supernatant, but leaving about 10uL for lightly dissolving the pellet by flicking the tube. Per sample, I diluted 1uL of APC-conjugated goat anti-human IgM antibody (Jackson Immunoresearch) in 200uL of PBS (with 0.5% FCS) and added to the cells. I incubated the cell-antibody mix on ice and in dark for at least 30min. I added 1mL of PBS (with 0.5% FCS) and pelleted down the cells with centrifugation. I aspirated the supernatant and washed the pellet once more with 1mL PBS (with 0.5% FCS). To observe cell viability, I further stained the cells with propidium iodine (PI). I resuspended the sample pellets in a solution containing 10uL PI (Invitrogen), 1uL of RNAse A (Invitrogen) and 200uL of PBS with 0.5% FCS. Upon PI addition, I incubated the samples for 5min at room temperature and kept them thereafter on ice until/during the analysis. Staining controls were one antibody and PI unstained sample, a PI-only stained sample and an antibody-only stained sample.

Sample preparation

I collected harvested cells that required IgM staining (Ramos cells) and prepared them as explained above. For samples with endogenous fluorescence that did not require additional staining, such as of the HEK mCherry-GFP cassette, I harvested pelleted cells (centrifuge 400xg for 5min at room temperature) and washed once with 1mL PBS supplemented with 0.5% FCS. For cells undergoing sorting I resuspended them in 500uL (per $2x10^6$ cells) of PBS with 2% FCS, while for FACS analysis in PBS with 0.5% FCS. All samples prior to analysis or sorting were filtered through Falcon 5 mL polystyrene test tubes, with cell strainer snap cap (purchased from Corning).

Instruments for FACS analysis and sorting

For FACS analysis I used the instruments FACSCalibur (BD) and Millipore Guava EasyCyte HT (ThermoFisher Scientific). In particular, I analyzed the samples stained with the anti-IgM staining protocol as discussed above, with FACSCalibur with the lasers: 640nm (FL4) to detect APC fluorescence and 488nm (FL2) to detect PI. I used Millipore Guava EasyCyte HT (ThermoFisher Scientific) for analyzing fluorescence with mCherry (561nm, Red-R detector) and eGFP (488nm, Green-B detector) from the HEK cell lines (see 3.3.1). For cell sorting I used the instrument FACSAria1 (BD) and in particular the lasers 561nm (Yellow-Green) for PI and mCherry, 640nm (Red) for APC, 488nm (Blue) for GFP. All instruments employed were provided, along with kind assistance in analysis and sorting, by the DKFZ Flow Cytometry Facility. For processing flow cytometry recorded data I used the software FlowJo v.10.6.2.

3.4 Methods Aim 3

3.4.1 SARS-CoV-2 sequence data

I overall retrieved 62 211 genome-wide SARS-CoV-2 sequences from the NCBI SARS-CoV-2 Resources portal (https://www.ncbi.nlm.nih.gov/sars-cov-2/; “NCBI SARS-CoV-2 Resources”), isolated from infected individuals (humans) in the USA from the first 15 months of the COVID-19 pandemic (collected between January 5th and March 31st, 2021).

3.4.2 Mutation calling and annotation

Mutations in the SARS-CoV-2 sequences (see subchapter 3.4.1) were reported by comparing the different sequences against the sequence of the first isolate from the original human infection in Wuhan, China (Accession number NC_045512, RefSeq; Wu et al., 2020), which I used as a reference genome. I aligned the different SARS-CoV-2 sequences using the software “VIRULIGN” (Libin et al., 2019), which reported mutations (also termed SNVs) in a codon-correct fashion. I further processed the different mutations called to evaluate the amino acid changes, employing R programming language (v. 4.0.2; R Core Team, 2020) and the package ‘Tidyverse’ (v. 1.3.0, Wickham et al., 2019), to therefore, note which mutations were missense or silent. I annotated all reported mutations according to the NCBI RefSeq SARS-CoV-2 genome annotation (NC_045512, RefSeq). I furthermore visually inspected alignments for validating most mutations called and their potential amino acid changes. As predominant mutations in aggregate throughout the dataset I considered those mutations which are present in at least 10% of the sequences (separately for 2020 and 2021). I further profiled the dataset for the different SARS-CoV-2 lineages reported thus far, in order to detect reported variants of concern (VOCs), using the “pangolin” tool (https://github.com/cov-lineages/pangolin), which follows the PANGO nomenclature (Rambaut et al., 2020). In order to detect low-frequency mutations in the Spike protein, which is key for the viral infectivity (see subchapter 1.2.4), or in the different VOCs, so as to evaluate their abundance shift in time, I considered mutations found in more than 0.1% of the sequences, so as to eliminate possible sequencing errors.

The students Mr. George Samaras, Ms. Alexandra Paulus, Ms. Gabrielle Whitehouse, Ms. Anna Jamison and Ms. Michelle Lee helped me by processing about 8 000 genomes altogether from 2020 to obtain mutation calls as described above, following the pipeline I established. The complete pipeline is presented and explained in the subchapters 3.4.2 - 3.4.4, as well as described in Tasakis et al., 2021. They furthermore contributed in the visual inspection, validation and curation of the mutation calls from the 8 000 genomes they processed.

3.4.3 Analysis of signatures of co-existing mutations

In order to evaluate whether SARS-CoV-2 mutations are gradually accumulating over time through host-dependent RNA editing activity (but also other co-existing mechanisms), I defined a set of all the different possible combinations of the predominant mutations. I used the combinations I found in the dataset, to infer a reference of putative mutational signatures overall. I focused on the signatures found in more than 0.1% of the viral isolates, for which I constructed time-scaled phylogenetic tree with the tool IQ-TREE 2 (Minh et al., 2020).

3.4.4 Downstream analysis and visualization

For all analyses, data processing, statistics and visualization I used the R programming language (v. 4.0.2; R Core Team, 2020), unless specified otherwise as in 3.4.1-3.4.3. I specifically used the package 'Tidyverse' (v. 1.3.0, Wickham et al., 2019) for data management and visualization, and the packages, 'msa' (Bodenhofer et al., 2015), 'treeio' (Wang et al., 2020) and 'ggtree' (Yu et al., 2017) for further analysis and visualization of the time-scaled phylogenetic tree.

4. Results and Discussion

4.1 Aim 1: ADAR1-mediated RNA editing correlates with acquisition of specific DNA mutations during Multiple Myeloma progression.

4.1.1 Preface

Multiple Myeloma (MM) is a cancer of antibody-secreting plasma cells amassing in the bone marrow, which may lead to marrow failure and/or bone destruction (Anderson et al., 2009). It is estimated that in 2021 in the United States alone, there will overall be about 34 920 newly diagnosed MM cases and 12 410 new deaths related to the disease (Siegel et al., 2021). The clinical image of MM is broadly heterogeneous, which has a great impact on the treatment of the disease; patients may present the relevant symptoms to different extends, while some may remain asymptomatic for extended periods of time prior to diagnosis (Alexanian and Dimopoulos, 1994). Apart from the phenotypic diversity characterizing the disease, patient-derived MM tumors are highly heterogeneous considering their genomic architecture, which comprises of chromosomal translocations, associated aberrant class-switch recombination, hyperploidy events, accompanied by elevated mutational load in key genes (Chng et al., 2007). Although there is an abundance of genetic events described in MM, there is no clear consensus of specific genetic drivers across MM patients. But what is rather specific, is that the vast majority of the aforementioned genetic events described in MM are aggregating in genomic loci critical for plasma-cell fate and longevity, as reviewed by Morgan et al., 2012.

A key characteristic of MM is that mutations or other genetic events in myeloma plasma cells are generating a diverse set of clones, which are being selected and they further evolve (Fakhri and Vij, 2016; Lagana et al., 2017; Corre et al., 2018). The most common chromosomal translocation in MM is the gain of additional copies of the chromosomal fraction 1q21, which is found in about 40% of newly diagnosed MM patients and is associated with poor disease outcomes (Nemec et al., 2010). Interestingly, 1q21 is also the genomic location of the *ADAR1* gene, and 1q21 gain is one way that MM tumors overexpress ADAR1 leading to elevated A-to-I RNA editing hyper-activity, which is associated with poor prognosis (Lazzari et al., 2017; Teoh et al., 2018). Previous analyses performed by Dr. Alessandro Laganà in a cohort of 590 MM patients (MMRF; see 3.2.1) validated that 1q21 gain is indeed associated with over-expression of ADAR1, elevated and aberrant RNA editing activity and poor prognosis in the aforementioned cohort of patients. Furthermore, Dr. Laganà found that MM patients without 1q21gain, may also over-express ADAR1 through interferon (IFN) induction, naturally accompanied by elevated RNA editing activity, which in the overall cohort appears to be uniquely associated with poor survival, regardless of whether the patients had the 1q21 gain or not. The overall RNA editing activity measured by the Alu Editing Index (AEI; Roth et al., 2019) correlated with the expression of ADAR1 and not ADAR2 (minimally expressed) in the cohort. The aforementioned findings by Dr. Laganà are published in our joint preprint (Tasakis et al., 2020). Relying on Dr. Laganà's previous findings, I considered that Multiple Myeloma tumors is the ideal scenario to test whether the aberrant A-to-I RNA editing activity by ADAR1 can correlate with acquisition of DNA mutations specific to ADAR1 activity. Therefore, from the aforementioned cohort of 590 patients (MMRF), I focused on a subset of 23 MM patients, who had matched RNA-seq and WES sequencing data from two successive timepoints of the disease. This cohort allowed me to address my hypothesis of whether ADAR1-dependent RNA editing in the earlier patient timepoints would lead to acquisition of DNA mutations by the same enzyme in the later timepoint upon selection.

4.1.2 ADAR1 as an RNA editor and likely a DNA mutator in Multiple Myeloma

As introduced in the chapter 1.2.1, ADARs are primarily known as RNA editing enzymes (also referred to as RNA editors), which deaminate Adenosines (A) to Inosines (I; recognized as Guanosines, G) within double-stranded RNA (dsRNA) instances (Nishikura, 2010). Recent *in vitro* data, notwithstanding, showed that ADARs can also deaminate As in the DNA when they are within double-stranded instances of DNA/RNA hybrids, which are also known as R-loops (Zheng et al., 2017). R-loops are abundantly present in mammalian genomes and can occur co-transcriptionally, when the nascent RNA chain exits the RNA polymerase and hybrids with the template negative-sense DNA strand, leaving the coding (positive-sense) strand unpaired (Sanz et al., 2016). Considering that since ADAR-mediated RNA editing can be co-transcriptional (Laurencikiene et al., 2006), I hypothesized (visualized in Figure 4.1) that ADARs may lose their touch with the dsRNA target and opportunistically access R-loops formed *in situ* to edit the RNA strand of the hybrid, which would be their primary job, but they may also deaminate As in the DNA strand of the R-loop (the template/negative-sense DNA strand of the locus). The last, should they not be corrected, they may lead to A-to-G mutations in the negative-sense strand, readable as T-to-C mutations from the positive-sense strand. Alternatively, should repair mechanisms have taken place, they may give rise to other mutations deriving from Ts.

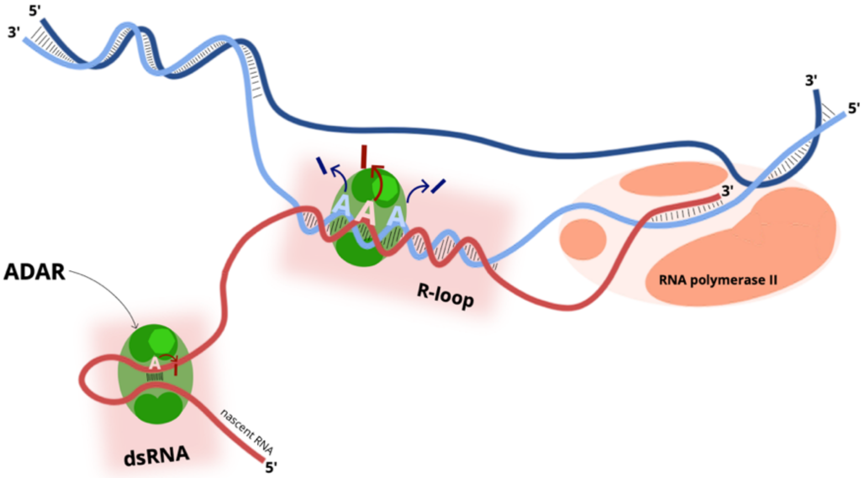


Figure 4.1 Hypothesis model in which an ADAR is an RNA editor of a certain transcript and a DNA mutator of its cognate gene. ADARs edit double-stranded RNA (dsRNA) co-transcriptionally. Hybrids of the nascent RNA (red strand) and the template DNA strand (light blue) may hybrid, forming R-loops within the transcription bubble, which I hypothesize that ADARs may access toward targeting the RNA but may also mutate DNA in the vicinity. Adapted figure from Tasakis et al., 2020. It is reused under the Creative Commons License 4.0. This illustration was created by myself.

Relying on the aforementioned proposed model (Figure 4.1), I tested my hypothesis by employing data from 23 MM patients, available from the MMRF study (see 3.2.1), who had matched RNA-seq and WES data available from two timepoints of the disease; tumors at presentation or diagnosis (Timepoint I or TP1) and tumors at relapse (Timepoint II or TP2). As shown in Figure 4.2A, I called A-to-I(G) RNA editing sites from TP1 and matched them to T-derived mutations (T-to-*) uniquely identified in TP2. I focused on the unique mutations in TP2, because these must be the ones that have been selected as advantageous for the relapsing tumors (TP2), considering the rules applying for the clonal evolution in MM tumors (Lagana et al., 2017), as introduced in the subchapter 4.1.1. I furthermore required the unique T-derived mutations (TP2) to be within 20bp distance up- or down-stream the RNA editing event, relying on the previous findings of Zheng et al., 2017 showing that the footprint of an ADAR on a double-stranded nucleic acid structure is about 20bp. Following this analysis pipeline (Figure 4.2A), I obtained a list of editing-to-mutation matches summarized in the heatmap shown in Figure 4.2B. In this heatmap, the editing-to-mutation matches are tallied by the genes they are mapped into (columns) and per patient (rows), while the candidates shown are the ones shared by at least 25% of the patients. I additionally profiled the editing-to-mutation matches with regards to their gene topologies and I found that the vast majority of the mutation candidates were mapped in non-coding regions such as introns or 3'UTRs, which is compatible with the transcript topology pattern that ADAR1 typically edits RNA within *Alu* repeats (Athanasiadis et al., 2004; Chung et al., 2018).

Within the top mutated candidates (Figure 4.2B), I found genes, the transcripts of which were previously suggested as targets of ADAR1-mediated editing in biological systems other than MM. The top candidate was *EIF2AK2* encoding for the Protein Kinase R (PKR), the transcript of which was previously shown as edited in B cells (Wang et al., 2013), and which I also validated in MM cell lines in collaboration with Dr. Violetta Leshchenko and Ms. Pavithra Nedumaran (see 4.1.3). Another top candidate was *MDM4*, which is known as a p53 inhibitor (Danovi et al., 2004), the transcript of which was also reported as edited (Hong et al., 2018). Additional candidates from the top quartile I found mutated, such as *LRRC28*, *ADAM19*, *COPE*, *EDARADD* had transcripts predicted as edited in the database REDIportal (Picardi et al., 2017). Pathway enrichment analysis (Figure 4.2D), which I performed with the pathway-level analysis tool SLAPenrich (Iorio et al., 2018), revealed that the correlative-to-editing mutation candidates affect pathways, such as p53, JAK-STAT, hematopoietic, proteasome signaling among others, previously suggested to be crucial for MM (Dehghanifard et al., 2018). It is overall encouraging to observe, that the majority of the top-candidates from my analysis have been previously suggested to be targets of ADAR1 on the transcript level, which validates the accuracy of the first necessary part of my analysis with regards to calling A-to-I RNA editing sites from TP1. It should be noted, that the RNA editing analysis I performed from the TP1 has revealed previously unknown targets of ADAR-mediated editing specifically in MM, notably those of *EIF2AK2* and *MDM4*. My correlative analysis that their cognate genes may be mutated on a later stage in the disease (TP2), provides a mechanistic insight of how mutations can be acquired during the course of the disease.

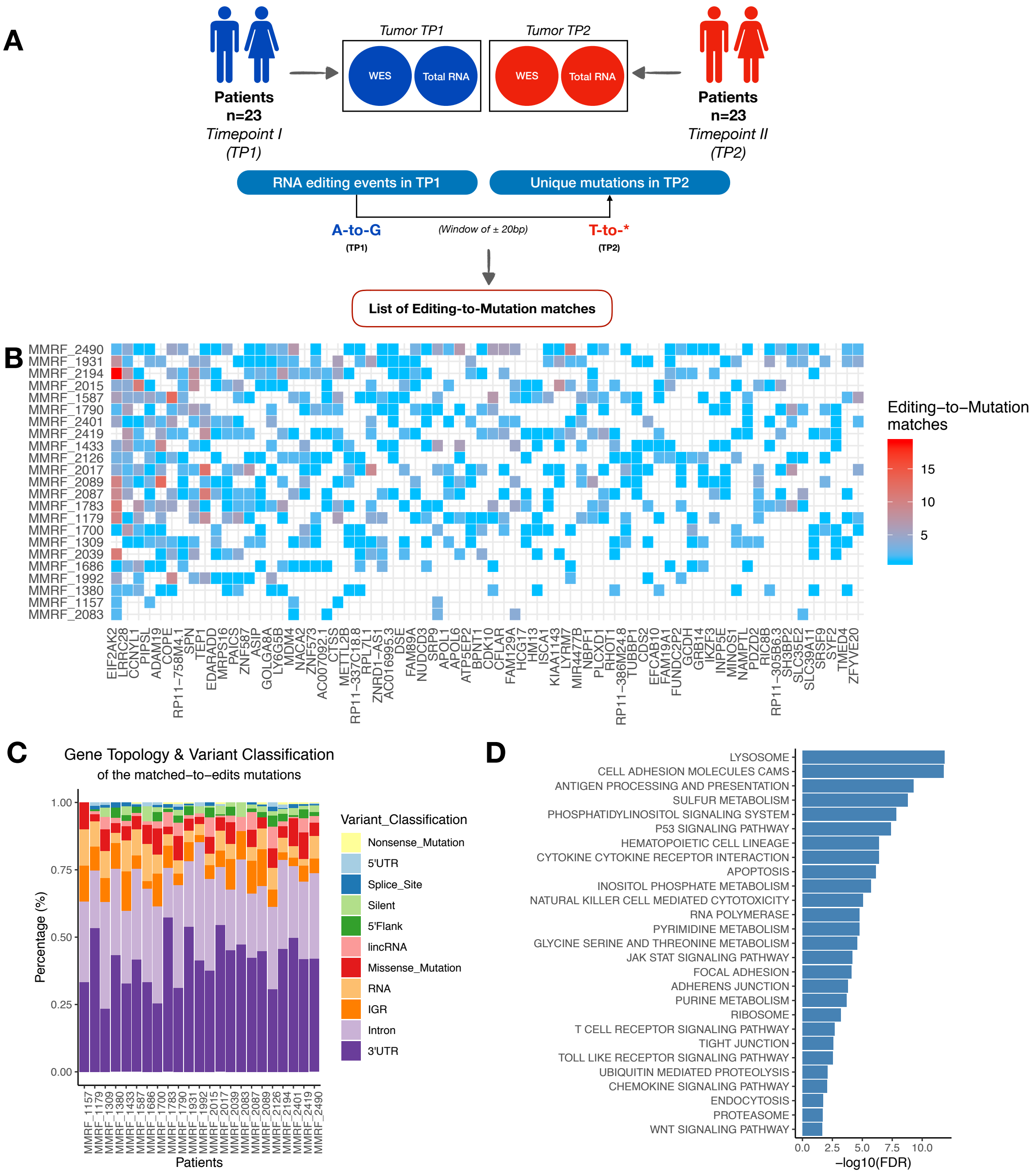


Figure 4.2 Correlation of ADAR1-dependent RNA editing and DNA mutation in Multiple Myeloma. (A) Schematic representation of the ADAR1-dependent editing-to-mutation analysis within the 23-patient cohort (see also 3.2.2). A-to-I(G) RNA editing sites were called through RNA/DNA(WES) comparisons in Timepoint I (TP1) and were matched to unique T-derived (T-to-*) mutations called from Timepoint II (TP2) within windows of ±20bp from the editing events. This analysis produced a list of "editing-to-mutation matches" correlating ADAR1-dependent RNA editing sites with mutation candidates per patient. (**B**) Heatmap summarizing the counts of editing-to-mutation matches per gene (columns) and per patient (rows). Only genes-candidates found in 25% or more the patients are shown. The top candidate was *EIF2AK2*, encodes for Protein Kinase R (PKR), among others in the top quartile previously suggested to be targets of ADAR1 on the RNA level, such as *MDM4* or *ADAM19*. (**C**) The mutation candidates were profiled for their gene topology and variant classification per patient. Most of the mutations were found in 3'UTRs or introns, where ADAR1 typically edits within *Alu* repeats. A non-negligible amount was also found within the CDS. (**D**) Significantly enriched pathways affected by the mutation candidates. All panels in this figure are from Tasakis et al., 2020 and are reused under the Creative Commons License 4.0. All data and illustrations were produced by myself.

4.1.3 *EIF2AK2* transcript is a target of ADAR1 in Multiple Myeloma cell lines.

Up to this point, I have presented correlative data that A-to-I RNA editing, most likely by ADAR1 (see 4.1.1), may lead to DNA mutation in a cohort of 23 MM patients. Here, I present experimental data supporting that the transcript of *EIF2AK2*, encoding for PKR, is indeed a target of ADAR1 in the context of MM. To address that, I employed two typical MM cell lines: KMS-11 and KMS-20. KMS-11 is a cell line representing MM in a later disease stage of B-cell differentiation (Namba et al., 1989). Upon isolating total RNA and gDNA from bulk cultures of KMS-11 and KMS-20 cells, I generated cDNA and gDNA amplicons for the 3'UTR of the *EIF2AK2* transcript and gene respectively, amplifying regions found as edited from my original *in silico* analyses from the patient data (see 4.1.2), following the methods described in detail in 3.2.3. Sanger sequencing chromatograms from the aforementioned cDNA amplicons revealed A-to-G double peaks, absent from the gDNA amplicons of the respective locus in both KMS-20 and KMS-11 cell lines (highlighted sites in Figure 4.3A). The last sites, were consistent with sites found as edited on the RNA from the human patient data, but I overall observed more sites dispersedly edited in the cDNA amplicons from the KMS-11 cell line.

As introduced in 1.2.1, ADAR1 has two isoforms: ADAR1-p110, which is nuclear and constitutively expressed, and ADAR1-p150, which shuttles between the nucleus and the cytoplasm and it is expressed through an interferon (IFN) inducible promoter (Lamers et al., 2019). Therefore, to address from a functional perspective that ADAR1 is editing the transcript of *EIF2AK2*, I challenged KMS-11 cells with 10U and 100U of IFNα with the help of Ms. Pavithra Nedumaran (see 3.2.3), and generated cDNA and gDNA amplicons 96h post-treatment. I found that RNA editing levels were increased in two dimensions. First, the per-site RNA editing increased gradually with IFNα treatment. For instance, as shown in the Figure 4.3B, the site chr2:37,327,859 was 68% edited on the transcript level in untreated cells, but reached nearly complete editing (~97%) when treated with 100U IFNα for 96h. Second, more sites were found significantly edited throughout the transcript, and again they gradually increased when IFNα doses were increased, as shown in Figure 4.3C. With this functional assay, I validate that *EIF2AK2* is indeed a target of ADAR1 on the RNA. Provided that my hypothesis is true, as envisioned in Figure 4.1, ADAR1 may also function as a DNA mutator of the cognate genes of its target transcripts, such as *EIF2AK2* which appears to be the top candidate of my *in-silico* analysis in the patient data of MM progression. To address that, I will first further explore the correlative *in silico* data (subchapter 4.1.4) and I will further present proof of concept experiments (subchapter 4.2), for which I target ADAR1 to a certain transcript and I look for subsequent acquisition of DNA mutation in its genomic locus.

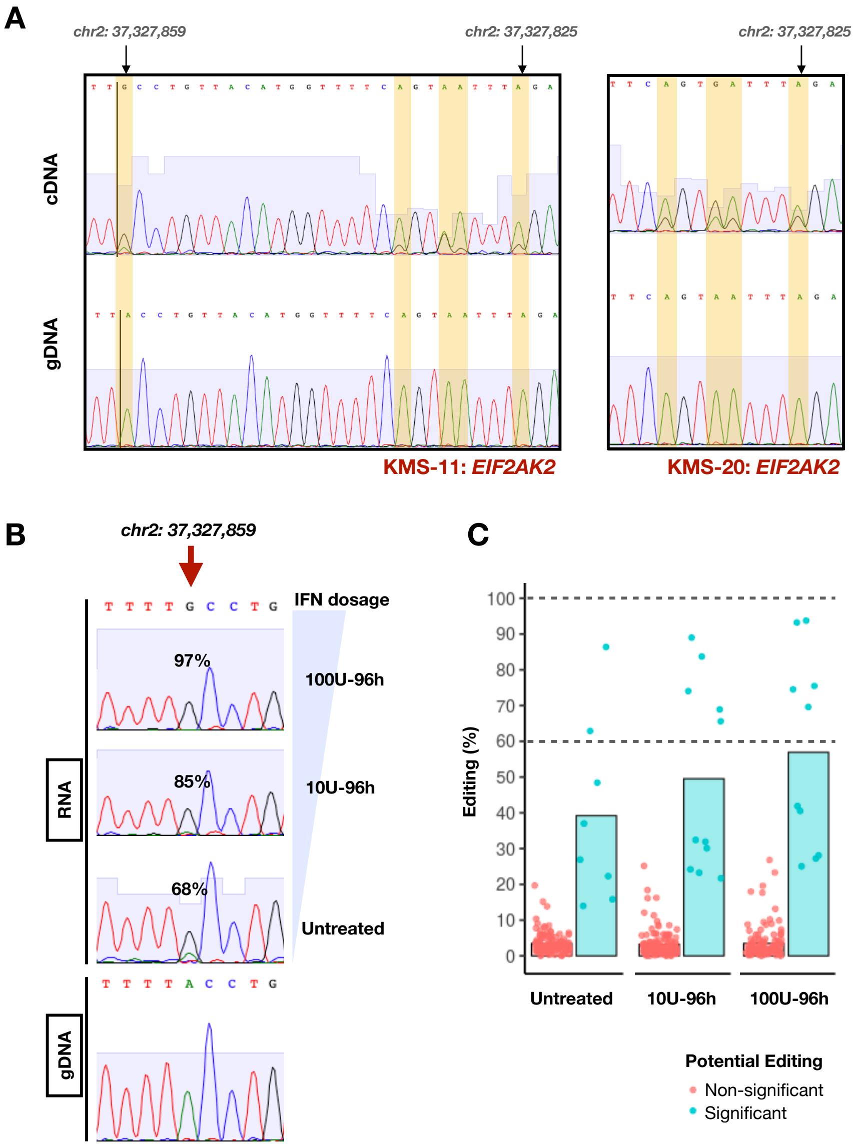


Figure 4.3 Experimental validation of ADAR1-mediated RNA editing in the *EIF2AK2* transcript. (**A**) Sanger sequencing from cDNA and gDNA amplicons of the 3'UTR of *EIF2AK2* (*Eif2ak2-Fw* and *Eif2ak2-Rv* primers; Appendix A) were generated from total RNA and genomic DNA (gDNA) respectively, which were simultaneously extracted from cultures of KMS-11 and KMS-20 cells. A-to-G double-peaks were detected in the cDNA amplicons (highlighted in yellow along with a range of genomic coordinates according to hg19), while absent from the respective positions of the gDNA amplicons, indicating A-to-I editing events. Overall, more A-to-G double-peaks were detected in the cDNA amplicons of KMS11 cells. (**B**) KMS-11 cells were challenged with 10U and 100U dosages of IFNα (IFN) for 96h and A-to-I editing, detected from cDNA (labelled as RNA) amplicons, in the position chr2:37,327,859 gradually increased from 68% (Untreated cells) to 85% in 10U-96h treated cells and then to 97% in 100U-96h treated cells. (**C**) In the samples treated with IFN (10U-96h and 100U-96h) more sites were found significantly edited (light-blue dots), while the levels of A-to-I editing (%) gradually increased as also shown in panel B. The bars indicate the mean value of the editing % across sites per sample. Editing quantification and relevant statistics were performed with the tool MultiEditR (Kluesner et al., 2021) following the default parameters. All panels in this figure are from Tasakis et al., 2020 and are reused under the Creative Commons License 4.0. All data and illustrations in this figure were produced by myself. Ms. Pavithra Nedumaran helped me with the IFN treatment of KMS-11 cells. Dr. Violetta Leshchenko maintained the cell cultures prior to IFN treatment.

4.1.4 ADAR1-dependent RNA editing and DNA mutation may jointly facilitate Multiple Myeloma progression.

I previously presented correlative data between ADAR1-dependent RNA editing and acquisition of specific DNA mutations during MM progression in a cohort of 23 patients (see 4.1.2). I furthermore showed that the top candidate from this analysis, *EIF2AK2* encoding for PKR, is indeed a target of ADAR1 in MM cell lines (see 4.1.3). Here, I employ the patient data from my aforementioned analysis to further explore the possibility that the acquisition of specific T-derived mutations is due to the dual role of ADAR1 as an RNA editor and also a DNA mutator in the context of MM. First, I focused on four candidates from the top quartile of the aforementioned analysis (Figure 4.2B), *EIF2AK2*, *MDM4*, *ADAM19* and *LRRC28*, and asked where the acquired T-derived mutations (in TP2) of interest would "localize" with regards to the original RNA editing events in TP1 per patient. As demonstrated in Figure 4.4A, I defined windows of 41bp centralized by an RNA editing site, in which I called the mutation candidates ±20bp up- or down-stream an RNA editing event. The newly acquired mutations uniquely found in TP2 and not in TP1, are shown as "lollipops" to their respective genomic coordinates in the same figure. The T-derived (or A-derived, depending gene-orientation), highlighted in red are mostly found within the editing-defined 41bp windows or near them, while the rest of the mutations (highlighted as grey) are randomly distributed. It should be noted at this point, that the majority of the newly acquired mutations in TP2 were of variation frequency values of between 1-5%, which is why the number of visualized mutations in Figure 4.4A is relatively high in some patients (for example MMRF_2194 for *EIF2AK2*). To purge potential sequencing errors, I cross-validated that these mutations were present in both WES and RNA-seq calls from the same timepoint (TP2).

According to my main hypothesis (model demonstrated in Figure 4.1), DNA mutations in genes by ADAR1 may not be the molecular purpose of this deaminase, but rather an off-target effect of their aberrant and hyper-editing activity which may present in tumors. As a first step toward further exploring this hypothesis, I aggregated the number of editing-to-mutation counts from my original analysis per patient and correlated this with their Alu Editing Index (AEI), the measure of their RNA editing activity. As shown in Figure 4.4B, the count of editing-to-mutation matches significantly correlates with the AEI in TP1 and not in TP2, which is consistent with my hypothesis that mutations may be introduced as off-target effects. However, the weak correlation of editing-to-mutation with the overall RNA editing activity in TP2, drove me to question whether this may also be due to changes of RNA editing activity from TP1 to TP2 per patient. Indeed, the 23-patient cohort I investigated is practically grouped in two: 14 patients significantly decrease their RNA editing activity from TP1 to TP2, while 9 patients show an increase (Figure 4.4C). This observation encouraged me to test whether there are discrepancies between the two groups with regards to the acquisition of T-derived mutation. Indeed, as demonstrated in Figure 4.4D, I found that the patients who decrease their editing activity, present a significant enrichment of T-derived mutations acquired in their ±20bp editing-defined windows, while the patients increasing their editing activity do not. When I extended the editing-defined windows to ±100bp the significance trend remained, in favor of the patients who decrease their editing activity. This may indicate that once DNA mutation is fixed there may be no necessity for RNA editing anymore, which can be mechanistically explained by the fact that the preferred motifs of ADAR1 on the RNA are altered due to mutation.

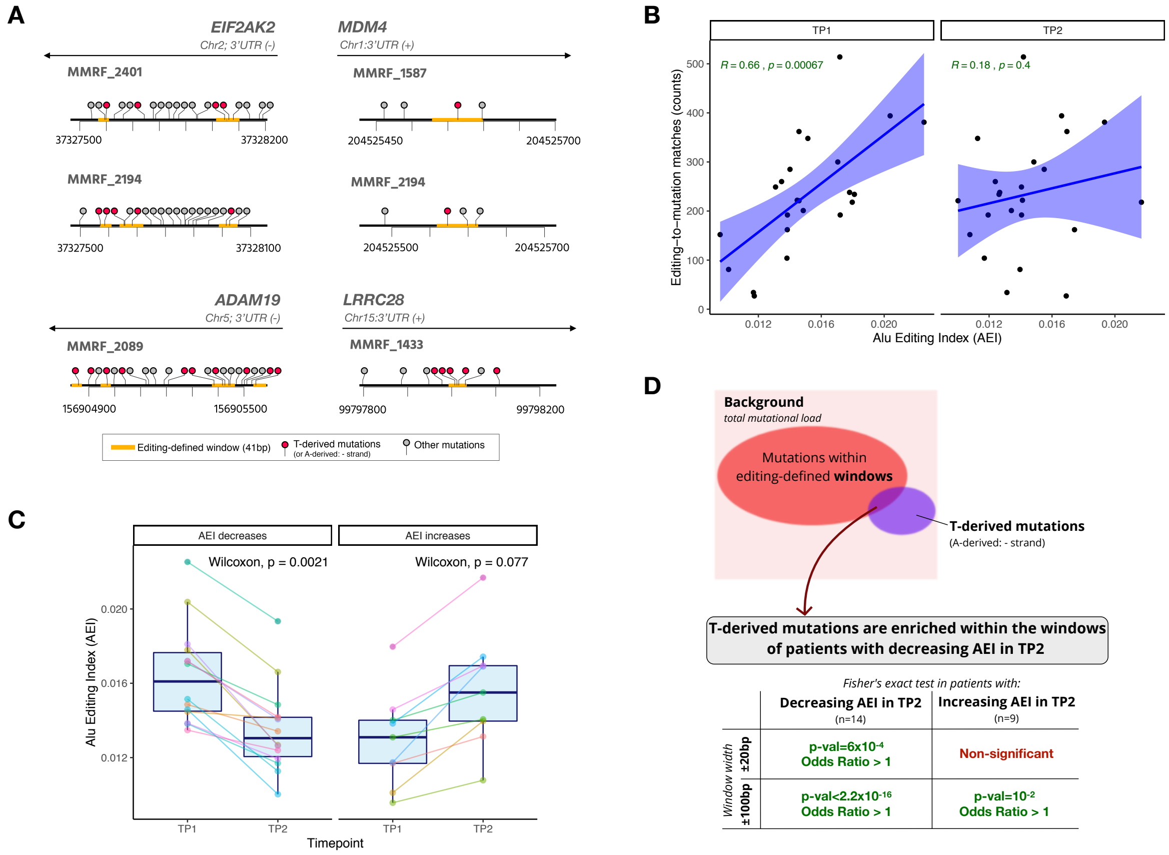


Figure 4.4 Multiple Myeloma progresses either through fixation of ADAR1-dependent DNA mutations or elevated RNA editing activity. (**A**) *EIF2AK2, ADAM19, MDM4, LRRC28* were within the top quartile of candidates for ADAR1-dependent DNA mutation (see Figure 4.2B). This schematic representation shows tracks of particular genomic regions for a few patients as an example, on which the yellow parts are the editing-defined windows from TP1. Editing-defined windows may be overlapping due to multiple RNA editing events in close proximity within a region. On the tracks, "lollipops" show the unique mutations found in TP2. Highlighted in red are all the T-derived mutations for the positively oriented genes (+) in the given regions and the A-derived for the negatively oriented genes (-). Most of the newly acquired T- or A-derived mutations, depending on gene orientation, in TP2 are found within or near the editing-defined windows. (**B**) The abundance of editing-to-mutation matches correlates with the RNA editing activity (measured with the Alu Editing Index; AEI) significantly in TP1 (p=6.7x10^{-4}, R=0.66), but not in TP2 (p=0.4, R=0.18) in the 23-patient cohort. R stands for the Pearson's correlation and p for the p-value. (**C**) The change of RNA editing activity (AEI) between TP1 and TP2 groups the patients in those who significantly decrease their AEI (n=14; Wilcoxon test, p=2.1x10^{-3}) and those who increase or keep their AEI to similar levels (n=9; Wilcoxon test, p=0.077 non-significant). The different patients are color-coded and lines between TP1 and TP2 are connecting the AEI values of each patient. (**D**) Fisher's tests for enrichment of the T-derived mutations (or A-derived depending on gene orientation) versus others within the ±20 editing-defined windows are significantly enriched (p-val=6x10^{-4}, Odds Ratio >1) for the patients decreasing their AEI from TP1 to TP2 and not those who increase their AEI. When editing-defined windows are extended to ±100bp, the trend of enrichment remains in favor of those patients who decrease their AEI (p-val<2.2x10^{-16}, Odds Ratio >1). All panels in this figure are from Tasakis et al., 2020 and are reused under the Creative Commons License 4.0. All data and illustrations were produced by myself.

4.1.5 Discussion

Multiple Myeloma (MM) is a hematological malignancy, accounting for about 10% of blood cancers, which entails a diverse genetic architecture and clinical image, often complicating the decision making in the clinic for the proper treatment regime (Alexanian and Dimopoulos, 1994; Morgan et al., 2012). One of the most frequent chromosomal abnormalities, detected in about 40% of newly diagnosed cases in MM, is the gain of 1q21 copies, associated with poor survival and disease outcomes (Nemec et al., 2010). 1q21 is also the chromosomal locus of *ADAR1* in the human genome and previous studies showed that 1q21 gain is a mechanism through which ADAR1 (both p110 and p150 isoforms) is overexpressed in MM, leading to transcriptome-wide and aberrant A-to-I RNA editing associated with poor survival (Lazzari et al., 2017; Teoh et al., 2018). However, Dr. Laganà showed that ADAR1 may also be overexpressed in patients without the 1q21 gain through interferon induction, leading to similar disease phenotypes (data published in our joint preprint, Tasakis et al., 2020). This mechanism of 1q21-dependent or independent ADAR1 overexpression, which I summarize in the first part of the scheme of Figure 4.5, has also been proposed in breast cancer (Fumagalli et al., 2015). Therefore, the principles of ADAR1-dependent RNA editing in MM and its crucial impact on the disease progression have been well studied and understood. Here, I explored a different possibility for ADAR1, particularly about its ability of mutating DNA.

It was previously reported from *in vitro* data that ADARs can deaminate Adenosines in the DNA within DNA/RNA hybrids, instances also known as R-loops (Zheng et al., 2017). In mammalian genomes, R-loops are formed genome-wide, they are conserved and have an average size of ~80-300bp (Sanz et al., 2016; Chen et al., 2019; Stolz et al., 2019; Malig et al., 2020). Although R-loops were originally considered to be transcriptional byproducts, they are in fact shown to be involved in fundamental processes, such as Class-Switch Recombination, and they have furthermore been presented as a potential threat to genome stability (Aguilera and García-Muse, 2012). R-loops are resolved by RNAse H, which degrades the RNA strand of the DNA/RNA hybrid (Amon and Koshland, 2016). Interestingly, it was recently shown in cancer cell lines that ADAR1, and in particular the nuclear isoform p110, deaminates unpaired Adenosines (to Inosines, recognized as Guanosines) against Cytidines in either the DNA or RNA strand of telomeric R-loops, enhancing RNAse H to resolve the R-loop and prevent their accumulation which may lead to genomic instability (Shiromoto et al., 2021). R-loops, however, are formed genome-wide co-transcriptionally between the nascent RNA and the template strand (Sanz et al., 2016). RNA editing can also be co-transcriptional (Athanasiadis et al., 2004; Laurencikiene et al., 2006). I therefore hypothesized that an ADAR may lose touch with its dsRNA target and access an R-loop formed between the template DNA strand and the nascent RNA, presumably to edit RNA, but it may also mutate the DNA strand of the hybrid in the vicinity of the original editing event (Figure 4.1). The result of an ADAR deaminating an Adenine in the DNA, will lead to Hypoxanthine ("behaving" like a Guanine and pairing with Cytosine) which may or may not be corrected by repair mechanisms (Budke and Kuzminov, 2006; Pang et al., 2012). This would lead to an

A-to-G mutation in the template (negative-sense) DNA strand or broadly A-deriving mutations if repair mechanisms have taken place. A-to-G (or A-deriving) DNA mutations from the negative-sense strand are read computationally as T-to-C (or T-deriving) mutations from the positive-sense reference strand.

In Multiple Myeloma, both ADAR1 isoforms (p110 and p150) are overexpressed in 1q21-positive tumors, while in 1q21-negative tumors ADAR1-p150 is overexpressed through IFN induction, and ADAR2 is generally expressed in very low levels, as others found (Lazzari et al., 2017) and Dr. Laganà confirmed (Tasakis et al., 2020). As introduced in 1.2.1, ADAR1-p110 is constitutively expressed and is primarily nuclear, while the interferon-inducible ADAR1-p150 is mostly cytoplasmic, though it can present in the nucleus (Lamers et al., 2019). Fundamentally, one would expect that ADAR1-dependent DNA mutation would be due to ADAR1-p110 activity, as others have also indicated (Shiromoto et al., 2021). I, therefore, focused on a set of 23 MM patients from a cohort of originally 590 patients with validated ADAR1 overexpression and subsequent elevated RNA editing activity (Tasakis et al., 2020). I focused on the set of 23 patients because each patient has with matched RNA-seq and WES data from two successive timepoints of the disease: tumors at diagnosis (TP1) and tumors at relapse (TP2). I called RNA editing sites from TP1 and matched them to T-derived DNA mutations unique for TP2 (Figure 4.2A), allowing me to focus on a set of mutations that were positively selected. I required that the mutations fall within a window of ±20bp from the editing event, following the findings of Zheng et al., 2017 about the footprint of an ADAR on double-stranded nucleic acid moieties. I found a number of genes mutated (in TP2), whose transcripts I previously found highly edited (in TP1), summarized in Figure 4.2B. I found that the vast majority of the subsequent mutations, correlated to RNA editing, were in Introns or 3'UTR encoding regions (Figure 4.2C), where ADAR1 usually edits RNA within *Alu* repeats (Athanasiadis et al., 2004), in genes crucial for tumorigenesis (i.e. p53, Apoptosis) and of high relevance for MM (i.e. Proteasome, Hematopoietic cell lineage), as my pathway analysis showed (Figure 4.2D).

The top gene-candidate I found mutated as correlated to editing was *EIF2AK2*, which encodes for the Protein Kinase R (PKR). PKR is a crucial regulator of a number of cellular pathways, notably involved in infection, inflammation (through the NFkB pathway) and tumorigenesis (Gal-Ben-Ari et al., 2019), the last of which is not surprising considering the fact that PKR interacts with p53, playing a prominent role in its tumor-suppressor function (Yoon et al., 2009). To validate that EIF2AK2 is indeed targeted by ADAR1 on the RNA level, I employed two representative MM cell lines (KMS-11 and KMS-20) and, first, generated cDNA amplicons for the 3'UTR of *EIF2AK2* in which I detected A-to-I editing events absent from the gDNA (Figure 4.3A). Second, I challenged KMS-11 cells with increasing concentrations of IFNα and I found proportionally higher levels of RNA editing per position, but I also found more positions edited (Figure 4.3B-C). This is causal to the ADAR1 activity and in particular to the interferon-inducible ADAR1-p150. Furthermore, it is overall encouraging to see within my top candidates, genes such as *EIF2AK2* as well as others (i.e. MDM4, ADAM19, LRRC28), whose transcripts were previously predicted to be edited by ADAR1 according to the REDIportal database (Lo

Giudice et al., 2020b; Picardi et al., 2017). Notably, MDM4, which is also a validated target of ADAR1 (Hong et al., 2018), is an inhibitor of p53 directly linked to tumor formation (Danovi et al., 2004). Therefore, provided that my hypothesis is true, which I experimentally address through proof of concept experiments in 4.2, aberrant ADAR1-dependent RNA editing may lead to generation of mutations in key components for MM progression.

I furthermore explored the patient data to evaluate the acquisition of specific T-derived mutations in TP2 post-relapse. I found that the newly acquired T-derived mutations were generally enriched within or near the editing-defined windows of TP1 (Figure 4.4A) and the abundance of editing-to-mutation events was strongly correlated with the ADAR1-dependent RNA editing activity in TP1, but not in TP2 (Figure 4.4B), which aligns with my original hypothesis that DNA mutation by ADAR1 may be generated as a "collateral damage" during its canonical RNA editing activity. I moreover wished to explore from the data available how the DNA mutation dynamic overlays with the RNA editing activity. I observed within my 23-patient cohort, that 14 patients decreased their RNA editing activity from TP1 to TP2, while 9 of them increased it or kept it at similar levels (Figure 4.4C). Therefore, I tested whether either of the group had an enrichment of T-derived mutations over the other. I found (Figure 4.4D) that patients who decrease their RNA editing activity had a significant enrichment of the newly acquired T-derived mutations in TP2, indicating that generation of DNA mutation by ADAR1 in key components for the disease *may* stabilize in the long run the effects of the editability of the cognate transcript. Mechanistically, this could be doable by the fact that the preferred target motifs of the enzyme have now an altered sequence preventing the enzyme to edit RNA *in situ*.

My findings thus far underscore a strong correlation between RNA editing and DNA mutation by ADAR1, a role which may have a strong impact on MM development and progression (summarized in Figure 4.5), and likely to other cancer types as ADAR1 is overexpressed in the vast majority of cancers (Han et al., 2015). I employed MM patient data from two timepoints of the disease and found newly acquired mutations in the later timepoint (TP2), attributable to ADAR1 editing activity in the earlier timepoint (TP1). The on-average variation frequency of newly acquired DNA mutations was about ~5%. The relatively low frequency of the newly acquired DNA mutations in TP2, may be explained due to the fact that these mutations are sub-clonal, because most of the individuals from the 23-patient cohort did not show high clonal expansion in their relapsed tumors, according to analyses of Dr. Laganà. Despite of the low frequency, such mutations may still be of high interest, because they provide the grounds for further selection and evolution, which is crucial for MM tumors (Walker et al., 2014; Corre et al., 2018). This phenomenon has been previously described as genetic surfing, according to which mutations may remain at low frequencies until they are brought to prominence through positive selection (Peischl et al., 2016). Additionally, it is important to recall the mutation rates of other deaminases presenting dual roles as RNA editors and DNA mutators; for instance, APOBEC1 mutates genomic DNA in a rate of about 1/10 000 bp (Saraconi et al., 2014).

All in all, I have presented patient data suggesting that ADAR1 may be both an RNA editor of certain transcripts, as well as a DNA mutator of their cognate genes. The latter may be an off-target effect or “collateral damage” to the genome, because of the aberrant RNA editing activity of ADAR1. This role may be shared by other deaminases, such as APOBEC1 (Saraconi et al., 2014) or APOBEC3A (Jalili et al., 2020), which can provide in-depth explanations about how cancers may expand their mutational spectra toward tumor generation, adaptation and evolution.

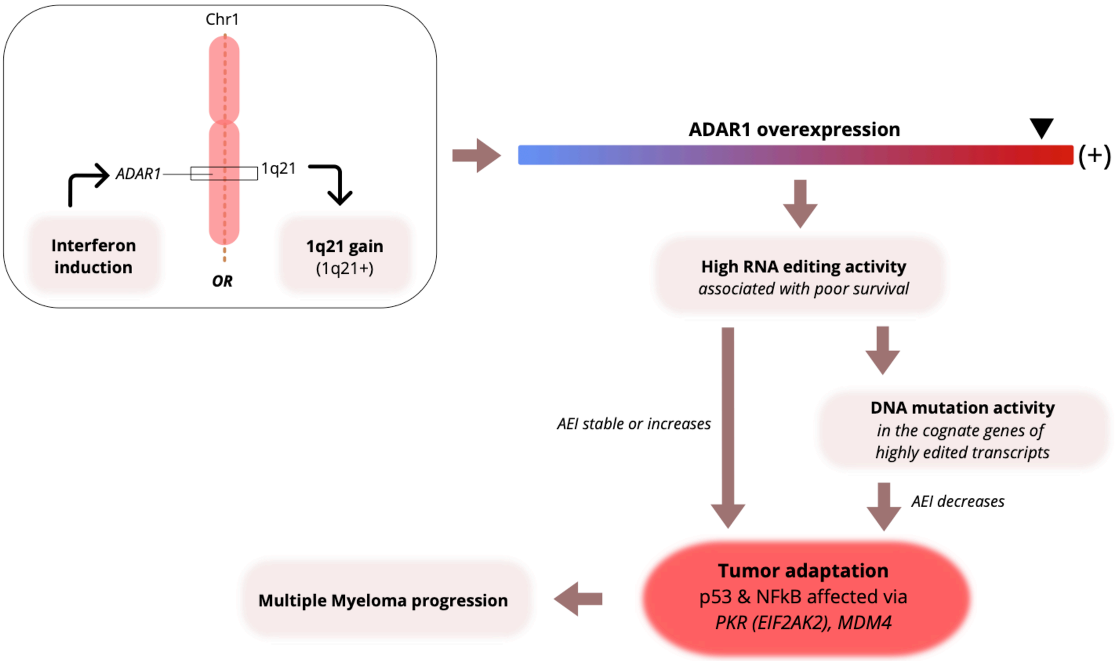


Figure 4.5 The proposed model of how the dual role of ADAR1 as an RNA editor and a DNA mutator may facilitate Multiple Myeloma progression. Multiple Myeloma (MM) tumors overexpress ADAR1 either through 1q21 copy-number gain (1q21+) or through interferon induction, leading to aberrant and elevated RNA editing activity, which is associated with poor survival. Aberrant RNA editing activity of highly edited transcripts may or may not lead to acquisition of ADAR1-dependent DNA mutations as “collateral damage” in their cognate genes resulting in decrease of RNA editing activity (AEI). Overall, ADAR1 may facilitate tumor adaptation in MM toward progression of the disease, both through RNA editing or DNA mutation in key components, such as *PKR* or *MDM4*, affecting the p53 or NFkB pathways among others. Figure from Tasakis et al., 2020. It is reused under the Creative Commons License 4.0. This illustration was produced by myself.

4.2 Aim 2: Experimental evidence of ADAR1-mediated mutagenesis

4.2.1 Preface

As introduced in chapter 1.3, there are currently several tools available for performing site-directed A-to-I RNA editing at specific adenosines in a transcript. In brief, this is chiefly possible with a small antisense oligoribonucleotide (RNA), complementary to the targeted transcript region, that transiently constitutes a dsRNA substrate for ADARs to act on. Such small RNA oligos are termed "guide" RNAs (hereafter gRNAs) and, depending on the tool employed, they either recruit endogenous ADARs or engineered enzymes, which typically incorporate the deaminase domains of ADARs and they are co-delivered with gRNAs (Casati et al., 2021). Overall, most of the site-directed mRNA editing tools available can edit efficiently on-target (adenosines on transcripts that are mismatched against cytidines on the gRNAs; see chapter 1.3), however off-target editing on the transcript, especially within the gRNA-defined targeted region, is also abundantly observed (Montiel-Gonzalez et al., 2013; Cox et al., 2017; Vallecillo-Viejo et al., 2017; Wettengel et al., 2017).

Recent findings revealed that ADARs can mutate DNA as well under certain circumstances. First, *in vitro* experiments with gRNAs targeting adenosines of the ssDNA M13 bacteriophage genome, showed that ADAR-mediated DNA editing is possible within DNA:RNA hybrids (Zheng et al., 2017). Then, it was recently shown also *in vitro* that in human telomeric repeats, which are prone to R-loop (DNA/RNA hybrid) formation, A:C mismatches in either the DNA or the RNA strand of an R-loop are resolved to I:C pairs through ADAR1 editing, allowing RNAse H2 to degrade the RNA strand of the hybrid and resolve the R-loop (Shiromoto et al., 2021). Indeed, depletion of ADAR1 in different telomerase-positive cancer cell lines led to telomeric R-loop accumulation and genomic instability, as shown in the same study (Shiromoto et al., 2021). Although it is evident from the aforementioned *in vitro* and cell-line data that ADAR1 is capable of mutating DNA, especially within the context of R-loops in human telomeres, it is not known how this mutagenic activity could be coordinated with its original function of being an RNA editor or whether ADAR1 can mutate DNA globally in the human genome.

I previously hypothesized (see 4.1.2), that ADAR1 can act both as RNA editor of a transcript and a DNA mutator of its cognate gene, within the context of R-loops formed co-transcriptionally. I tested this hypothesis in 23 Multiple Myeloma patients, who each had matched RNA-seq and WES tumor data from two timepoints of the disease, at diagnosis (pre-relapse) and at relapse (see 4.1), and I showed that newly acquired and specific mutations at relapse were predominantly found in genes, whose transcripts were highly edited by the overexpressed ADAR1 at diagnosis (pre-relapse). This data suggests that ADAR1 can potentially be a global DNA mutator, with related DNA mutations being acquired as *collateral damage* of its aberrant RNA editing activity. To prove that ADAR-mediated mutagenesis is a consequence of its RNA editing activity, I employ a series of site-directed mRNA editing experiments, for which I target specific transcripts and look for DNA mutations in their genomic loci. For targeting ADAR to specific transcripts, I leveraged the power of three available tools;

LEAPER, which promises to recruit the endogenously expressed ADAR1 with unmodified gRNAs (Qu et al., 2019; introduced in chapter 1.3.2), the 4λN-ADAR tool for which an ADAR deaminase domain, bound to λN-peptides, co-delivered with a gRNA that contains BoxB loops to tether the deaminase via the λN-peptides (Montiel-Gonzalez et al., 2013; introduced in 1.3.1) and the RESTORE tool, which recruits the endogenously expressed ADAR1 with gRNAs containing GluR2 loop motifs (Merkle et al., 2019; introduced in 1.3.2).

4.2.2 Loss of IgM through DNA mutation in the V region of Ramos cells

Activation Induced Cytidine Deaminase (AID) protein is the product of Aicda gene and is in principle responsible for somatic hypermutation (SHM) and class-switch recombination (CSR) of antibody genes (Muramatsu et al., 2000). Ramos Burkitt's lymphoma B cell line is a model for SHM; Ramos cells express IgM antibodies, which are further diversified only through constitutive hypermutation of the immunoglobulin V gene (V_H) at a rate of $2.8x10^{-3}$/bp mutations (Sale and Neuberger, 1998). Diversification of IgM through AID-dependent mutation in the V_H can be easily detected as IgM loss through FACS with about 10-20% of the cell population being IgM^- (Upton and Unniraman, 2011). Indeed, I stained Ramos WT cells with an anti-human IgM antibody and I observed loss of IgM at about 19.4% of the overall cell population (Q4, AID WT, Figure 4.6). When I stained AID-/- Ramos cells, in which expression of AID is lost, they also lose the ability to hypermutate and therefore I showed that they lose surface IgM expression (Figure 4.6 AID-/-). This has been previously reported as a measure of hypermutation (Cook et al., 2007). Finally, when I transfected Ramos AID-/- cells with plasmid vectors (mAID-cDNA-pl, Appendix 3) that express the cDNA of AID, they lose IgM again (Figure 4.6, AID-/- +mAID), as also shown in (Al-Qaisi et al., 2018).

Overall, these findings validate that the V_H of Ramos cells can function as mutation reporter, since mutations in that gene can lead to IgM loss, which is a phenotype that can be easily detected through FACS. Therefore, I leverage the power of this system to estimate whether ADAR-dependent RNA editing can lead to DNA mutation. In the next experiments (chapter 4.2.4), I target the V_H region of Ramos AID-/- cells with a site-directed mRNA editing tool (LEAPER, Qu et al., 2019), which recruits the endogenously expressed ADAR1.

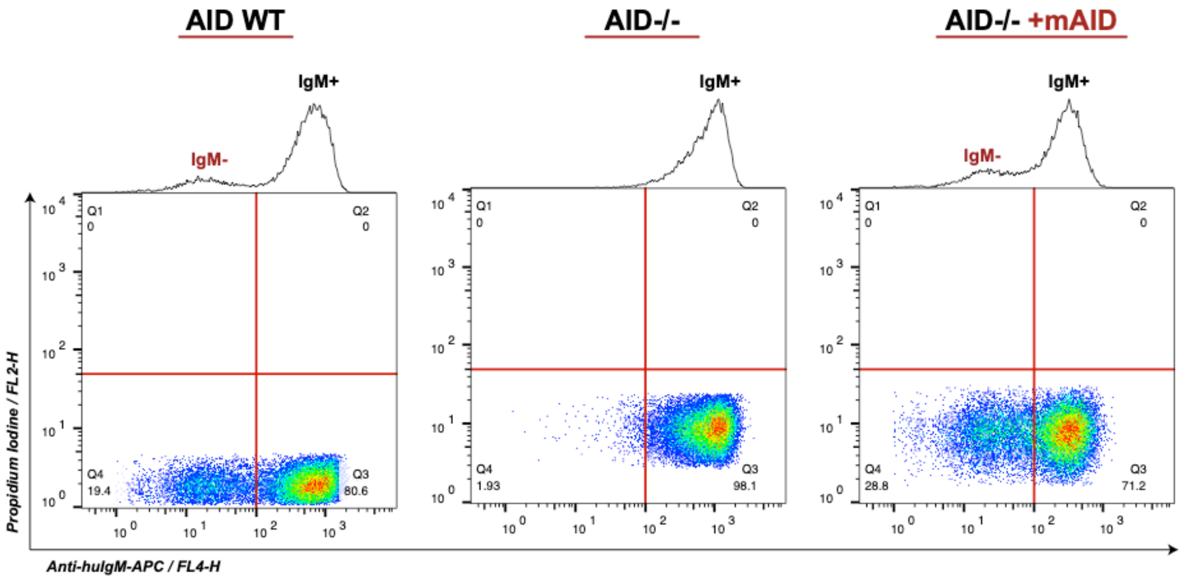


Figure 4.6 Loss of IgM in Ramos B cell line due to mutation in the immunoglobulin V gene by AID. Ramos wild-type (WT) cultures expressing AID have both IgM+ and IgM- populations (Q4, 19.4%). AID-/- Ramos cells, only have IgM+ populations, independently of how long they have been cultured (IgM- in Q4, <2%). Transient expression of mouse AID (mAID) with vectors expressing its cDNA, reconstitute the IgM- populations in AID-/- cells (Q4, 28.8%). Cells were stained with anti-human IgM antibody with APC fluorophore (anti-huIgM-APC) and Propidium Iodine to exclude dead cells. The gating strategy for the FACS analysis is provided in Appendix D, Part 1. Adapted figure from Tasakis et al., 2020. It is reused under the Creative Commons License 4.0. The processed data and illustration were produced by myself. The raw FACS data were generated jointly with Ms. Dimitra Stamkopoulou.

4.2.3 ADAR1 is the major A-to-I deaminase in Ramos cells

As introduced in chapter 1.2.1, A-to-I RNA editing is catalyzed by two enzymes of the ADAR family, ADAR1 and ADAR2. ADAR1 is ubiquitously expressed and is mostly responsible for the vast majority of A-to-I RNA editing in the transcriptome. Site-directed mRNA editing tools, and in particular the LEAPER tool which I employ here (see 4.2.4), reportedly recruit the endogenous ADAR1 for on-target editing with gRNAs, as discussed in chapter 1.3. Therefore, I validated with an expression and functional assay that ADAR1 is the major A-to-I deaminase in Ramos cells. Expression of ADAR1 and ADAR2 was measured from the cDNA of Ramos AID-/-, WT and ADAR1-/- cells with qPCR (Figure 4.7A). ADAR1 is abundantly expressed in Ramos AID-/- and WT cells, while absent in the ADAR1-/-. ADAR2 is expressed at very low levels in all Ramos cell lines. Low levels of ADAR1 detected with qPCR in the ADAR1-/- cells is due to the fact that the *ADAR1* locus that was knocked out (see 3.3.1) may still be transcribed, but ADAR1 is not functional in the same cell line. Indeed, when I performed Sanger sequencing of the endogenous MAVS cDNA, a known target of ADAR1 (Li et al., 2021), showed no A-to-I(G) RNA editing sites, while the same sites were edited in both Ramos WT and AID-/- cells. Overall, this data show that ADAR1 is expressed and functional in the Ramos AID-/- cells, in which I will target ADAR1 in their V_H region aiming to induce DNA mutations that will be reported as loss of IgM.

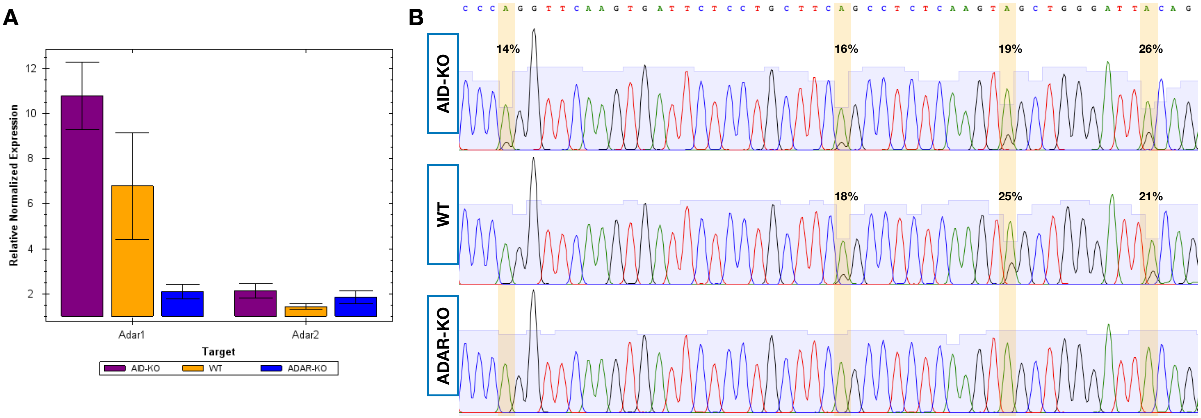


Figure 4.7 ADAR1 is the major A-to-I deaminase in Ramos cells. (**A**) qPCR from the cDNA of Ramos AID-/- (AID-KO), Ramos WT (WT) and ADAR-KO (ADAR-/-) cells for the transcripts of ADAR1 (primers *qADAR1-Fw* and *-Rv*, Appendix A) and ADAR2 (*qADAR1-Fw* and *-Rv*) showed abundant ADAR1 expression in AID-/- and WT Ramos cells and very low expression levels of ADAR2 in all cell lines. Expression was normalized with the housekeeping genes beta Actin (*qActb-Fw* and *-Rv*) and GAPDH (*qGAPDH-Fw* and *-Rv*). (**B**) RNA editing sites detected in cDNA amplicons of MAVS (*Mavs-Fw* and *Mavs-Rv*) in AID-/- (AID-KO) and WT Ramos cells were absent from the ADAR1-/- cell line, validating that ADAR1 is functional in Ramos AID-/- and WT cells. RNA editing sites are highlighted in yellow with the percentage (%) of editing on top of each site on the chromatogram.

4.2.4 Loss of IgM after targeting the V$_H$ transcript and gDNA with gRNAs

Up to this point, I have presented experiments that demonstrate that introduction of DNA mutations at the V$_H$ gene of Ramos AID-/- cell leads to IgM loss, which is an easy read out detectable by FACS. Furthermore, I have validated that ADAR1 is the predominantly A-to-I deaminase in Ramos AID-/- (and WT) cells in terms of expression and function. Here, I employ a site-directed mRNA editing tool, LEAPER, which has been previously shown to recruit the endogenously expressed ADAR1 with an antisense oligoribonucleotide (gRNA) to the target-transcript (Qu et al., 2019). However, the gRNAs employed to target a transcript are also complementary to the sequence of the coding strand (positive-sense) of the cognate genomic locus. It is known from *in vitro* experiments (Zheng et al., 2017), that ADARs can deaminate DNA within DNA/RNA hybrids. Therefore, I hypothesize that gRNAs can temporarily hybrid with the coding DNA strand of a genomic locus, whose transcript is targeted with a gRNA, allowing ADAR to mutate the coding strand (Figure 4.8A). Similarly, that should also happen when antisense oligoribonucleotides hybrid with the template DNA strand (negative-sense) and perhaps more efficient than antisense oligoribonucleotides to the template DNA strand, because there is no complementarity competition with the transcript.

I designed unmodified gRNA and oligos (78-81nt long) - hereby termed as arRNAs (ADAR-recruiting RNAs) - according to (Qu et al., 2019) for targeting the V$_H$ region against the transcript or coding strand, as well as against the template strand of the V$_H$ genomic locus. Scheme in Figure 4.8B summarizes all arRNAs employed and the regions of targeting within the V$_H$ locus. The sequences of all arRNAs are provided in Appendix B. arRNAs targeting the template DNA strand were complementary to the target throughout the sequence, while arRNAs targeting the coding strand, and therefore the transcript, were designed for an A:C mismatch centered within the target:arRNA hybrid.

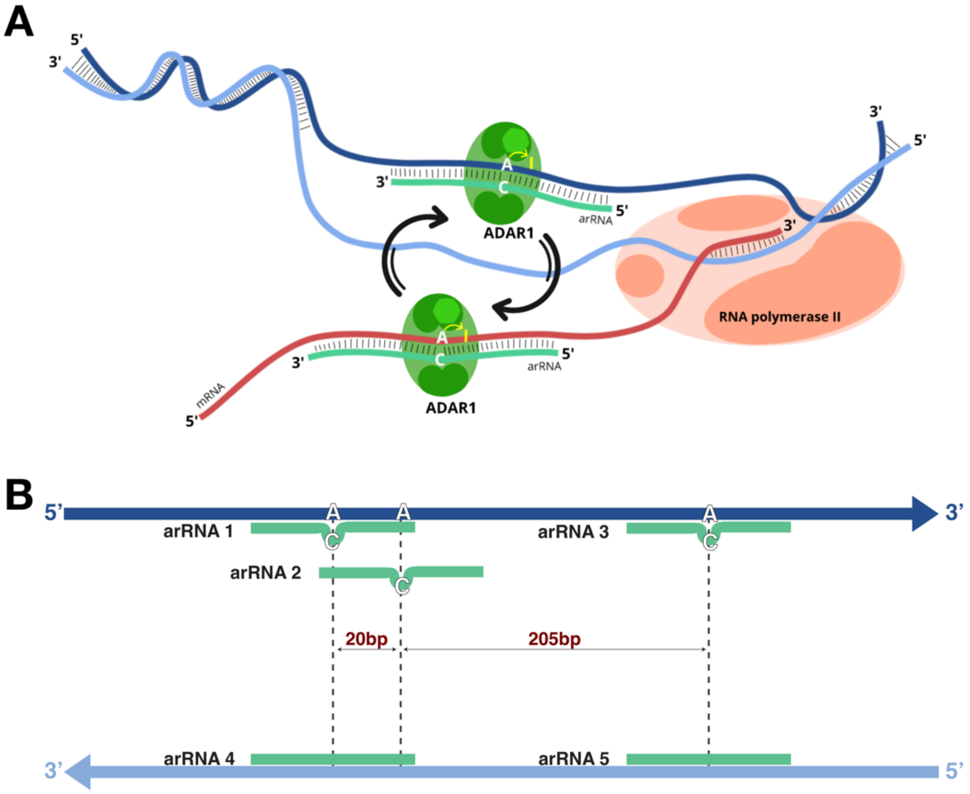


Figure 4.8 Targeting of the V_H gDNA with site-directed mRNA editing strategies. (**A**) Transcripts targeted with an antisense oligoribonucleotide (here arRNA, recruiting the endogenous ADAR1) may also temporarily hybridize with the coding strand of the of the genomic locus. ADAR1 may edit the coding DNA strand of the hybrid, as it has been shown *in vitro* that it can deaminate DNA within DNA:RNA hybrids (Zheng et al., 2017). (**B**) The layout of 5 arRNAs recruited to edit the V_H transcript or mutate the cognate genomic locus. arRNAs targeting the coding strand (dark blue) are antisense also to the transcript of the V_H locus and form an A:C mismatch at the center of the DNA:arRNA hybrid. arRNAs targeting the template strand (light blue) are designed to be complementary throughout the hybridized region of DNA:arRNA. arRNAs 1, 2 and 4 were 81nt long and arRNAs 3 and 5 were 78nt long. The sequence of arRNAs is provided in Appendix B.

Although the capability of ADARs mutating DNA has been shown *in vitro* (Zheng et al., 2017) and correlated with cell phenotypes in cell lines (Shiromoto et al., 2021), the potential mutation rate of ADAR1 is not known. It is expected that the mutation rate is lower than the rate of RNA editing by the same deaminase (Saraconi et al., 2014) and, thus, it is very likely that mutations by ADAR1 may be undetectable with the strategies employed in the field thus far. This is after compiling knowledge from other RNA editors showed to also be DNA mutators (i.e. APOBEC1 from Saraconi et al., 2014) as well as from the previously presented correlative editing-to-mutation data in Multiple Myeloma (Chapters 4.1.2 and 4.1.4, Tasakis et al., 2020). In Ramos cells, the mutation reporter system I employ, it has been observed that loss of IgM through SHM is often accompanied by DNA strand breaks (DSBs) within the V_H (Sale and Neuberger, 1998). Therefore, I employ pairs of arRNAs (shown in Figure 4.8B) targeting the V_H region of Ramos AID-/- cells at distances within the range of 205-225bp (by the center of the arRNA), so as to enhance loss of IgM through DSBs, if ADAR1 has mutated the targeted genomic locus of V_H. I employed different pairs of the arRNAs, targeting the coding strand (arRNA1+3, aRNA2+3), template strand (arRNA4+5) or both strands (arRNA1+5, arRNA2+5). arRNAs were expressed by the same U6-vector (gRNA-pl, Appendix C) transfected in Ramos AID-/- cells. As a positive control, I transfected a vector expressing the cDNA of mAID (mAID-cDNA-pl, Appendix C) and as negative control an arRNA (hereafter as arRNA(-)) with no target in the genome (Appendix B). An additional negative control was a mock culture of Ramos AID-/- cells. Upon transfection, I cultured the cells for 5 weeks in bulk. FACS analysis of viable cells stained with an anti-human IgM antibody bound to APC fluorophore is summarized in Figure 4.9A and I performed it jointly with Ms Dimitra Stamkopoulou. I observed loss of IgM in two cultures in which the pairs of arRNA1 and arRNA5 (arRNA1+5) or arRNA2 and arRNA5 (arRNA2+5) were co-delivered. The IgM-negative (IgM^-) population for arRNA1+5 was 14.7%, while for arRNA2+5 was 26.7%. gDNA amplicons of the V_H (primers *Vh-gDNA-Fw* and *-Rv*, Appendix A) followed by Sanger Sequencing showed a signal of T-derived mutations in the vicinity of the on-target RNA editing site (Figure 4.9B). Signal of mutations was absent from the arRNA(-) control (Appendix D, Part 3).

Both samples in which I observed loss of IgM, had a common layout with regards to the pair of arRNAs, which is summarized in Figure 4.9C; an 81nt-long arRNA antisense to the coding strand (and also the cognate transcript) centered by a C in its sequence towards the A:C mismatch within the hybrid of target (arRNAs 1 or 2), and a 78-nt long arRNA antisense to the template strand (arRNA 5) at a distance of about 200bp from their centers. To explore the possibility of IgM loss due to DNA mutation or DSBs within the window of ~200bp, in the two samples that prominently showed IgM^- populations, I performed deep-amplicon NGS in V_H amplicons from the bulk gDNA of both samples (arRNA1+5 and arRNA2+5) and the AID-/- negative control (Mock). Although in the predominant IgM^- samples more sequences (a sequence is a read-pair covering the entire amplicon) appeared heterogeneously mutated, the specificity of mutation in certain sites was minimal (Figure 4.9D). Similarly, even though there were widespread small gaps (2-8bp) per sequence in the IgM^- samples compared to the control, large deletions (>30bp) were not abundant (Figure 4.9E).

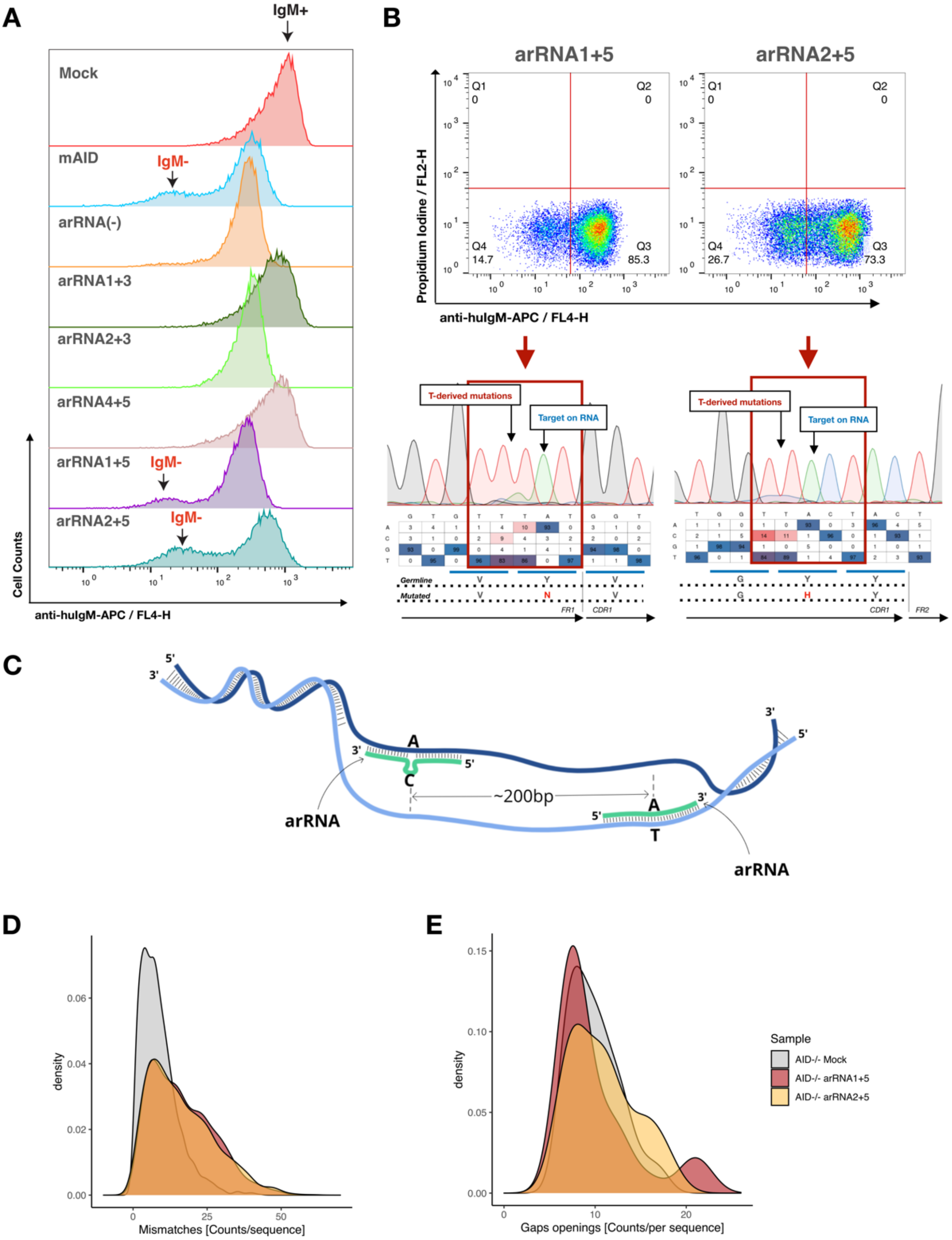


Figure 4.9 Loss of IgM upon arRNA delivery and associated outcomes in the gDNA of V_H. (**A**) Histogram summarizing the IgM^+ population is noted at the Mock negative control (Ramos AID-/- cells from culture) and IgM^- populations in the positive control (Ramos AID-/- transfected with the cDNA of mouse AID - mAID) and the samples arRNA1+5 and arRNA2+5 in which abundant IgM loss was observed. The IgM^- population is virtually absent from AID-/- and very minimal in the arRNA(-) negative control (arRNA with no target), while the IgM^- subpopulation peaks emerge in the arRNA1+5 and arRNA2+5 samples 5 weeks after transfection. Dotplots are available in Appendix D, Part 2. (**B**) AID-/- cells transfected with pairs emerge in the arRNA1+5 and arRNA2+5 show abundant loss of IgM versus cells transfected with the control arRNA. T-derived mutation signal is reported in V_H amplicons from gDNA, which was

extracted from bulk cultures, in the region where ADAR1 was targeted to on the V_H transcript. For the arRNA1+5 transfected cells, signal of T-to-A mutation right next to the on-target A, lead to a Y-to-N amino-acid change (tyrosine-to-asparagine) altering the last codon of the Framework Region 1 (FR1; according to IMGT). While for the arRNA2+5 transfected cells, T-to-C mutation signal in the two upstream positions from the A-target on the RNA are observed, which are in the range of the last three codons of CDR1 by IMGT. This mutation leads to a Y-to-H (tyrosine-to-histidine) amino acid change. These mutation signals were absent from the arRNA(-) sample (see Appendix D, Part 3). (**C**) arRNAs combinations presenting the most prominent IgM$^-$ populations have the same layout of bi-stranded targeting: an 81nt-long arRNA is antisense to the coding positive-sense strand with an A:C mismatch in hybrid and a 78-nt arRNA antisense to the template (negative-sense) strand. The arRNAs are in ~200bp distance measured by the center of their sequences. (**D, E**) Deep amplicon NGS data from the highly IgM$^-$ populated samples (arRNA1+5 and arRNA2+5) blasted by read-pair against the germline sequence of Ramos V_H, reveal numerous mutations, as demonstrated by a shift in the distributions of counts of mismatches per sequence, when compared to the Mock negative control. No site-specific mutations were detected in depth. The same data also revealed a higher count of small gap openings (2-8bp) in sequences (read-pairs) from the arRNA-transfected samples. The gating strategy for Ramos B cells is provided in Appendix D, Part 1 and FACS data are down-sampled to 20000 cells per sample for all the samples. Panel B is adapted from Tasakis et al., 2020 under the Creative Commons License 4.0 and was produced by myself. The raw FACS data presented in panels A and B were generated jointly with Ms Dimitra Stamkopoulou.

The data I presented thus far, have given indications of potential ADAR1-dependent DNA mutations as reported by the loss of IgM in AID-/- Ramos B cells. However, these observations have further raised two major questions. First, I observed a transient but significant drop of IgM expression (Figure 4.9A), which rebound later. This, may be due to either RNAi (RNA interference; Hannon, 2002) effect for the arRNAs antisense to the transcript (arRNA1, 2 and 3) or also due to short R-loop formation with the template strand, thus generating a substrate for RNAseH (Shiromoto et al., 2021). To test this, I transfected the same arRNAs directly in Ramos cells upon *in vitro* transcription (IVT) and, indeed, I observed loss of IgM within a 3-week window post-transfection (Figure 4.10), as further discussed below. Second, the loss of IgM post-rebound I present here, derive from bulk cultures at 5 weeks post-transfection with arRNAs expressed from transfected plasmid vectors (gRNA-pl, Appendix C). However, in those cultures I did not detect evidence of targeted V_H RNA editing early in transfection (between 24h and 120h post-transfection), again quite possibly because of the overall drop in IgM transcript levels, so that the transcripts actually targeted for editing would also be removed by RNAi or RNAseH. These experiments offered valuable insights into the system, which I have redesigned accordingly and I provide an alternative solution in the subchapter 4.2.5.

As I mentioned above, I observed an overall decrease in IgM expression, which may be due to RNAi effects of the arRNAs antisense to the transcript or resolution by RNAseH of a temporary short R-loop formed between the arRNA and the template (negative-sense) strand. To evaluate whether absence of RNA editing from cDNA amplicons of the V_H region, may be due to such effects, I transfected Ramos AID-/- cells with each arRNA produced by IVT separately and I monitored loss of IgM between 24h and 21 days post-transfection. As positive control, I included cells from Ramos AID WT culture. For negative controls, I transfected the same arRNA(-) to Ramos AID-/- cells. Apart from the mock sample as before (culture of AID-/- cells), I included an additional negative control (mock-trf), which is cells without any arRNA undergoing the transfection process. The latter was added because in the previous experiment there was a minimal IgM$^-$ population in the arRNA(-) sample (Figure 4.9A, Appendix D, Part 2, panel arRNA(-)) and I wanted to confirm that this was background noise, perhaps due to transfection. Indeed, as I show in Figure 4.10, both the mock-trf and mock samples showed similar minimal abundance of IgM$^-$ populations (~5%). As expected, Ramos WT cells presented IgM$^-$ populations between ~13-26%, while AID-/- culture was purely IgM$^+$. At 24h post-transfection, I did not observe IgM$^-$ populations, but 7days or even 14 days post-transfection IgM$^-$ populations reached up to ~96%. At 21 days post-transfection most of the populations retreated back to being primarily IgM$^+$, except for arRNA5 which showed an increase from 21% to 40.7% compared to the previous timepoint. I did not detect RNA editing or DNA mutations in the V$_H$ cDNA and gDNA amplicons from the bulk for neither of the timepoints.

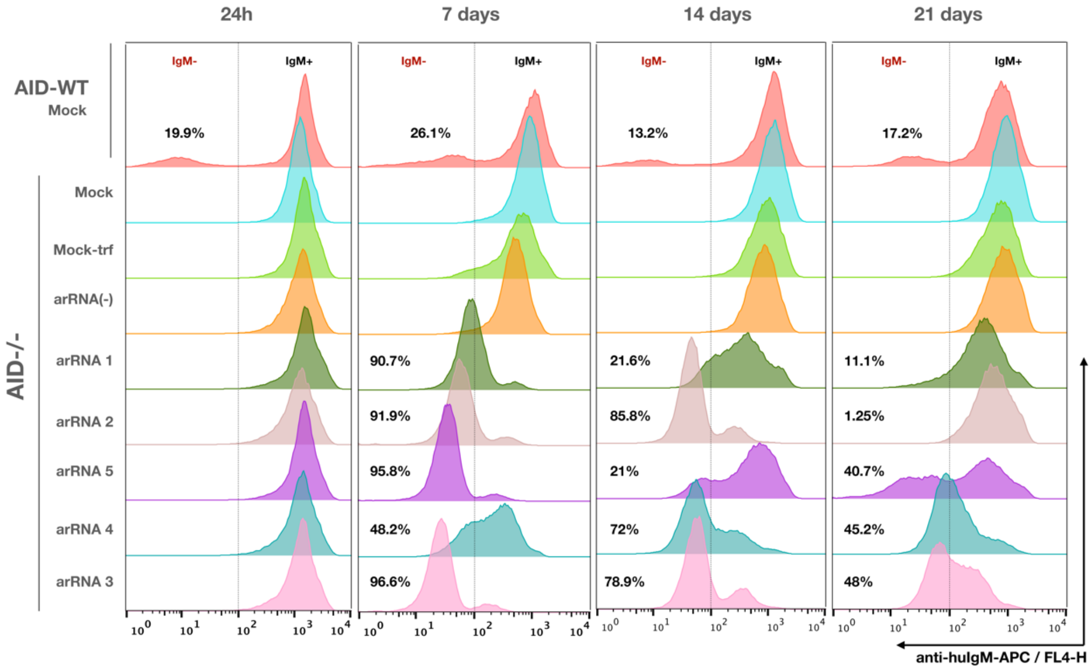


Figure 4.10 IgM loss through transient mechanisms independent of deamination. Previous observations of overall decrease in IgM expression indicated that there might be mechanisms, such as RNAi or resolution of transient R-loops between arRNAs and the template negative-sense DNA strand by RNAseH. arRNAs were delivered directly upon *in vitro* transcription to Ramos AID-/- cells and loss of IgM was monitored between 24h and 7 days post-transfection. Abundant loss of IgM at 7- and 14-days post-transfection in combination with IgM gain at 21 days post-transfection, indicate that arRNAs can interfere with expression of the locus or the transcript independent of deamination. Positive control was Ramos WT (AID WT) cells from culture. Negative controls were Ramos AID-/- cells from the culture (Mock), Ramos AID-/- cells transfected without RNA (Mock-trf) and Ramos AID-/- cells transfected with arRNA(-). Percentages in the IgM$^-$ panels of the histograms note the percentage of the IgM$^-$ population. arRNA sequences are provided in Appendix B. Gating strategies of Ramos cells are provided in Appendix D, Part 1.

My findings highlight that loss of IgM may be a valuable system for reporting DNA mutation, as previously shown (Sale and Neuberger, 1998), but it is not ideal at the current form for RNA editing, as it may be masked by various parameters as shown above, such as transient loss of IgM, perhaps due to an RNAi-like effect (Hannon, 2002) or even a downregulation due to RNAseH mediated resolution of R-loops (Shiromoto et al., 2021). While I reckon that the Ramos system I presented thus far is valuable, it needs to be engineered into one whose readout is **gain**, rather than loss of IgM, where gain is the result of a **stop codon reversion**, a situation to which ADAR-mediated editing is naturally suited (Montiel-Gonzalez et al., 2019). Such a system would be to engineer a Ramos AID-/- cell line that is IgM$^-$ due to a premature stop codon UAG in the V_H region, such that upon U**A**G>U**G**G editing it would lead to IgM gain, also detectable via FACS. This would ensure that the original population is edited on the RNA and only this population would be further evaluated for DNA mutation. This experiment is currently ongoing and such cell line is being created by exploring and propagating the IgM$^-$ cells upon transient transfection of AID cDNA in Ramos AID-/- cells, followed by single-cell clonal expansion and selection of a clone with a specific stop codon which can then be reverted using relevant gRNAs (Figure 4.6, panel AID-/- +mAID). This was inspired by a cell line experiment (HEK293T-W58X, see 3.3.1), in which a premature UAG is present within an inactivated eGFP gene that is transcribed; upon editing (U**A**G>U**G**G) an active eGFP protein is produced. This system for reporting editing was originally described in (Montiel-Gonzalez et al., 2013) and in the next subchapter I will present data that indicate permanent activation of eGFP by ADAR1, very likely through DNA mutation in clones that originated from a priming population of cells, purely editing the eGFP activation site on the RNA.

4.2.5 eGFP activation: a gain of function strategy to report ADAR-dependent mutations

As I concluded above, one of the major parameters for detecting RNA editing in endogenous transcripts is to eliminate the background of cells, which were not successfully transfected with the components required for site-directed mRNA editing. To achieve that, I present an alternative experimental set up, originally employed in Montiel-Gonzalez et al., 2013, which provides an easily readable output for cell populations with successfully edited endogenous transcripts. For this experiment I employed a HEK cell line (HEK-293T-W58X, see chapter 3.3.1), which expresses a cassette comprising of a gene expressing mCherry fluorescent protein and a gene expressing an inactivated eGFP protein due to a premature stop codon UAG, separated by the sequence of a self-cleaving 2A peptide (Figure 4.11A). Site-directed A-to-I(G) editing targeting the aforementioned UAG stop codon on the transcript, transiently changes the stop codon sequence to a tryptophan codon (U**A**G>U**G**G), leading therefore an actively fluorescent eGFP protein (Figure 4.11A). This system, providing a readout that can be easily monitored through FACS, allowed me to isolate a priming cell population, in which RNA editing reportedly occurs, and which I focused on for detecting ADAR-dependent DNA mutation on-target.

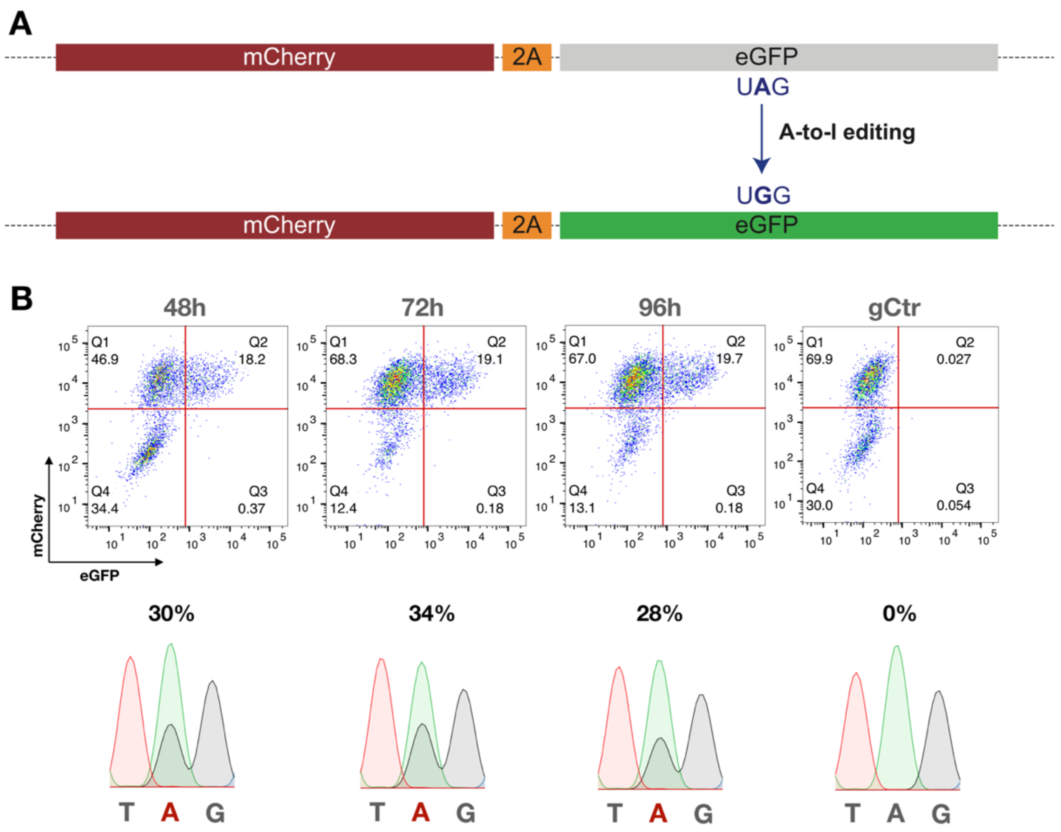


Figure 4.11 eGFP activation through A-to-I site-directed RNA editing. (**A**) Stably transfected HEK cells expressing a cassette of mCherry, 2A and inactivated eGFP due to a premature UAG stop codon were employed to monitor site-directed mRNA editing. Cells with successfully A-to-I edited UAG>U**G**G codons express an activated fluorescent eGFP. This system has been originally described in Montiel-Gonzalez et al., 2013. (**B**) The λN-ADAR tool with gRNAs targeting the eGFP activation site (gGFP) was employed and RNA editing was monitored in the transfected cell populations at 48h, 72h and 96h post-transfection. GFP-activated populations (Q2) were ranging between 18.2% and 19.1% of their original populations. As negative control, cells were transfected with 4λN-ADAR and a gRNA with no target against the transcriptome (gCtr). RNA editing was quantified for the gGFP samples by Sanger sequencing chromatograms of cDNA amplicons (*pmCherry-Fw* and *peGFP-Rv*, Appendix A) of the mCherry and eGFP cassette from the total RNA of the GFP-activated populations (Q2: mCherry-positive and eGFP-positive) and for the gCtr control by the Q1 population (mCherry+). On-target RNA editing was 30% at 48h, 34% at 72h and 28% at 96h post-transfection, while no editing was detected from gCtr. Negative controls were set for all timepoints, the gCtr included in the figure is from 72h post-transfection. Gating strategies for the FACS analyses are available in Appendix D, Part 4.

For the following experiments I employed the λN-ADAR site-directed mRNA editing tool (Montiel-Gonzalez et al., 2013 and chapter 1.3.1, for which a gRNA with BoxB loops recruits an ADAR deaminase domain, which is bound to λN peptides. In this set-up I particularly employed a λN-ADAR that has 4 λN peptides, expressed by a plasmid (4λN-ADAR-pl, Appendix C) upon transfection in HEK cells (see 3.3.3). Additionally, a gRNA with two BoxB loops targeting the aforementioned UAG stop codon of eGFP (gGFP) and a control gRNA with no target in the transcriptome (gCtr) were employed (sequences in Appendix B). gRNAs were also expressed by plasmid vectors, under the same promoter and characteristics (gRNA-pl, Appendix C). I transfected HEK cells expressing the mCherry-2A-eGFP cassette with co-delivered plasmid vectors for 4λN-ADAR and gGFP (at ratios 1:5) and I monitored between 48h and 96h post-transfection for eGFP activation with FACS. Negative controls were set for all timepoints by co-delivering 4λN-ADAR and gCtr in the same amounts at the same number of cells. As shown in Figure 4.11B, the GFP-activated cell populations were ranging between 18.2%-19.7% (Q2) throughout the timepoints tested, while entirely absent from the negative controls (gCtr). I sorted the double-positive for mCherry and eGFP cells and I generated cDNA amplicons (primers *pmCherry-Fw* and *peGFP-Rv*, Appendix A) for the mCherry-2A-eGFP cassette from the total RNA of each population. On-target RNA editing in the UAG stop codon, quantified from Sanger sequencing chromatograms, was ranging between 28% and 34%, peaking at 72h post-transfection (Figure 4.11B). No on-target editing was detected in the gCtr negative controls (Q1 population, mCherry-positive) in amplicons from the same transcript.

As shown above, there is a discrepancy between the levels of editing measured by the eGFP activation on the cellular level and the percentage of editing measured on the transcript level. This is because even low levels of edited transcripts expressed by the mCherry-2A-eGFP cassette will produce active eGFP protein within a cell. It is also not entirely clear, up to date, whether RNA edited transcripts are differentially translated compared to unedited transcripts from the same locus. I previously discussed (chapters 4.1.2 and 4.1.4) that DNA mutation by ADAR1 is more likely to rise in genes of highly edited transcripts. Therefore, employing the eGFP-activation system presented thus far, I looked for ADAR-dependent mutations in cells primed by a cell population with the highest editing levels on the transcript thus far. Therefore, I repeated a transfection of HEK cells expressing the mCherry-2A-eGFP cassette, with the same layout and controls (see chapter 3.3.3) and 72h post-transfection, I sorted the strictly double-positive (mCherry+ and eGFP+) population (25 000 cells) in bulk (Figure 4.12A). No eGFP+ cells were recorded in the control (gCtr). I validated on-target RNA editing from a second replicate in the gGFP sample at 33% editing 72h post-transfection, and no RNA editing by the gCtr (Figure 4.12A). I further cultured in bulk the sorted double-positive cells (25 000) from the sample that showed eGFP activation due to editing and the same number of cells from the gCtr were further cultured for a total of 2 weeks post-transfection.

A window of time, such as 2 weeks post-transfection, is crucial for the eGFP-activated priming population to expand and meanwhile lose the plasmids of the original transfection. Especially in this context this is important, because the goal is to obtain permanent GFP-activated clones due to U**A**G>U**G**G DNA mutation on-target by ADAR. After 2 weeks post-transfection, I sorted the propagated cells from the original GFP-activated population (gGFP – 2 weeks) and the control (gCtr – 2 weeks) as single cells in 96-well plates. As demonstrated in Figure 4.12B very few cells remained double-positive (~5 /10 000 recorded cells) from the GFP-activated priming population (gGFP – 2 weeks), while the control remained virtually clear (gCtr – 2 weeks). I managed to rescue about 400 cells in total from the gGFP – 2 weeks sample. After 2 weeks, I screened the clones which successfully grew (~90) for eGFP with optical microscopy. I observed that 1 clone presented a heterogeneously eGFP+ population. Since the relevant plasmids were not under selection, this observation suggests a DNA mutation event. Indeed, transformation of *E. coli* competent cells with lysate of the aforementioned clone, growing in the appropriate antibiotic resistant agar plates for the respective plasmid (Appendix C), did not reveal any bacterial colonies.

I expanded the aforementioned clone for an additional week and I sorted at 5 weeks overall post-transfection, again for the double-positive population, but this time this population included more intensely eGFP+ cells (Figure 4.12C). To confirm that this is a DNA mutation event, I sorted the double-positive population (~850 cells) in lysis buffer and I extracted gDNA as well as total RNA were simultaneously by the lysate. I generated and sequenced amplicons from the cDNA and the gDNA of the mCherry-2A-eGFP. I detected a double A-to-G peak in the on-target codon indicating 18% base change from the cDNA, though not from the gDNA amplicon. Although this observation would traditionally indicate an RNA editing event and not a DNA mutation event, it is unlikely that this is the case. First, because I did not detect the plasmids encoding the required components for RNA editing in the clone as mentioned above, and secondly, it is unknown how many copies of the cassette are in the cells of the clone. After a closer inspection of the FACS plots for the heterogeneously GFP-activated clone (Figure 4.12C), it is obvious that mCherry fluorescence for the double-positive population is also very high, while this cell line alone robustly presents a distribution of mCherry intensities (y axis in all FACS plots of Figures 4.11 and 4.12). These observations indicate that the mCherry-2A-eGFP cassette is heterogeneously present in different number of genomic copies in the cell population and, therefore, it is very likely, for example, if only one genomic copy has a U**A**G>U**G**G mutation and it is expressed, while there are more unmutated that are not expressed (or at least not all), then detection of this mutation is not possible with Sanger sequencing. In a publication I have also contributed to, we have previously shown that the limit of accurate detection for base editing is 5% (Kluesner et al., 2021).

To recapitulate the findings, I isolated 25 000 eGFP-activated cells through A-to-I RNA editing 72h post-transfection and I further expanded them for growth. After 2 weeks of expansion, in which the originally transfected plasmids should be minimally present, I observed that very few cells remained eGFP-activated. I single-cell sorted about 400 cells and ~25% of them grew to clones. 1 clone presented heterogeneous eGFP-activation. I simultaneously extracted gDNA and total RNA from the eGFP+ cells from the lysate of that clone. I generated cDNA and gDNA amplicons for the mCherry-2A-eGFP cassette. 5-weeks post-transfection, I detected 18% A-to-G base change on-target in the UAG codon which activates eGFP from the cDNA amplicon.

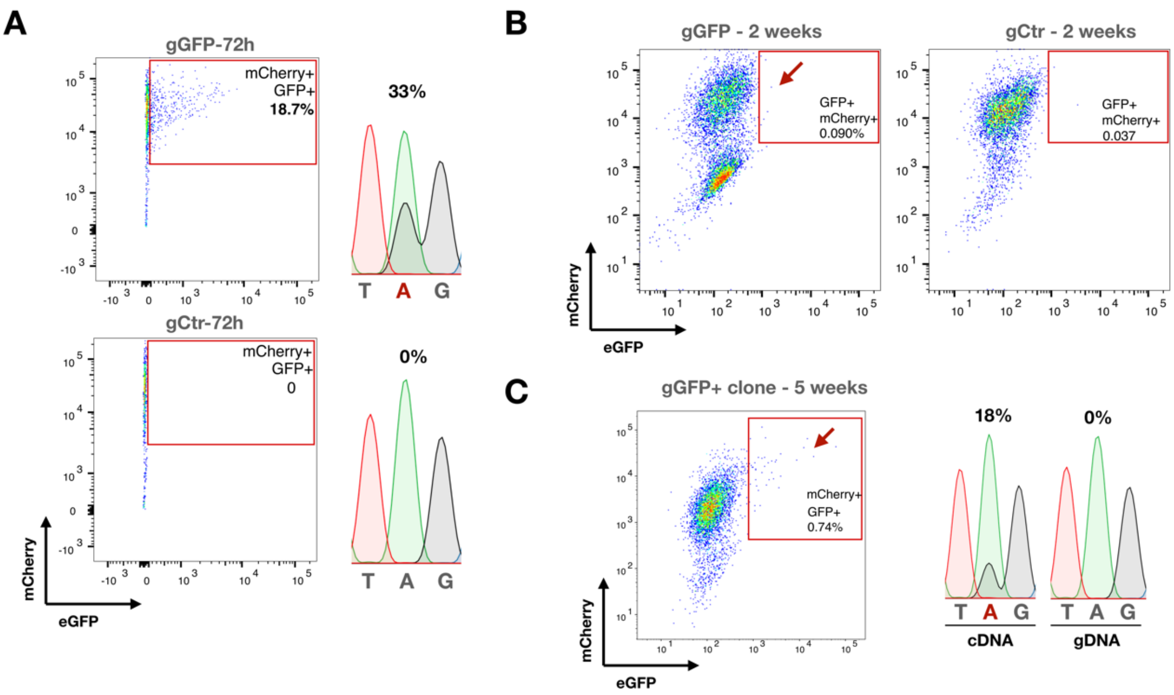


Figure 4.12 Strategy for detection of ADAR-dependent DNA mutation through eGFP-activation in HEK cells. (**A**) HEK cells expressing the mCherry-2A-eGFP cassette, which encodes for an inactivated eGFP due to a premature UAG stop codon (see figure 3.6A) were transfected with 4λN-ADAR and a gRNA targeting the UAG stop codon towards eGFP activation (gGFP). Upon 72h post-transfection the double-positive (mCherry+ GFP+) cells, in which 33% editing was detected on target from cDNA amplicons of the cassette, were sorted in bulk (25 000 cells). A negative control in which a gRNA with no target (gCtr) was transfected instead, did not show double-positive cells and no on-target editing from cDNA of the same cassette. (**B**) Double-positive cells from the gGFP sample and same number of cells from the gCtr sample were propagated for further growth for 2 weeks. Upon propagation, very few cells remained double positive (red arrow in panel gGFP – 2 weeks), while the same cells were virtually absent from the gCtr sample (gCtr – 2 weeks). Double-positive cells were sorted in 96-well plates as single cells, rescuing about 400 double-positive cells in total. (**C**) After 2 weeks, clones from about 25% of the original number of single cells grew. Clones were screened for eGFP+ signal and 1 clone with heterogeneous eGFP signal was rescued. The clone was cultured for an additional week and sorted, revealing a few double positive cells (red arrow, mCherry+ GFP+, 0.74%). Double-positive cells were sorted in lysis buffer and total RNA and gDNA was extracted simultaneously. Amplicons from the cDNA and gDNA of the cassette (*pmCherry-Fw* and *peGFP-Rv*, Appendix A) were sequenced and A-to-G base change was detected from the cDNA amplicon at 18% on-target. Gating strategies for FACS data are provided in Appendix D, Part 4.

My preliminary findings presented thus far indicated that ADARs *may* mutate genomic DNA in a rate of 1 in 25 000, estimated by the number of GFP-activated clone at the endpoint that rose from a population 25 000 edited cells. Of course, this was only an initial observation. In order to show DNA mutation from gDNA amplicons and additionally to validate the aforementioned rate, I repeated the experiment as described above and upscaled the priming population of edited cells to 350 000 at 72h post-transfection. This time, I kept them in culture for additionally 2 more weeks prior to single-cell sorting for clone generation. At the endpoint (7 weeks overall post-transfection), I obtained 7 clones which were again heterogeneously GFP+ (Figure 4.13A), from which I detected 23% A-to-G base change from the clone IV, again from the cDNA amplicon of double-positive sorted cells (Figure 4.13B). In this repeat, the respective rate would be 1 in 50 000. Although these findings were encouraging, I did not detect DNA mutation from the gDNA amplicons. Therefore, I wondered whether this had something to do with the fact that the λN-ADAR tool employs an engineered deaminase that needs to be exogenously co-delivered with the gRNAs (Montiel-Gonzalez et al., 2013; Montiel-González et al., 2016). For this reason, I performed another experiment (see 4.2.6) following the same strategy as described above, but this time with the tool RESTORE (Merkle et al., 2019) which recruits the ADAR1 for site-directed editing.

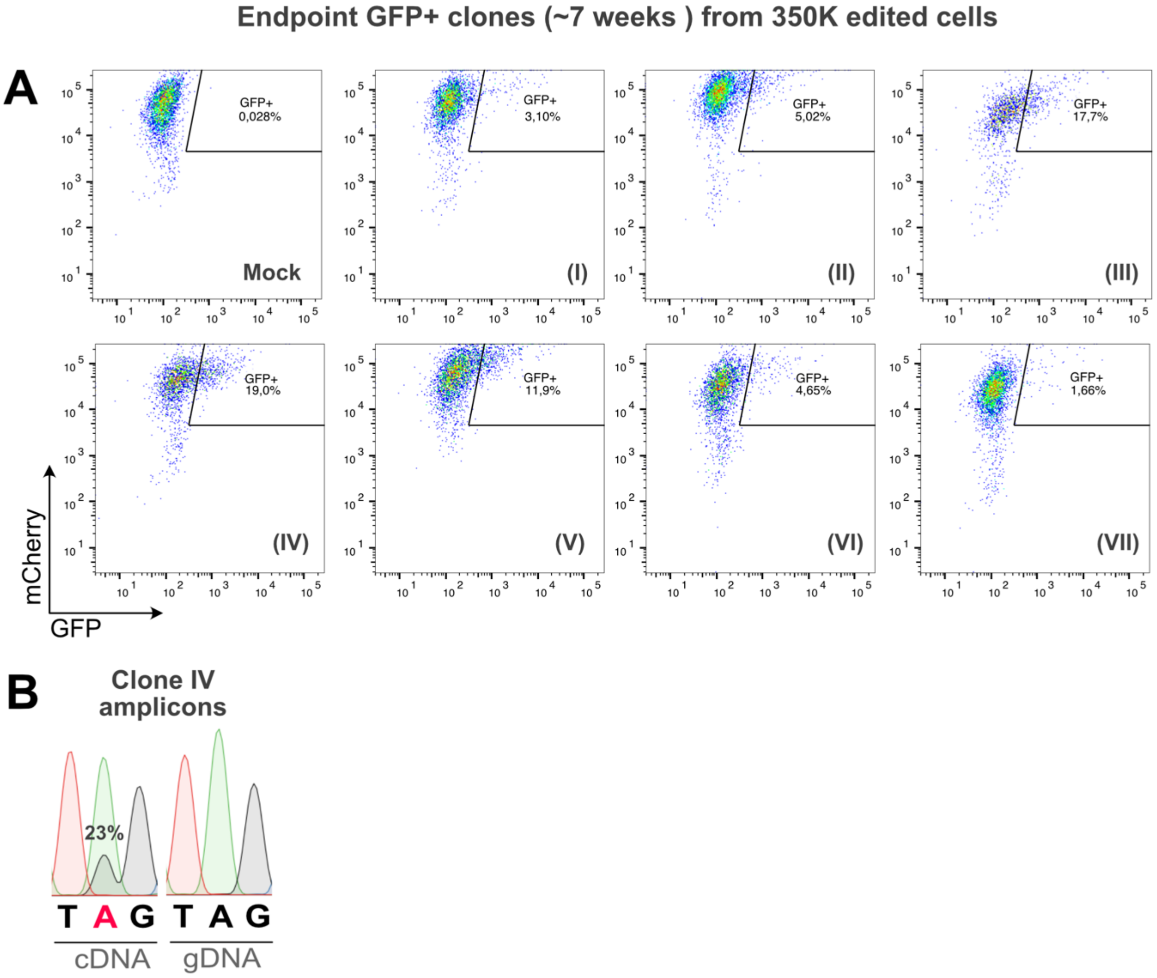


Figure 4.13 Upscaled repeat for detecting ADAR-dependent DNA mutation with the λN-ADAR tool. Following the same strategy as described and summarized in figure 4.12, (**A**) 7 clones were detected as heterogeneously GFP+ at the endpoint (7 weeks post-transfection), originating from 350 000 (noted as 350K) cells as priming edited population. The double-positive cells (mCherry+ and GFP+, noted in the FACS plots as GFP+) ranged from about 2% up to 19% of the overall population in the respective clones. Mock was a negative control of propagated cultured cells upon transfection with the gCtr (Appendix B). (**B**) 20 000 double-positive cells from the upscaled clone IV were sorted in lysis buffer for total RNA and gDNA extraction. cDNA and gDNA amplicons (*pmCherry-Fw* and *peGFP-Rv*, Appendix A) from the cassette were generated and sequenced. 23% A-to-G base change was detected from the cDNA on-target. Gating strategies for the FACS data followed as described in Appendix D, Part 4.

4.2.6 Validation of ADAR1-dependent DNA mutation upon eGFP activation through RNA editing in HEK cells

Up to this point I have shown that gain of function methods (i.e. activation of GFP by reversing a premature stop codon, see 4.2.5) are more reliable to report site-directed RNA editing than loss of function methods (i.e. loss of IgM, see 4.2.4). By employing the cell line HEK-293T-W58X (see chapter 3.3.1), which expresses a cassette of mCherry and an inactive GFP due to a premature stop codon (Figure 4.11), I showed that gain of GFP signal through site-directed mRNA editing and monitoring potentially subsequent DNA mutation events is easily detectable through FACS, as also validated from sequencing. I previously showed (see 4.2.5) employing the 4λN-ADAR system that the potential mutation rate by ADAR could be 1 in 25 000 (number of endpoint GFP+ clones in the total number of priming edited cells). However, the desired A-to-G base change was from cDNA amplicons but not from gDNA. Here, I employ a different tool, RESTORE (Merkle et al., 2019 and chapter 1.3.2), which recruits the endogenously expressed ADAR1 as reported by Merkle et al., 2019, which is the major A-to-I deaminase in HEK cells, with ADAR2 being minimally expressed in the same cell line (Schaffer et al., 2020). My colleagues, Dr. Riccardo Pecori, Ms Beatrice Casati, previously produced a plasmid vector missing the last 5bp from the CMV promoter and with that being the only difference with the gRNA-pl (Appendix C) vector I have used throughout my experiments. Their version of the gRNA-pl, expressing a gRNA targeting the premature UAG stop codon toward GFP activation (gGFP_R, sequence in Appendix B), showed higher RNA editing efficiency at 72h post-transfection at on the transcript level than the usual gRNA-pl vector. Therefore, I used their vector-based gGFP_R gRNA to recruit ADAR1 to edit the transcript of the mCherry-GFP cassette toward GFP activation in the HEK-239T-W58X cell line.

At 72h post-transfection with the gGFP_R guide, I sorted 25 000 double-positive (mCherry+ and GFP+) cells, establishing the priming population of edited cells on which I focus on detecting a subsequent ADAR1-depenent DNA mutation (Figure 4.14A Stage I). As a negative control I used a gRNA with no target (gCtr, Appendix B), which showed absence of double-positive cells at 72h post-transfection like the gGFP_R did (5% double positive cells, Figure 4.14B). I propagated for 2 weeks the priming population of 25 000 (noted in Figure 4.14A as 25K) edited cells and a culture of equal number of cells from the bulk gCtr-transfected culture. As I explained in 4.2.5, propagating the culture is crucial for allowing the plasmid to be diluted out and eventually lost from the culture, as I apply no selection pressure. Because of the fact that in the preliminary experiment (see 4.2.5) I obtained heterogeneously GFP+ clones at the endpoint, I decided to repeat the sorting of the remaining double-positive cells in bulk (36 000 cells rescued from the original gGFP_2-transfected culture) in order to increase the chances of eliminating the non-mutated cells (Figure 4.14A). After 3 weeks in culture (5 weeks overall post-transfection), I single-cell sorted the double-positive cells (Figure 4.14A, Stage II) in ten 96 well-plates containing conditioned media (see 3.3.1 and 3.3.5). The cells I sorted in a single-cell layout were <0.1% relatively intensely double-positive from the overall population (Figure 4.14C, red arrow), while absent from the negative control (gCtr – 5 weeks, Figure 4.14C). The seeded 96 well-plates were further cultured for growth for 3 weeks and 70% of the wells seeded (673 out of 960) turned viable clones. I screened the grown clones for GFP+ signal with optical microscopy and 61.2% of them had mostly heterogeneously GFP+ clones, similarly to what I observed in the endpoint of the preliminary experiment presented in 4.2.5. I selected 56 clones, which showed >50% GFP+ signal, and expanded them for further growth in 24WP for 3 additional days (Figure 4.14A, Screening). 16 out of the 56 expanded clones remained highly GFP+, with one of them being 100% GFP. I further detached and isolated the cells from all the grown clones (overall 8 weeks post-transfection) and half of the clones underwent FACS analysis, while I simultaneously extracted RNA and gDNA from the rest hald of the clones. I generated cDNA and gDNA amplicons (primers *pmCherry-Fw* and *peGFP-Rv*, Appendix A) from the mCherry-GFP cassette and sequenced them. In Figure 4.13D (Stage III), I summarize histograms for the GFP signal from the FACS analysis from the endpoint clones (I - XVI). As shown in the relevant figure panel (4.14D), all clones presented heterogeneously GFP+ signal, absent from the control (gCtr), with one clone (IV) being almost entirely GFP+. As shown in Figure 4.14E, I detected 100% A-to-G base change from both the cDNA and gDNA amplicons of the entirely GFP+ clone (GFP+ 98.5%), within the on-target codon that reverts the UAG stop codon to a tryptophan-encoding codon (UGG), absent from the respective amplicons of the negative control gCtr. It should be noted that in the rest of the clones I could not detect on-target A-to-G base change from neither cDNA or gDNA amplicons from RNA and gDNA isolated from the unsorted clones. It is not unlikely that due to presence of unedited/unmutated on the UAG target GFP$^-$ cells, the base calling did not pass the limit of detection. To shed a light on this, I am currently having experiments in which I am to sort the GFP+ cells and sequence amplicons for their cassette with deep-amplicon NGS sequencing.

All in all, with this experiment I have obtained evidence that ADAR1 may mutate genomic DNA in a rate of 1 in 25 000, accounted by the number of endpoint GFP+ clones (with validated A-to-G DNA mutation on-target) to the number of edited cells compiling the priming population. For this experiment I employed the tool RESTORE, which recruits the endogenously expressed ADAR1 (Merkle et al., 2019), to induce RNA editing reported as a gain-of-function for GFP activation, on which I focused on and detected a subsequent DNA mutation by ADAR1 on target. In ongoing experiments, I am performing repeats in which I am upscaling the priming population (similarly to what I did in 4.2.5) to validate the mutation rate, while I will include an additional negative control which will be a HEK-293-W58X ADAR1-/-. The last cell line is currently being generated by the student Mr. George Samaras.

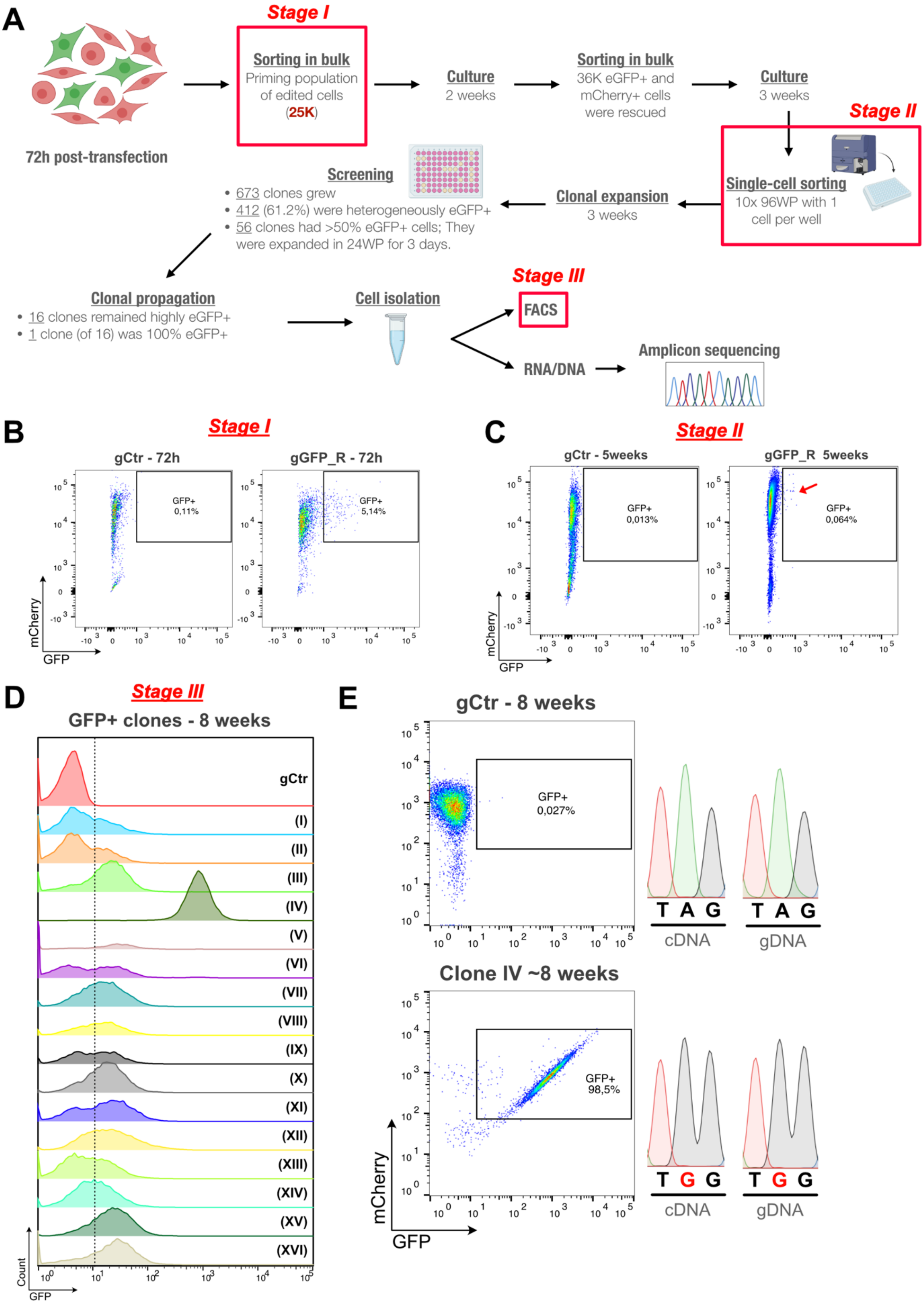


Figure 4.14 Experimental validation of ADAR1-dependent mutation through eGFP activation upon site-directed mRNA editing with the tool RESTORE. (**A**) A schematic representation of the strategy followed along with the main findings or observations. See main text of 4.2.6 for the detailed explanation. (**B**) FACS data 72h post-transfection representing the generation of the priming population (25 000 or 25K as shown in Stage I of panel A) of the double-

positive (mCherry+ and GFP+) cells upon transfection of the gGFP_R (GFP-activating gRNA, designed according to RESTORE). The rate of double-positive cells was 5.14% of the overall gated population. As a negative control gCtr guide with no target against the transcriptome was used and a double-positive population was absent. The number of events recorded in the original population was 10 000 cells. (**C**) 5 weeks post-transfection (Stage II in panel A), the remaining double-positive cells (0.064%) from the propagated population from the priming edited cells (gGFP_R) was sorted in conditioned media in 96 well-plates in a single-cell layout toward clone generation. Equal number of cells from the propagated gCtr original sample was analyzed as a negative control. The number of gated cells was 50 000. (**D**) Upon clone generation, 16 clones from the ones that successfully grew remained highly GFP+. The 16 clones (I - XVI) were upscaled in 24 well-plates for three days. 8 weeks overall post-transfection, half of each clone underwent FACS analysis (Stage III in panel A) and the other half for RNA/DNA extraction and amplicon generation. In the multi-histogram plot, the GFP+ signal is summarized per clone (I – XVI). As a negative control, a gCtr fraction from the original negative control was included. The number of overall recorded events per clone was 10 000, besides clone V for which 3 000 cells were available (**E**) The clone IV from the endpoint (Stage III in panel A, also shown in panel D), showed an almost entirely (98.5%) double-positive signal, a phenotype which is explained by a DNA mutation (shown from both cDNA and gDNA amplicons of the mCherry-GFP cassette, primers in Appendix A) which is 100% on-target and has reverted the stop codon UAG (shown as TAG) to UGG (shown as TGG). No double-positive population or on-target A-to-G base change was detected in the negative control (gCtr). The number of recorded events was 10 000 for both the clone IV and the gCtr. Gating strategies followed as in Appendix E - part 4. This observation validates an ADAR1-dependent DNA mutation in a rate of 1 in 25 000. For this experiment I employed a vector, which Dr. Riccardo Pecori and Ms. Beatrice Casati designed, to efficiently express the gGFP_R.

4.2.7 Discussion

As previously discussed throughout this dissertation, A-to-I RNA editing is a widespread modification in the human transcriptome, found in transcript topologies with repetitive sequences that form dsRNA, such as *Alu* elements (Athanasiadis et al., 2004). Both ADAR1 and ADAR2 have demonstrated deamination activity, but the major driver of A-to-I editing in the transcriptome is ADAR1, which is ubiquitously expressed across tissues in two isoforms ADAR1-p150 and ADAR1-p110 (Zinshteyn and Nishikura, 2009). ADAR1-p110 is constitutively expressed and resides in the nucleus, while ADAR1-p150 expression is induced by interferon (IFN) through an IFN inducible promoter, found in a Z-DNA binding domain, which is absent from the ADAR1-p110 isoform (Lamers et al., 2019). Although ADAR1-p150 can be temporarily in the nucleus, it is mostly cytoplasmic due to a nuclear export signal (NES), which is also present in the Z-DNA binding domain (Poulsen et al., 2001). ADAR1 functions, primarily attributed to the isoform ADAR1-p150, have been primarily associated with preventing cellular response to endogenous self-dsRNAs as non-self through A-to-I RNA editing (Liddicoat et al., 2015, 2016). Furthermore, ADAR1 is found overexpressed in the vast majority of tumors (Han et al., 2015), leading to elevated A-to-I RNA editing activity and enhancing transcriptomic heterogeneity in tumors (Paz-Yaacov et al., 2015), often associated with poor prognosis as discussed, for instance, for multiple myeloma in chapter 4.1 and also shown by Lazzari et al., 2017; Tasakis et al., 2020; Teoh et al., 2018.

More recently, it has been shown that ADAR1 can deaminate DNA within DNA/RNA hybrids (R-loops) *in vitro* (Zheng et al., 2017), which is shown to be a crucial function of ADAR1-p110 ensuring telomeric stability in cancer cells (Shiromoto et al., 2021). R-loops, however, are formed genome-wide (Yan et al., 2019) and in chapter 4.1, I presented data correlating RNA editing and DNA mutation by ADAR1 with the context of R-loops in Multiple Myeloma. I hypothesized that DNA mutation by ADAR1 may not necessarily be the primary function of the enzyme, but rather a collateral damage of elevated RNA editing activity, which may be a genome-wide phenomenon in tumors. In this chapter, I have presented a series of experiments as proof of this concept. First, I leveraged the power of the V_H immunoglobulin gene of Ramos AID-/- cells and showed that it can function as a reporter for DNA mutation through loss of IgM (chapter 4.2.2). Then, I ensured that ADAR1 is the major A-to-I deaminase expressed in Ramos B cells and therefore in the AID-/- cells of the same cell line (chapter 4.2.3). To report the potential mutagenic role of ADAR1 in Ramos AID-/-, I recruited the endogenous enzyme with the site-directed mRNA editing tool LEAPER (Qu et al., 2019) with multiple pairwise combinations of gRNAs, so as to induce loss of IgM (chapter 4.2.4). Loss of IgM was observed in Ramos AID-/- cells, the V_H genomic locus of which was targeted with a gRNA targeting the coding positive-sense strand (and therefore the transcript) and a gRNA about 200bp downstream targeting the template negative-sense strand (Figure 4.8C). It has been recently discovered, that the footprint of an ADAR dimer for editing dsRNA is 50bp *in cis* (Song et al., 2020), which would translate to 170Å distance along the DNA strand. However, it is not entirely impossible that similar distance can be achieved by an ADAR dimer *in trans* in a genomic locus, such as the V_H region. Provided that the experiments presented here get validated (such as with the reversion of the stop in the IgM$^-$ Ramos AID-/- cell line, suggested in 4.2.4), it is worth engineering new gRNAs which target the V_H in different parts of the genomic region, so as to explore the optimal distance required between the gRNAs for ADAR1 to mutate gDNA.

Despite the encouraging observations of IgM loss upon targeting the V_H region with gRNAs, no specific mutations or double-strand breaks in the V_H could be associated with this phenotype nor RNA editing was detected in the early stages of the experiment. I observed an overall drop in IgM expression and, therefore, I tested potential side-effects of the gRNAs. I delivered gRNAs directly upon *in vitro* transcription and I observed abundant loss of IgM through 14 days post-transfection, which at 21 days IgM populations were regained (Figure 4.10). No RNA editing or DNA mutation data supported the observed loss of IgM. This may be due to two reasons: 1) the gRNAs targeting the positive-sense coding strand and, therefore the transcript, may function as small-interfering RNAs (siRNAs) and 2) for the gRNAs targeting the negative-sense template strand this may be due to resolution of transient R-loops between the gRNA and the template strand by RNAseH, as reported by Shiromoto et al., 2021.

These limitations encouraged me to reconsider the system for reporting that RNA editing may lead to DNA mutation. First, it is crucial to eliminate the background of cells in which RNA editing did not occur and then propagate the specifically RNA-edited cell population to look for DNA mutations. As discussed in chapter 4.2.5, I employed a HEK cell line, previously described in Montiel-Gonzalez et al., 2013, which expresses a cassette of mCherry and an in inactivated eGFP due to a premature UAG stop codon (Figure 4.11A). I employed the λN-ADAR tool, developed by the same group to edit the eGFP transcript, which lead to U**A**G>U**G**G editing with a specific gRNA (gGFP) and, therefore, produced a fluorescent eGFP protein (Figure 4.11B). I sorted eGFP-activated cells after 72h post-transfection, the timepoint that showed the highest level of editing (33%) on the transcript. 25 000 eGFP-activated cells were the priming population, which I focused on to look for DNA mutation (Figure 4.12A). I propagated this population for further growth for 2 weeks, allowing the loss of the plasmids with the editing components (4λN-ADAR-and gGFP). The very few cells that remained eGFP-activated were sorted in single cells (Figure 4.12B) and 2 weeks after I screened the successfully grown clones for eGFP signal. 1 clone showed heterogeneously eGFP-activation and was expanded for an additional week. I generated and sequenced cassette amplicons from cDNA and gDNA of the eGFP-activated cells from the clone sorted (Figure 4.12C). I detected 18% of A-to-G base change in cDNA amplicons, but not from gDNA amplicons 5 weeks post-transfection. Although base change is reported from the cDNA, this observation may still be a DNA mutation. Throughout my FACS analyses, the HEK cells expressing the mCherry and eGFP cassette, showed diversity in mCherry fluorescence and the eGFP-activated cells were mostly at high mCherry fluorescence. This means that the cassette is likely incorporated in the genome of HEK cells in definitely more than one copies, which may be the reason why this mutation does not pass the limit of detection from gDNA amplicon chromatograms. If this is true, my findings are the first indication for the mutation rate by ADAR in a cell population, occurring at 1 in 25 000, accounted by the number of eGFP-activated clones in the number of the priming GFP-activated cell population. To validate the mutation rate and hopefully detect A-to-G mutations from the gDNA amplicons of the mCherry/eGFP cassette, I upscaled the priming population of eGFP-activated cells due to RNA editing to 350 000. Following the same strategy as the original experiment (but increasing the incubation time for 2 more weeks prior to single-cell sorting), I obtained 7 clones at the endpoint being heterogeneously eGFP+ (Figure 4.13A), similar to what I observed before, which accounts the mutation rate for this experiment at 1 in 50 000. However, I detected again A-to-G base change (~23%) from the cDNA and not gDNA amplicons of the cassette (Figure 4.13B). I therefore wondered whether my observations thus far (particularly not being able to detect A-to-G base change from the gDNA amplicons), has something to do with the fact that the λN-ADAR tool is employing an engineered deaminase domain and not an endogenously expressed ADAR.

To address whether ADAR1-dependent RNA editing would lead to DNA mutation by the endogenous enzyme, I employed RESTORE, a site directed mRNA editing tool which recruits the endogenously expressed ADAR1 with a gRNA containing a GluR2 loop (Merkle et al., 2019 and chapter 1.3.2). My colleagues, Dr. Riccardo Pecori and Beatrice Casati, previously designed an efficient plasmid vector (see 4.2.6, gGFP_R) which activates eGFP in the HEK cell line expressing the mCherry/eGFP cassette I have used thus far. I transfected this cell line with the aforementioned gGFP_R plasmid vector, along with the appropriate controls, and I repeated the experiment following an optimized experimental pipeline to properly eliminate the background of unedited cells (summarized in Figure 4.14A and explained in 4.2.6). 72h post-transfection I generated 25 000 eGFP-activated cells through RNA editing as the priming population, on which I later focus to detect DNA mutation (Figure 4.14B). After 2 weeks in culture I sorted the remaining double-positive (eGFP and mCherry) cells in bulk to eliminate the background of unedited cells and propagated them for further growth for three weeks with no selection pressure for keeping the gGFP_R plasmid. 5 weeks post-transfection, I performed a single-cell sorting experiment (Figure 4.14C) for the remaining small double-positive population (<0.1%). 8 weeks overall post-transfection about 70% of the wells seeded with single cells successfully grew, of which ~61% were heterogeneously eGFP+ after screening with optical microscopy. 56 clones were highly eGFP+ (>50% of the clone), which I further expanded in 24 well-plates for 3 days. 16 clones from the propagated ones remained eGFP+, with one being purely eGFP+ (Clone IV; FACS data shown in Figure 4.14D). I isolated RNA and gDNA from the bulk of the clones (not sorted for double-positive) and I generated cDNA and gDNA amplicons for the mCherry/eGFP cassette. As I show in the figure 4.14E, I detected 100% A-to-G base change from the gDNA amplicon of the clone IV within the on-target UAG codon, which is reverted to UGG through mutation and explains the phenotype, absent from the negative control. This experiment validates for the first time a likely ADAR1-dependent DNA mutation with a rate of 1 in 25 000 (rate counted as 1 endpoint clone arising from the overall number of priming edited cells). In ongoing experiments, I am upscaling the priming population of edited cells to validate the mutation rate of ADAR1, while I will include an additional negative control, which is an ADAR1-/- HEK cell line containing the mCherry/eGFP cassette.

To conclude, the experiments I presented thus far, have clarified a number of points with regards to the appropriate strategies that should be employed to detect DNA mutations by ADARs, while I reported that ADAR1-dependent DNA mutation may occur in a rate of 1 in 25 000. These experiments are a proof of concept to support the correlative data I drew between RNA editing and DNA mutation from MM patiens (see 4.1) and ultimately validate my original hypothesis, according to which ADAR1 may mutate genomes as a collateral damage of its original editing activity (Figure 4.1). Finally, my experimental findings also provide a better insight for genome editing tools incorporating ADARs, which are currently in development.

4.3 Aim 3: RNA deaminases drive SARS-CoV-2 genome evolution

4.3.1 Preface

As introduced in the chapter 1.2.4, the ssRNA(+) genome of SARS-CoV-2 or its dsRNA intermediates is targeted by deaminases from both the ADAR and APOBEC families, with C-to-U and A-to-G changes compiling about 65% of the documented mutations thus far (Giorgio et al., 2020; Klimczak et al., 2020; Poulain et al., 2020; Wang et al., 2020). In particular, it has been shown that enzymes from the APOBEC3 subfamily are likely driving the C-to-U mutagenesis in SARS-CoV-2 (Poulain et al., 2020). As per the A-to-G mutagenesis, the cytoplasmic isoform ADAR1-p150 is the one generally editing the double-stranded intermediates of RNA viruses replicating into the cytoplasm (Lamers et al., 2019), which appears to be the case for other RNA viruses as well, such as HIV-1 (Doria et al., 2009). It is therefore clear that RNA deaminases are mutating the SARS-CoV-2 genome, as well as how they do that mechanistically. However, what remains unclear is how the variation due to RNA deaminases is shaping the evolution of the viral genome. And in the case of SARS-CoV-2, the virus responsible for the ongoing COVID-19 pandemic (Cucinotta and Vanelli, 2020), this is crucial to know as several variants of concern (VOCs) appear to have been evolved worldwide, being able to bypass the population immunity and further trigger and prolong the pandemic.

Here, I hypothesize that RNA deaminases are actually responsible for the variant evolution of SARS-CoV-2 in the following way: as infection from one individual to another occurs, an RNA deaminase may mutate the viral genome introducing a founding mutation, which if selected, it will pass to the next individual infected; and when inter-individual infection continues, more mutations will be selected and co-exist in the genome. From an evolutionary perspective, this is similar to how tumors also select and build their mutational loads, only from selection and expansion of mutations from one cell to the other within the entity; a phenomenon called clonal evolution (Greaves and Maley, 2012). A clear example is Multiple Myeloma, the focus of my first aim in the present dissertation (see 4.1), in which clonal evolution upon a strict selection of myeloma plasma cells is key in disease progression (Lagana et al., 2017). Therefore, here I am leveraging the power of publicly available SARS-CoV-2 sequence data from the United States (see 3.4.1) to address my bigger question from a fast-track evolution perspective, which is no other than exploring whether it is in the destiny of RNA deaminases to damage genomes, by introducing mutations opportunistically, which upon selection may be damaging.

4.3.2 SARS-CoV-2 genome gradually accumulates specific mutations over time.

I obtained overall 62 211 SARS-CoV-2 genome wide sequences, from which I called single-nucleotide variations (SNVs), comparing them against the reference sequence of SARS-CoV-2 (see 3.4.1), isolated from the first individual infected in Wuhan China in December 2019 (Wu et al., 2020). The sequence cohort I employed represents infections in the United States from the first 15 months of the COVID-19 pandemic (see 3.4.1). First, I calculated the number of SNVs per sequence and, as shown in Figure 4.15A, the median number of SNVs per sequence (also termed here as viral isolate) and per date is gradually increasing with time, indicating that SARS-CoV-2 is indeed gradually accumulating mutations as infection rates continue to rise. To obtain the intrinsic pattern of mutations, I asked how many kinds of different SNVs were found in the viral genome, counting an SNV from each genomic coordinate once. I found (Figure 4.15B), that C-to-U (shown as C>T) was the predominant SNV, followed by G-to-U (G>T), A-to-G (A>G) and U-to-C (T>C) at very similar levels, as well as G-to-A (G>A), as other studies also found (Giorgio et al., 2020; Klimczak et al., 2020). C>T mutations are attributed to APOBEC3 activity (Poulain et al., 2020), while A>G and T>C to ADAR1 activity, with the latter to have raised as an A>G change in the negative-sense strand of the dsRNA viral intermediate as suggested by Giorgio et al., 2020, similarly to what I described for Figure 4.1. I then tallied robustly the different kinds of SNVs separately for each Open Reading Frame (ORF) of the SARS-CoV-2 genome (Figure 4.15C), and I observed that different SNVs were enriched for different ORFs. For instance, C>T SNVs were enriched over others in ORF1a, 1b, 5(M), 8 and 9(N), A>G in ORF2 encoding for the Spike protein, and G>T in ORF3a. As shown in the Figure 4.14C, it is not surprising that longer ORFs are aggregating more mutations, however the enrichment of specific kinds of SNVs in individual ORFs indicates the aggregation of specific mutations.

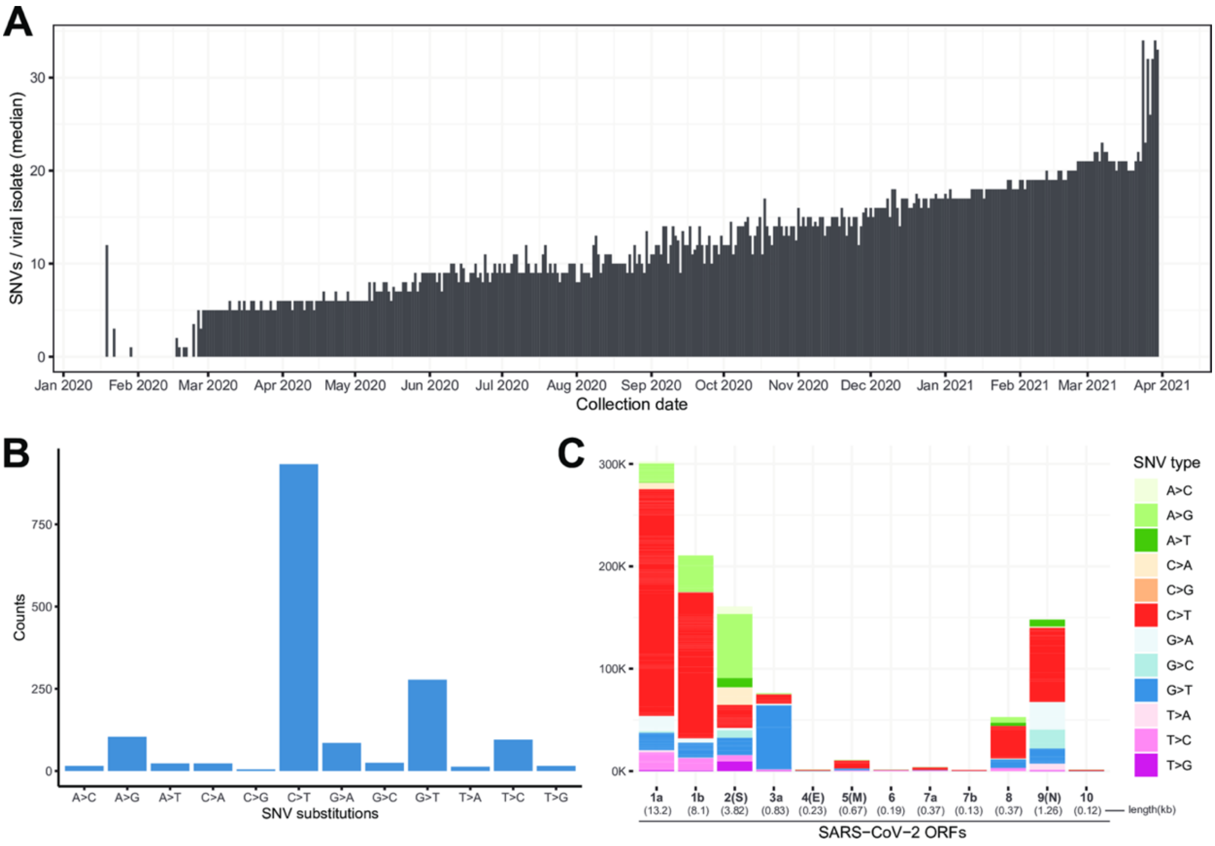


Figure 4.15 The genome of SARS-CoV-2 gradually accumulates specific kinds of SNVs over time. (**A**) Histogram demonstrating that the median of SNVs per sequence (viral isolate; y axis) is steadily increasing over time (Collection Date in x axis). Collection dates with at least 5 sequences per day were considered. (**B**) Bar plot showing the intrinsic pattern of mutations in the SARS-CoV-2 genome. SNV substitutions (x axis) were counted throughout the viral genome once per genomic coordinate. (**C**) Different Open Reading Frames (ORFs; x axis) are showing differential mutational

bias for example for C>T (ORF 1a, 1b, 5(M), 8, 9), A>G (ORF2(S)), likely through selection of specific mutations. The lengths of the different ORFs are given in parentheses below their labels across the x axis in kb. Figure from Tasakis et al., 2021. Reused under the Creative Commons License 4.0. The illustration was produced by myself, for the mutation calls I received help for a part of the raw data generation as explained in 3.4.2.

My findings thus far, underline that SARS-CoV-2 is not keeping a stable genome, but instead it gradually accumulates different mutations. As shown in Figure 4.15C, the distributions of different SNVs tallied robustly per ORF were different than the intrinsic pattern of mutation of the overall viral genome (Figure 4.15B), which indicated that the enrichment of specific SNVs in different ORFs is due to the recurrence of specific mutations. Indeed, Dr. Marilyn Diaz and Prof. Laurent Verkoczy observed that the ratio of missense to silent mutations per SNV (for A>G, T>C, C>T and G>A) was >1, indicating positive selection for specific mutations (Tasakis et al., 2021). Therefore, I explored the entire viral genome for specific mutations and calculated their frequency within the cohort of sequences. First, I focused only in the cohort of sequences from the first 12 months of the pandemic in the United States (calendar year 2020) and I found 14 specific missense mutations present in at least 10% of the sequences, summarized in Table 4.1 and shown in Figure 4.16A.

Table 4.1 The 14 predominant mutations found in the SARS-CoV-2 genome in 2020. This table summarizes information for the predominant mutations found in at least 10% of the genomes in 2020, and in particular: their SNV change to the nucleotide level with the respective position in the genome, the ORF they are found in, the subsequent amino acid change and the associated protein function, along with the frequency (%) in the cohort. The mutations shown are ranked by Frequency and are also shown in Figure 4.16A. Table from Tasakis et al., 2021 and is reused under the Creative Commons License 4.0. This table was produced by myself. I received help for generating mutation calls from a fraction of the raw data as explained in 3.4.2.

Change (Nucleotide)	ORF	Change (Amino Acid)	Protein Function	%Frequency
C14408T	1b	P323L	RNA-dependent RNA polymerase	82.03%
A23403G	2 (S)	D614G	Spike protein; between the RBD and S2 domains	80.76%
G25563T	3a	Q57H	APA3 viroporin – accessory protein	57.62%
C1059T	1a	T85I	Nsp2	48.79%
C27964T	8	S24L	Ig-like protein	15.22%
T28144C	8	L84S	Ig-like protein	14.07%
C10319T	1a	L89F	Peptidase C30	12.98%
A17858G	1b	Y541C	DNA/RNA helicase domain	12.34%
C17747T	1b	P504L	DNA/RNA helicase domain	12.19%
A18424G	1b	N129D	Nsp14; 3’-5’ exonuclease	11.08%
C21304T	1b	R216C	Nsp16	10.93%
C28472T	9	P67S	Nucleocapsid	10.78%
G25907T	3a	G172V	Viroporin	10.76%
C28869T	9	P199L	Nucleocapsid	10.53%

The predominant mutations I found in SARS-CoV-2 in 2020 (>10% of the genomes; shown in Figure 4.16A and Table 4.1), primarily because of C>U(T) or A>G nucleotide changes pinpointing to APOBEC and ADAR activity respectively, had frequencies that varied. Some of them were significantly more abundant than others, such as P323L or D614G present in more than 80% of the genomes, which drove me to consider that some of them (ie D614G or P323L) perhaps appeared earlier than others and were selected. Therefore, I tested their densities within the cohort over time and I found that they follow 4 distinct patterns (Figure 4.16B). Mutations of pattern A (Figure 4.16B), appeared later in 2020, while mutations of patterns B or D appeared generally from the beginning and remained in the cohort with variation of abundance throughout the year and mutations of pattern C were present abundantly in the beginning and later on disappeared. Here, I give an overview of the mutations grouped by their patterns:

A. As mentioned above, pattern A is a group of mutations that appeared later in the year at the cohort I investigated. Therefore, it is not surprising that they are not within the top ones in the list of predominant mutations for 2020. The most frequent of the group (12.98%) and the one that appeared the earliest was L89F (Leucine to Phenylalanine), recoding the peptidase C30 encoded by ORF1a, also reported by (Wang et al., 2020a). Two additional mutations were the N129D (Asparagine to Aspartic-acid at 11.08%) and R216C (Arginine to Cysteine at 10.93%), which co-appeared about the same time, recoding the Nsp14 (3'-5' exonuclease) and Nsp16 respectively from ORF1b as also reported by (Pater et al., 2021). Three more mutations, the least frequent, from the group were G172V (Glycine to Valine at 10.76%; also found in Hassan et al., 2020) recoding the viroporin expressed by ORF3a, as well as P67S (Proline to Serine at 10.78%) and P199L (Proline to Leucine at 10.53%) recoding residues from the viral nucleocapsid encoded by ORF9(N), and found in (Pater et al., 2021).

B. This pattern groups the most abundant mutations in the cohort. The D614G (Aspartic acid to Glycine at 80.76%) recodes residues from the Spike protein (ORF2) and it has been extensively discussed in the literature, as it was the first mutation to be associated with increased viral infectivity (Hou et al., 2020). The P323L (Proline to Leucine at 82.03%; also reported by Toyoshima et al., 2020) recodes the viral RNA-dependent RNA polymerase encoded by ORF1b and is the first mutation detected in the cohort I investigated and the most abundant. Two additional mutations in the group are T85I (Threonine to Isoleucine at 48.79%; Laha et al., 2020) recoding the non-structural protein 2 (Nsp2) encoded from ORF1a and the mutation Q57H (Glutamine to Histidine at 57.62%; Hassan et al., 2020) recoding the viroporin expressed by ORF3a.

C. The mutations that follow the pattern C appeared early in the cohort of 2020, but disappeared halfway the year. Two mutations were recoding the DNA/RNA helicase domain by ORF1b were P504L (Proline to Leucine at 12.19%) and Y541C (Tyrosine to Cysteine 12.34%). Furthermore, an L84S (Leucine to Serine at 14.07%) mutation was recoding the Ig-like protein encoded by ORF8. They were previously also found in Wang et al., 2021.

D. The last pattern I found, contains only one mutation the S24L (Serine to Leucine at 15.22%; Wang et al., 2021) which also recodes the Ig-like protein. This mutation presents a very interesting pattern, according to which it “surfs” at small-to-medium range frequencies overtime since the beginning of the 2020.

My findings thus far show that the different predominant mutations in 2020 as described above, can potentially co-occur in groups within the cohort of sequences I investigated. I found 4 distinct patterns, three of which group at least three mutations (A-C), while the fourth (D) could tag along periodically with either of the three above. Therefore, my results thus far indicate that in 2020 there were at least 3 different major versions (variants) of the SARS-CoV-2 genome in the United States.

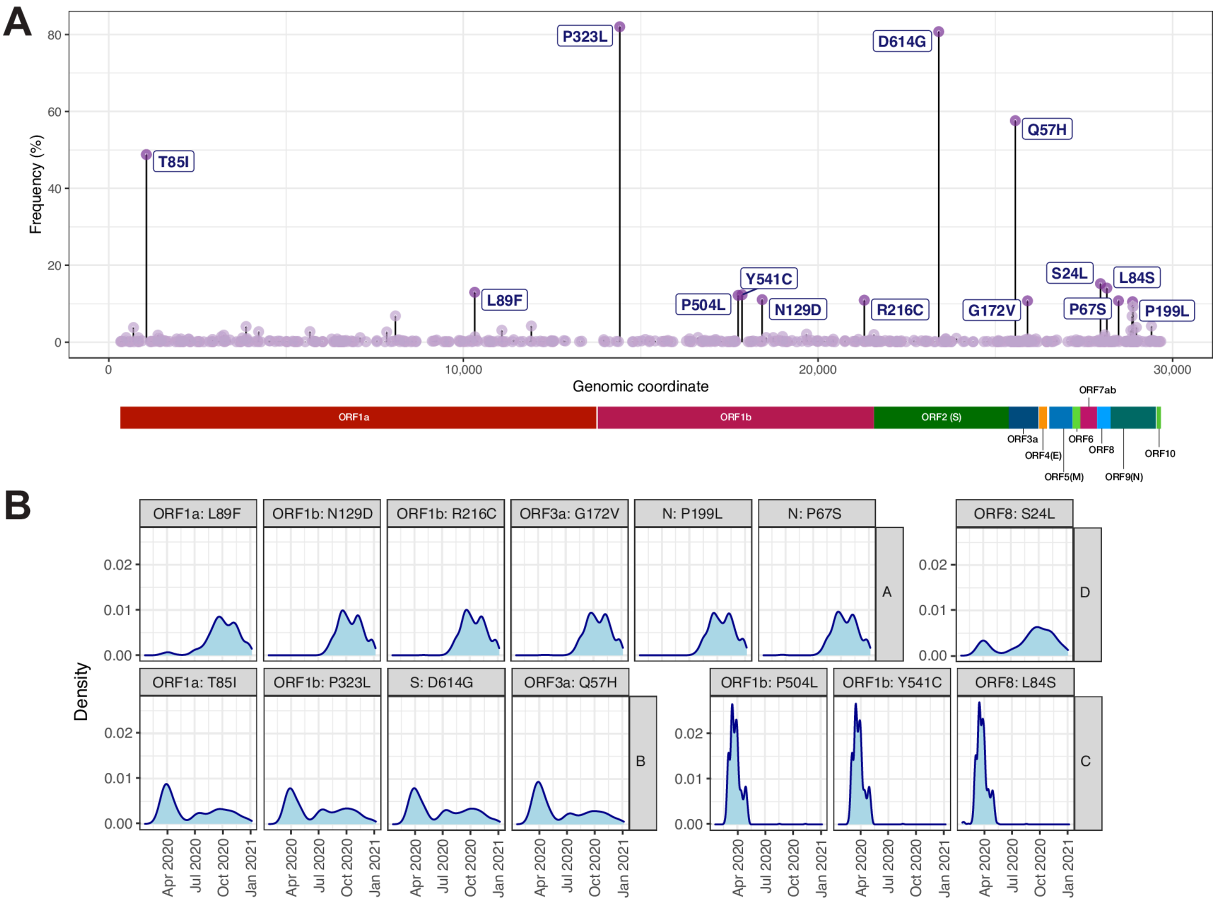


Figure 4.16 Selected mutations in the SARS-CoV-2 genome and their dynamic shift of abundance throughout the 2020. (**A**) Frequency plot (% frequency in y axis) of the missense mutations detected in the cohort across the SARS-CoV-2 genome (x axis), organized in distinct ORFs. The predominant missense mutations detected in 2020 in SARS-CoV-2 are highlighted with dark-purple and labelled according to their amino acid change and protein position. A detailed description of these mutations is given in Table 4.1, as well as in the main text. (**B**) Density plots showing the abundance throughout the year 2020 for each of the predominant mutations highlighted in panel A. These mutations are grouped by four different patterns (A-D) with regards to their abundance in the cohort over time, indicating that mutations may co-occur in at least three different variant versions of the SARS-CoV-2 genomes. Adapted figure from Tasakis et al., 2021 under the Creative Commons License 4.0. Data and illustration were produced by myself. I received help in generating mutation calls for a fraction of the raw data as explained in 3.4.2.

4.3.3 Signatures of co-occurring mutations can explain the SARS-CoV-2 genome evolution

As I showed above, the different mutations predominantly found in the SARS-CoV-2 genome can be grouped within different patterns of presence through time within the cohort of sequences I investigated for 2020. In particular, I detected 3 main patterns (A-C) of mutations, according to which some mutations were present throughout the cohort, some appeared later and some appeared early but disappeared early in 2020. I, therefore, hypothesized that mutations may co-exist in the circulating viral genomes and if that is true, they can provide an in-depth resolution to study the genome evolution of SARS-CoV-2 toward variant establishment in the population. To address that, I employed an alternative approach, which is unbiased to previous nomenclatures proposed to describe the variation of SARS-CoV-2 (i.e. PANGO, as in suggested by Rambaut et al., 2020). My approach is inspired by the COSMIC signatures, which in cancer biology is employed to profile the mutational spectra of tumors (Alexandrov et al., 2020, 2013). First, I profiled every genome in the cohort to obtain information of whether the relevant sequence had either of the predominant mutations. Relying on that, I compiled a signatures "thesaurus", which contains all the putative signatures of the co-existing predominant mutations detected in my dataset. This is demonstrated in the heatmap of Figure 4.17, in which a signature (columns) is defined by whether the predominant mutations (summarized in Table 4.1 and rows of the heatmap) where found (dark blue) or not (light yellow). I overall detected 48 distinct signatures (s1-48), along with the signature s0, which stands for the viral genomes that do not contain any of the predominant mutations in 2020. It should be noted, that the signatures were originally compiled with 8 000 SARS-CoV-2 genomes (see 3.4.2) employed for the analysis due to data availability. When more data from 2020 became available and I included them, there was no change for the predominant mutations found in 2020, nor a proportional difference of each signature to the overall cohort.

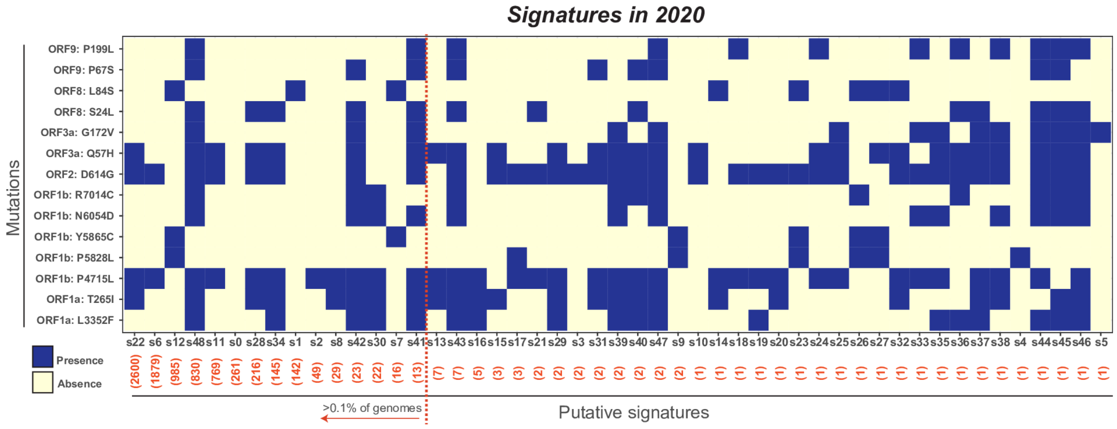


Figure 4.17 Putative signatures of co-occurring mutations in 2020. Unique combinations of the 14 predominant mutations detected in 2020 (rows) are compiling 48 distinct putative signatures (columns; s1-48) detected in the cohort of SARS-CoV-2 genomes. s0 is the non-variant genome as per this analysis and is also included in the heatmap. The number of genomes profiled with each signature is given in the red labels across the x axis. Signatures are ranked by their abundance and the ones present in more than 0.1% of genomes were further considered for the downstream analyses. The presence or absence of a mutation in a given signature is colored with blue or yellow respectively. Figure from Tasakis et al., 2021 under the Creative Commons License 4.0. The data and illustration were produced by myself.

For the downstream analysis, I focused on the signatures that were found in 0.1% or more of the cohort of genomes in 2020, representing the top 15 most abundant signatures I found in 2020, including the non-variant s0. I obtained the first sequence found in the cohort from each of the top 15 signatures, along with the reference sequence for SARS-CoV-2 isolated from the patient-zero in Wuhan, China (Wuhan-Hu-1) and performed a real-time phylogenetic analysis (see 3.4.3). In Figure 4.18, I present the real-time phylogenetic tree that derived from the analysis, rooted to the Wuhan-Hu-1 (purple dot), noting the first (light blue) and last (red) sequence detected (by date) per signature in the cohort, along with the mutations acquired (red) or lost (grey) in the different clades. I found that the s0 was one of the first detected, along with signatures s1, s12 and s17, containing the mutations of pattern C (ORF1b: P504L, ORF1b: Y541C and ORF8: L84S). All those signatures were not detected ever since late May 2020. Instead, they were replaced by other signatures, which gradually accumulated mutations. One of the first acquired were the ORF1b: P323L and ORF2(S): D614G, leading to s6, followed by ORF3a: Q57H and ORF1a: T85I, leading to s22, followed by ORF8: S24L and ORF1a: L89F toward s34. A mutational burst of ORF1b: N129D, R216C, 3a: G172V, 9(N): P67S and P199L has led to one of the most predominant signatures s48. Throughout this process a number of other related signatures appeared, but the most prevalent (also in abundance; see Figure 4.17) toward the end of 2020, was the signatures s6, s22 and s48. These three signatures likely represent three major variants circulating the United States as I predicted above relying on the patterns A-D.

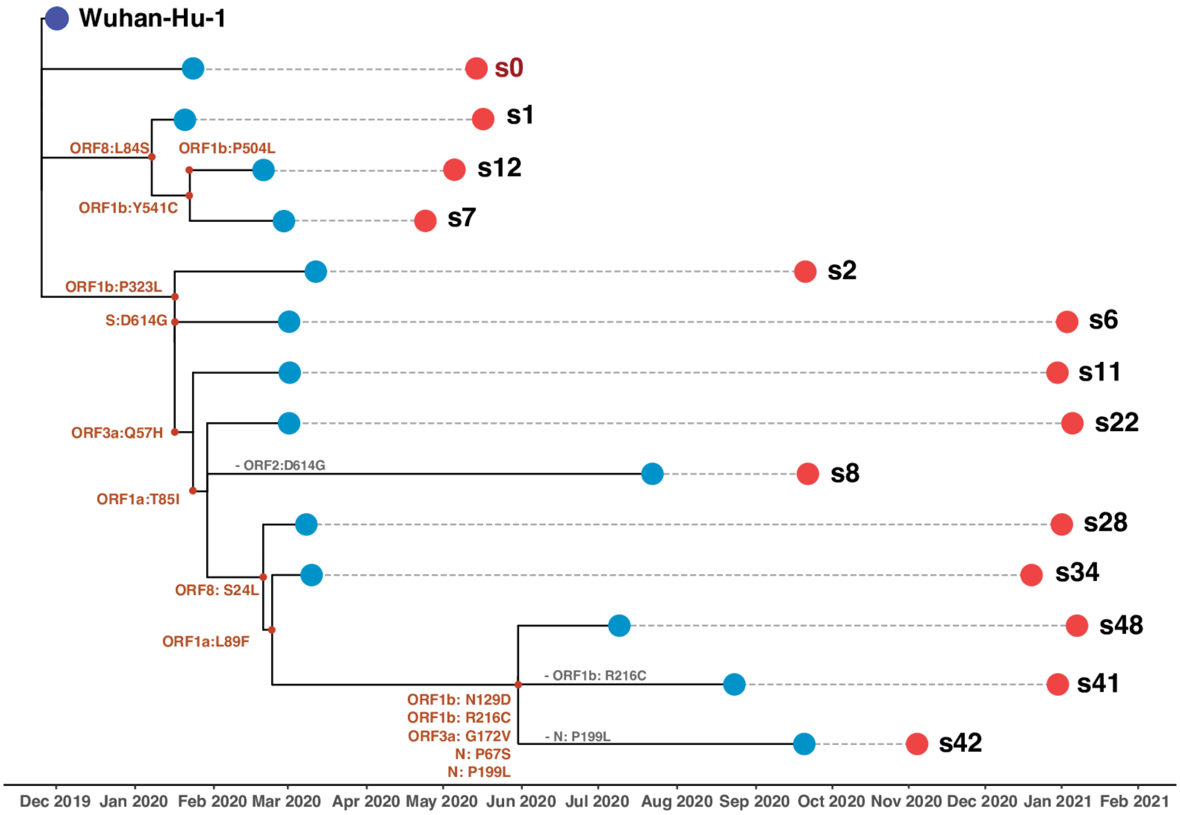


Figure 4.18 Time-scaled phylogenetic analysis relying on the signatures detected in 2020. 14 signatures of co-occurring predominant mutations in 2020 and the non-variant (s0) as per this analysis, were found in more than 0.1% of the SARS-CoV-2 genomes. The first sequences found in 2020 from each signature were employed for a time-scaled phylogenetic analysis (see 3.4.3), which revealed a gradual and processive acquisition of mutations (red labels in the

tree branches), with occasional losses (grey labels) toward more complex signatures over time. The original isolate of SARS-CoV-2 (reference genome, see 3.4.1), noted as "Wuhan-Hu-1" (purple dot) in the phylogenetic tree was the root for this analysis. The first and the last viral genomes detected from each signature in 2020 are annotated on the tree with blue and red dots respectively. The non-variant s0 (labelled in red) and the less complex signatures s1, s12 and s7 defined by mutations from pattern C (see Figure 4.16B), disappeared early on by June 2020. Most of the viral genomes thereafter were profiled with signatures of increased complexity, which all contain the S:D614G (besides the shortly present s8) and ORF1b:P323L. The first group that appeared was the one of signatures s2, s6, s11, s22, compiled by mutations of pattern B. Downstream, the viral genomes further acquired mutations of patterns A and D (see Figure 4.16B), first revealing the signatures s28 and s34, and a burst of additional mutations from pattern A, likely originating from June 2020, formed the signatures s48, s41 and s42. By the end of the year 2020 or very early 2021 the predominant signatures were overall s6, s22 and s48. Adapted figure from Tasakis et al., 2021 and under the Creative Commons License 4.0. The data and illustration were produced by myself.

4.3.4 Appearance of Variants of Concern is shifting the SARS-CoV-2 mutational profile from 2020 to an entirely different in 2021.

Thus far, I have shown that the SARS-CoV-2 genome aggregates mutations progressively, most of which are attributable to ADAR and APOBEC activity as introduced in 1.2.4 and 4.3.1. As host-to-host infections continue to occur, mutations are being generated, selected and further expand setting a different genomic profile for SARS-CoV-2 that is rapidly changing. In fact, the non-variant genomic profile (Wuhan-Hu-1 or s0 in Figure 4.17) is not circulating since late May 2020. My data have shown that by the end of 2020 there are three major variants (s6, s22 and s48) that were circulating the United States. These variants have been generated through gradual accumulation of selected mutations within the population. However, since late 2020 a number of variants were already reported, which were noted as variants of concern (VOCs) because of accumulated mutations in the spike protein, a key component in SARS-CoV-2 infection (see 1.2.4; Rambaut et al., 2020). The presence of such variants, being evolutionarily fit to potentially bypass the population immunity and out-crowd other less infectious variants from the population (Darby and Hiscox, 2021), is another parameter that I consider in my analysis. Here, I apply the signature-based methods I have described thus far from 2020 to mid-2021, but I also consider that VOCs may be masking the mutations of the homegrown variants I explored thus far.

I first focused on the ORF2 genomic region of SARS-CoV-2, which encodes for the spike protein, which is crucial for the viral infectivity and mutations in the spike raise concern for generation of more infectious variants (Darby and Hiscox, 2021). The only mutation as predominant in the spike protein thus far is the D614G (Table 4.1). However, I found a number of mutations in very low frequencies. In Figure 4.19A, I present the low frequency spike mutations (LFSM; >0.1% of the genomes) per quarter of the year 2020 (Q1-Q4), highlighting the different subunits of the spike: RBD (Receptor-binding domain), FCS (Furin Cleavage site), FP (Fusion Peptide), HR1 (Heptad Repeat 1) and HR2. Although in low frequencies, the abundance of mutations increases dramatically from Q1 to Q4, especially in the RBD which is binding to the host's receptor ACE2 for initiating the infection (see 1.2.4 and Yi et al., 2020). Interestingly, I also observed that near the FCS there is a hypermutable region

(675-681), which even has a single residue Q677 being mutated in multiple ways (Q677H by different nucleotide changes or Q677R). This LFSM is a dynamic pool providing the grounds for selection and, indeed, a few of those mutations rose to prominence in 2021 (Figure 4.19B). I performed the same analysis for 2021 (as I did for 2020), in order to report the predominant mutations of 2021. In Figure 4.19B, I summarize the mutations found in >10% of the genomes in 2021 alone and I highlight with red the mutations that were novel for 2021. Additionally, I noted with an asterisk (*) the mutations which I previously observed as LFSM and are carried in the cohort because of VOCs that rose to prominence, as I present in the next paragraph.

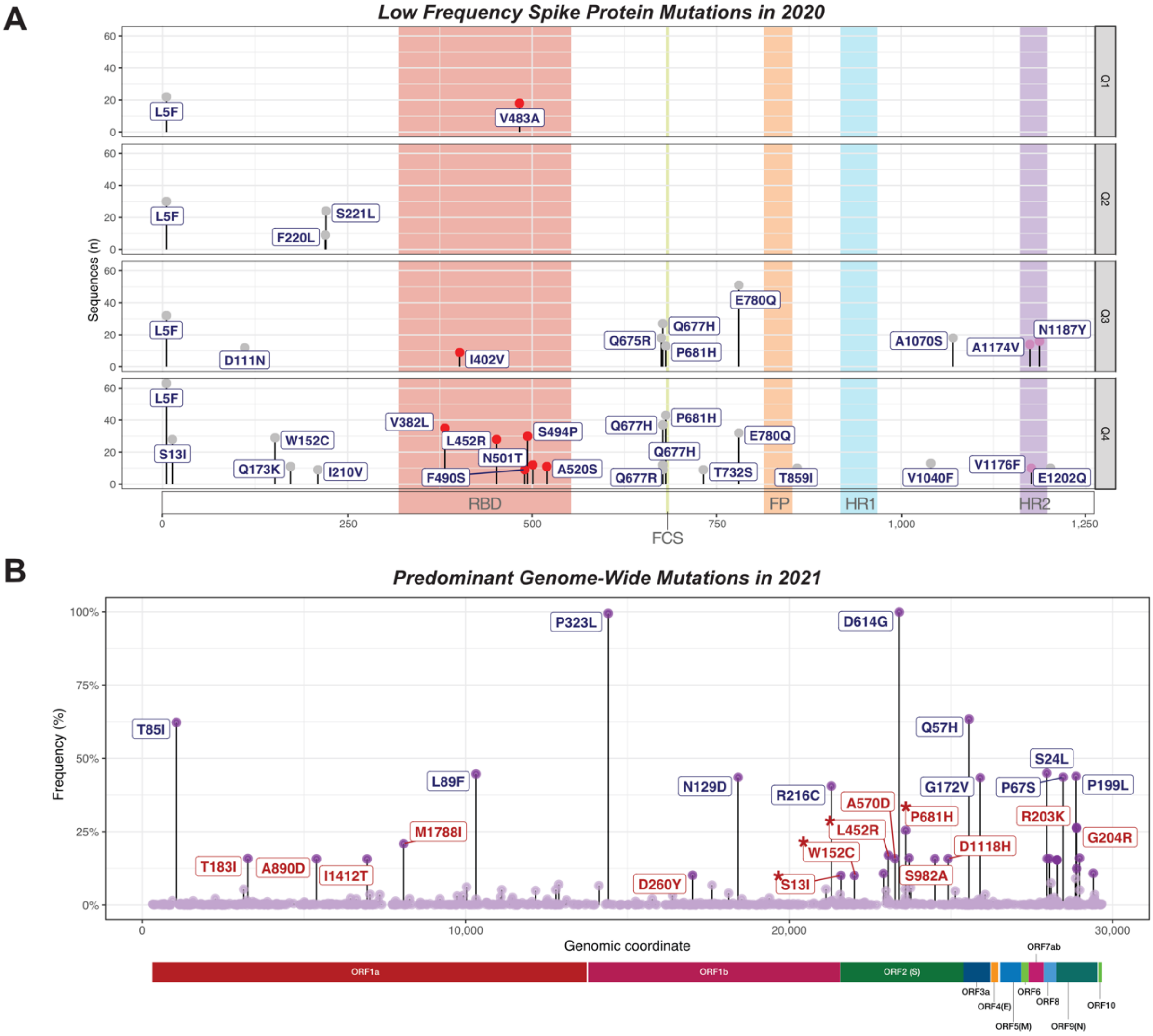


Figure 4.19 A fraction of low frequency spike mutations become predominant in 2021. (**A**) Throughout 2020 there was only one predominant mutation (>10% of the genomes) in the Spike protein (D614G; shown in Figure 4.16A), however a number of low frequency spike mutations (LFSMs; >0.1% frequency) was being accumulated throughout the year 2020. This plot shows the number of sequences (n; SARS-CoV-2 genomes) that each LFSM was found in the different quartiles(Q1-Q4) of 2020. The different mutations are discussed in the main text of 4.3.4. The different domains of the Spike protein are color-coded: in red is the receptor binding domain (RBD), in green the furin cleavage site (FCS), in orange the fusion peptide (FP), in turquoise is the heptad repeat region-1 (HR1) and in violet the heptad repeat region-2 (HR2). (**B**) Frequency plot of the predominant missense mutations (n=34; >10% frequency) detected from the SARS-CoV-2 genomes isolated in 2021. 12 of the previously identified predominant mutations in 2020 are still predominantly

found in 2021 (shown in purple labels) and 22 new ones prominently appeared (highlighted in red). A few of the LFSMs found in the Q4 of 2020 became prominent Spike mutations in SARS-CoV-2 in 2021 (noted with asterisks), which is due to the abundance of variants of concern (VOCs) in 2021. Not all predominant mutations are labelled. The complete set of mutations is in the heatmap of Figure 4.21. Adapted figure from Tasakis et al., 2021 under the Creative Commons License 4.0. The data and illustration were produced by myself based on suggestions by Dr. Marilyn Diaz (for panel A).

As I showed in Figure 4.19B, in 2021 alone I overall detected 34 predominant mutations (>10% of the genomes), 12 of which were previously reported as predominant in 2020 (Figure 4.16A). From 2020 to 2021, the mutations P504L and Y541C found in the ORF1b lost their prominence, but 22 new ones rose to prominence in 2021 (labelled in red in Figure 4.19B and summarized in Figure 4.21). By the end of 2020 a number of VOCs were reported worldwide (Sanyaolu et al., 2021), a handful of which I also found the cohort I investigated covering genomes till the end of March 2021, after profiling my dataset with the pangolin tool (see 3.4.2) for the PANGO lineages (Rambaut et al., 2020; O'Toole et al., 2021). The variants I detected with the number of genomes (n) in parentheses are summarized here below and their defining mutations in the Spike-encoding ORF2 are shown in Figure 4.20:

- B.1.1.7 (n=5166): first detected in September 2020 in the United Kingdom and rapidly spread worldwide, being identified as very infectious and able to escape antibody neutralization (Davies et al., 2021; O'Toole et al., 2021; Planas et al., 2021),
- B.1.429 (n=2285) and B.1.427 (n=1030): first detected in California, United States in July and June 2020 respectively (McCallum et al., 2021),
- B.1.526 (n=150): detected first in New York City, United States (West et al., 2021),
- P.1 (n=52): first reported in Brazil in February 2020 as a highly concerning variant due to being able to escape neutralizing antibodies (Maggi et al., 2021; P. Wang et al., 2021),
- B.1.351 (n=39): first reported in South Africa in October 2020 and shown to also be able to escape antibody neutralization (Planas et al., 2021).

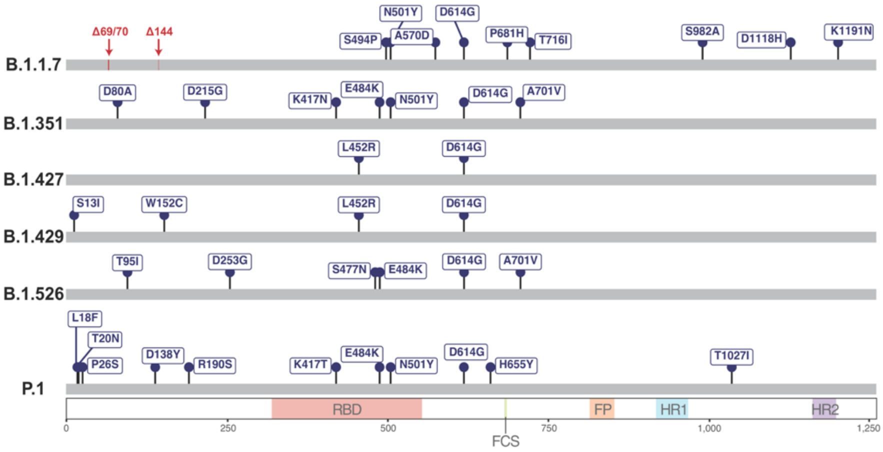


Figure 4.20 Variants of concern detected and their defining spike mutations. 6 different variants of concern (VOCs) were detected in the cohort of SARS-CoV-2 genomes in 2021. In this schematic representation the different Spike mutations, which define each VOC, are shown across the protein sequence. The different domains of the protein are color coded, same as in Figure 4.19A. Figure from Tasakis et al., 2021. It is reused under the Creative Commons License 4.0. This illustration was produced by myself.

After comparing the defining Spike mutations of the different variants shown in Figure 4.20, with the predominant mutations I found in the first three months of 2021 (Figure 4.19B and 4.21), I realized that the mutations reported as predominant were mostly because they were carried in due to the increasing abundance of the VOCs. For example, the mutations S13I, W152C and L452R are predominant due to the B.1.427/B.1.419, while A570D, S982A, P681H, D1118H, are due to B.1.1.7 and N501Y is due to B.1.1.7, B.1.351, and P.1 together. For the same reason, the mutations T183I, A890D, I1412T are in ORF1a are also predominant due to B.1.1.7 and the D260Y in ORF1b due to B.1.427/B.1.419 (Tzou et al., 2020; O'Toole et al., 2021). However, I noticed that there are mutations, such as the ORF1a: M1788I, which were not found in any VOCs from the ones I detected in my dataset. Therefore, I extended the signatures analysis I described above and, first, projected the signatures I found in 2020 to 2021, and second, I called new signatures from the mutations found prominent in 2021 alone in the exact same way as I described before. I found that the signature s48 (Figure 4.17 and 4.18) was the only one from 2020 that was still abundant (about 20% of the genomes in 2021) across the United States, while a few others were barely detectable. Furthermore, I found 348 new putative signatures in 2021 (starting with 'i'; summarized in heatmap of Figure 4.21) and I realized that the VOCs were appearing in my dataset as different, yet related, signatures (Figure 4.22). For example, B.1.1.7, the most abundant VOC in my dataset (up to March 2021) had primarily the signature i342, but also as i335 or i300 etc. (Figure 4.22).

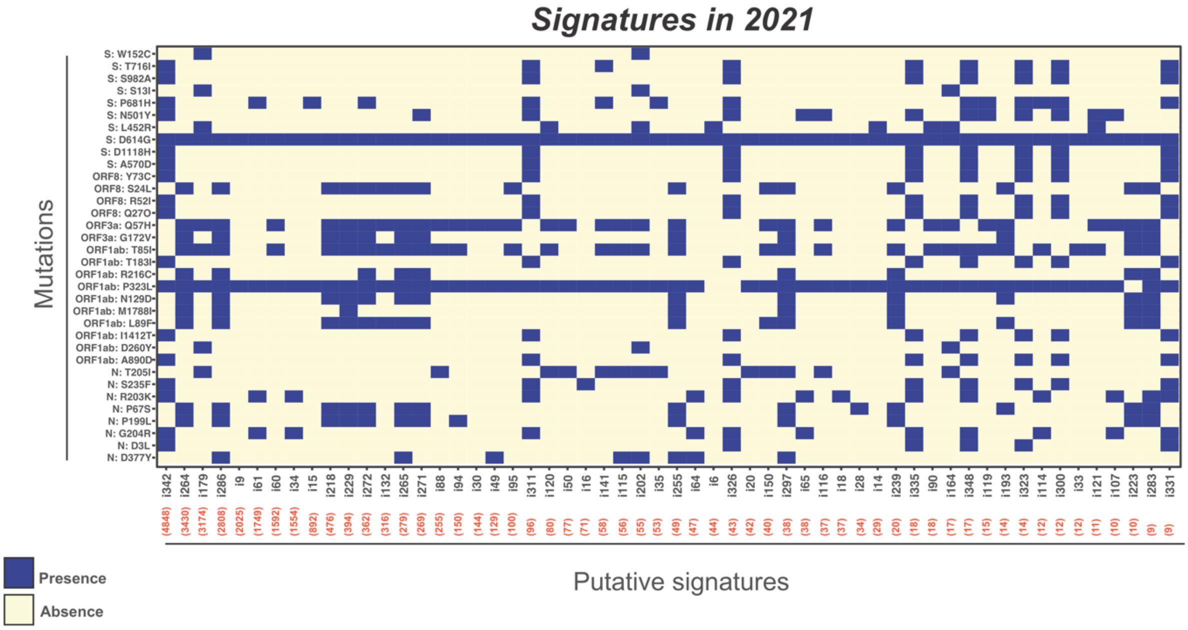


Figure 4.21 Putative signatures of co-occurring predominant mutations in 2021. Heatmap summarizing the unique combinations of mutations predominantly found in SARS-CoV-2 genomes in 2021 (rows), which compile a set of overall 348 new signatures (columns). Presence or absence of mutation in a given signature is noted with blue or yellow respectively. The number of genomes supporting a signature is given in a red label across the x axis and only signatures with more than 9 genomes are demonstrated. The signature s48 (not shown), first detected in 2020, was also abundantly present in 2021 as the top signature (5838 genomes). Adapted figure from Tasakis et al., 2021 under the Creative Commons License 4.0. The data and illustration were produced by myself.

The signature-approach I propose here, which relies on the unique combinations of co-existing mutations, is a method that provides an in-depth resolution with regards to the acquisition of mutations, which in the case of SARS-CoV-2 is dynamic. As I present in the Figure 4.22, all the VOCs I detected in the data covering through March 2021, have been profiled with more than one signatures. B.1.1.7 is mainly present in my dataset with the signature i342, and less frequently with a number of other signatures (i.e. i300-348). In this case i342 is the typical B.1.1.7 variant with its defining mutations as in (O'Toole et al., 2021) and the other signatures are representing B.1.1.7 sub-variants, perhaps specific to the United States. The same applies for the variants B.1.427/B.1.429 primarily sharing the signature i179, but also appearing in other signatures (i.e. i144-200). However, despite of the presence of the VOCs in the first quarter of 2021, the thus-far non-concerning lineage B.1.2 was still abundant in the United States, representing 1/3 of the total number of genomes in 2021. The signature s48 that "grew" in the United States in 2020, as I showed earlier (see 4.3.3), and is still prevalent in the dataset in 2021, accounting about ¼ of the genomes profiled with B.1.2. However, more signatures, such as i264 and i286, which are related to s48 are abundant for B.1.2 as well and likely evolved from s48. Despite the fact that this clade of the lineages has not been identified yet as problematic, it is not unlikely that it may in fact be one, since there is a clear positive selection for s48 and related signatures.

To conclude, the genomic profile of SARS-CoV-2 has significantly changed ever since the beginning from the pandemic and the jump from 2020 to 2021 was crucial, especially considering the presence of VOCs. In Figure 4.23, I focus in States that had an adequate number of sequences covering the 15 first month of the pandemic I investigated. The first striking observation is that the non-variant s0 version of the virus is absent as of June 2020 in all States, if present. For instance, according to my findings, Florida (FL) or Maryland (MD) never had the non-variant version. Furthermore, from the beginning of the pandemic and before 2021, there were a number of different signatures of co-existing mutations circulating already, but by the end of 2020 a few were being selected. For example, in California (CA), Massachusetts (MA) and Maryland (MD), s6, s22 and s48 were the ones present mostly by the end of 2020, while in Wisconsin (WI) and Washington (WA) s6 and s48 were the major ones. In Florida, s48 was prominent only as of early 2021, which may have happened through migration of viral genomes. From 2020 to 2021, the only signature that unanimously remains is the s48, while a plethora of new signatures appears in 2021 (starting with 'i'), shifting virtually entirely the mutational profile of the SARS-CoV-2. The same time, VOCs were abundantly present and in many different mutational profiles (Figure 4.22). B.1.1.7 was the major VOC I found in early 2021, representing in about 20% of the viral genomes (highlighted signature colors in red, Figure 4.23). Even more abundant, however, and more diverse as per signatures, was the "non-concerning" lineage B.1.2 (signature labels in blue, Figure 4.23). Overall, the viral genome has changed dynamically ever since the beginning of the pandemic, as many different variants appear through gradual accumulation of mutations, which is likely through the continuous infections. This variation, especially in the spike protein, but also important in the other genomic regions of SARS-CoV-2 is crucial for vaccination strategies or emerging therapeutics.

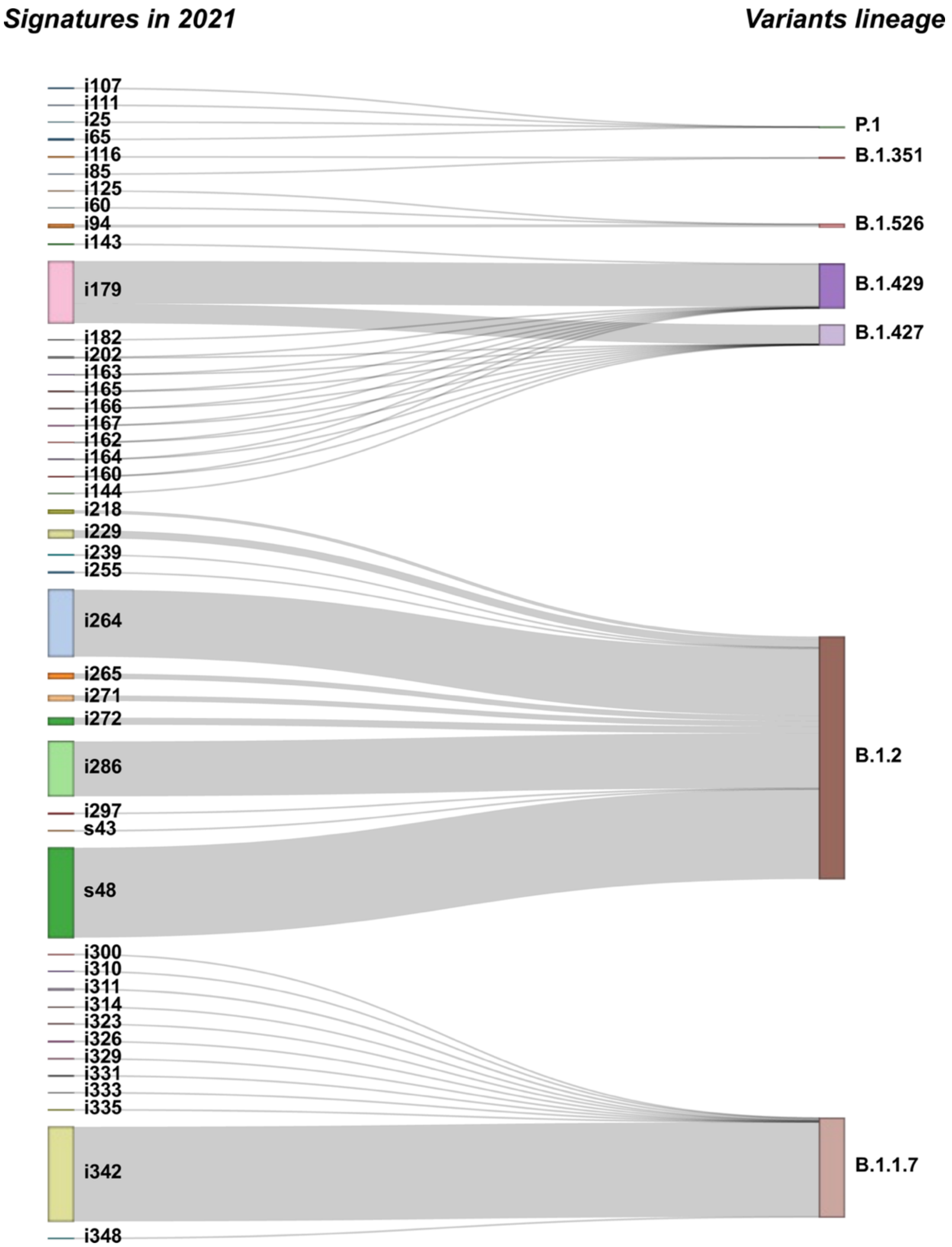


Figure 4.22 A mutational signature approach provides high resolution in SARS-CoV-2 genome evolution. 6 different variants of concern were abundantly detected in 2021 within the cohort of SARS-CoV-2 genomes: B.1.1.7, B.1.429, B.1.427, B.1.526, B.1.351 and P.1. However, the predominant lineage of genomes (1/3 of the cohort) is the non-concerning B.1.2 thus far. This Sankey diagram shows the different signatures of co-occurring mutations detected in 2021 in >0.01% of the cohort (left) and to which variant lineages their respective genomes belong to (right). Although there are some major signatures profiling a variant lineage (i.e. i342 for B.1.1.7), there is a number of additional signatures per lineage. The thickness of the different connections corresponds to the abundance of the genomes supporting the connection. The number of genomes per signature are shown in Figure 4.21. Figure from Tasakis et al., 2021. It is reused under the Creative Commons License 4.0. The data and illustration were produced by myself.

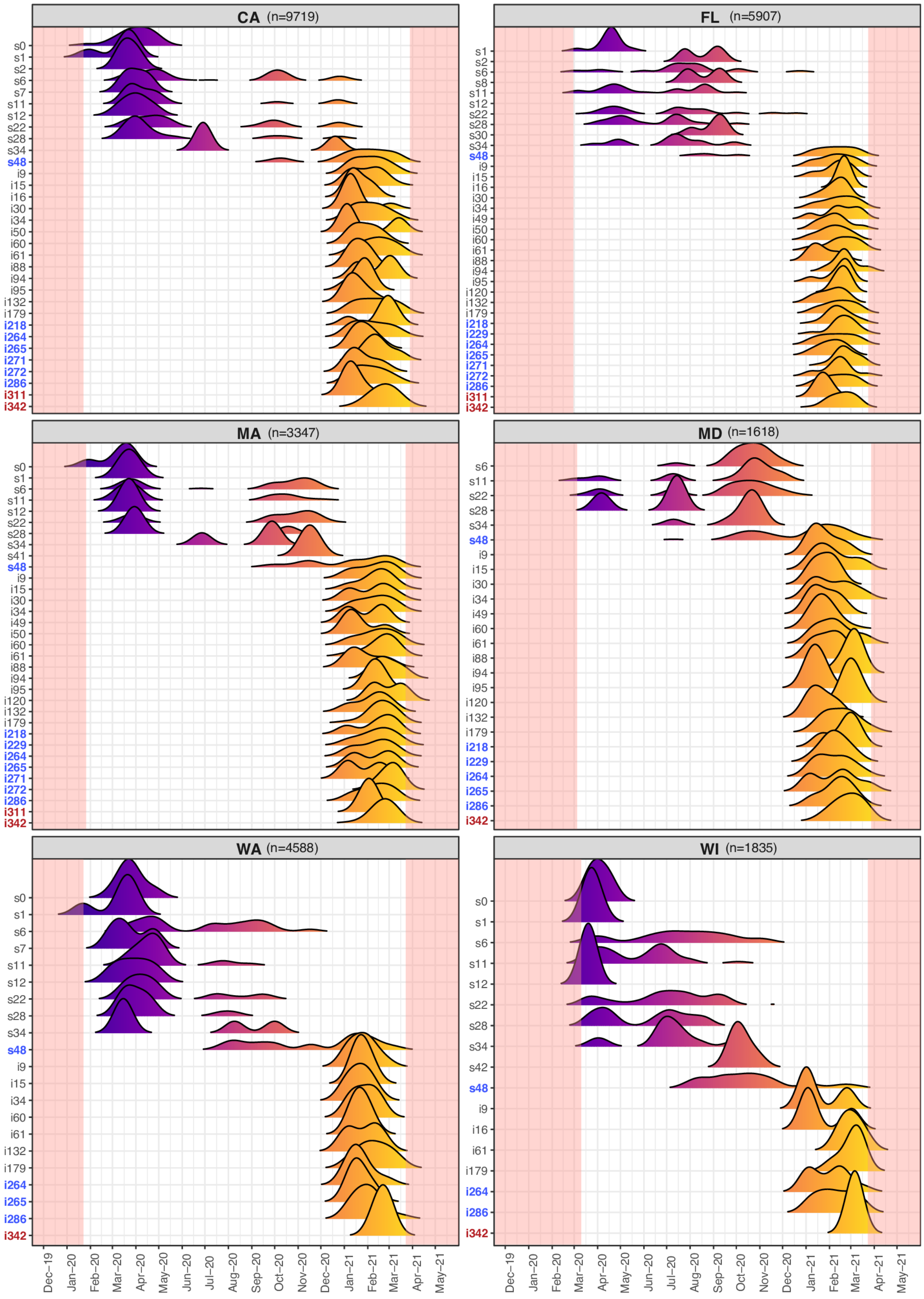


Figure 4.23 The mutational signatures of SARS-CoV-2 underline a dynamically changing profile, which is diverse even between different geographic locations of the United States. This multi-faceted density plot shows the abundance of each signature (y axes; found in at >0.1% of genomes) over time (x axes) throughout the year 2020 and the first quartile of 2021 for 6 states. The states of California (CA), Florida (FL), Massachusetts (MA), Maryland (MD), Washington (WA) and Wisconsin (WI), were selected because they had adequate number viral genomes (n) covering

the complete collection time of the cohort (the entire 2020 and first quartile of 2021). The shaded areas at the beginning and the end of each plot correspond to the time in which limited data were only available. Density curves are filled gradually changing colors to show transition in time. Overall the mutational profile of SARS-CoV-2 has utterly changed from 2020 to 2021. The non-variant s0 is absent since June 2020, while it was not detected at all in FL and MD. About the same time s0 disappeared, the signature s48 among others emerged and in 2021 s48 remained the predominant signature. s48 belongs to the B.1.2 lineage and in 2021 more related signatures of the same lineage emerged (highlighted in blue), which evolved by acquiring new mutations. The utter shift in 2021 is also partly due to the appearance of VOCs due to migration, and in particular the B.1.1.7 (related signatures to this variant are highlighted in red). Figure from Tasakis et al., 2021. Reused under the Creative Commons License 4.0. The data and illustration were produced by myself.

4.3.5 Variants of concern accumulate additional Spike mutations and are further evolving.

I previously found that VOCs, such as B.1.1.7, B.1.427/B.1.429, B.1.526, P.1 and B.1.351, were present in my dataset (see Figure 4.20 and associated text) and were profiled with different signatures of co-occurring mutations in their genomes (Figure 4.22). This suggests that VOCs may accumulate mutations and further evolve. In the cohort I investigated, most of the VOCs appeared in the United States by the end of 2020 and were abundant in the first quarter of 2021. As shown in Figure 4.24, the most abundant VOC is the B.1.1.7 (n=5166 genomes), while B.1.429 and B.1.427 were also abundant with n=2285 and n=1030 genomes respectively, followed by B.1.526 (n=150), P.1 (n=52) and B.1.351 (n=39). Such variants warrant surveillance due to the spike mutations they carry, which may interfere with immunization strategies, such as vaccinations (Darby and Hiscox, 2021). Therefore, I focused on the spike protein and identified spike mutations other than their defining ones (shown in Figure 4.20). As shown in Figure 4.24, I identified a number of spike mutations some of which were sporadic (blue labels; low frequency but >0.1% of VOC genomes) and others were more frequent and recurrent (red labels). The latter ones are noted as of the date of their appearance and detected thereafter in overall frequency of >1% of genomes per VOC.

Interestingly, the mutation L5F, which I first detected as LFSM "surfing" throughout 2020 in low frequencies (see Figure 4.19A), now appears recurrently in B.1.1.7, B.1.429, B.1.526 and P.1 genomes. This mutation has now been described thus far as problematic, but there is a clear positive selection for it, as my data show. Furthermore, B.1.427 genomes have acquired S13I and W152C mutations, which are part of the defining mutations of B.1.429. It should be noted, that these two separate variants are now appearing as a merged B.1.427/B.1.429 (McCallum et al., 2021). Because of examples as such, I looked into the possibility that SARS-CoV-2 may recombine, as others in the literature have noted this possibility (Gallaher, 2020), but I did not detect in my cohort of genomes evidence of recombination. Moreover, mutations such as Q677H or Q677R, T859I, V1040F, V1176F or E1202Q, which I also found as LFSM in 2020 are appearing in low frequency mutations in the Spike of VOCs in 2021, which indicates that there is a driving force for recurrent mutations. This can potentially be possible due to ADAR or APOBEC activity, considering that they have preferential deamination motifs (Giorgio et al., 2020; Mourier et al., 2021), which could explain mutations in specific genomic positions.

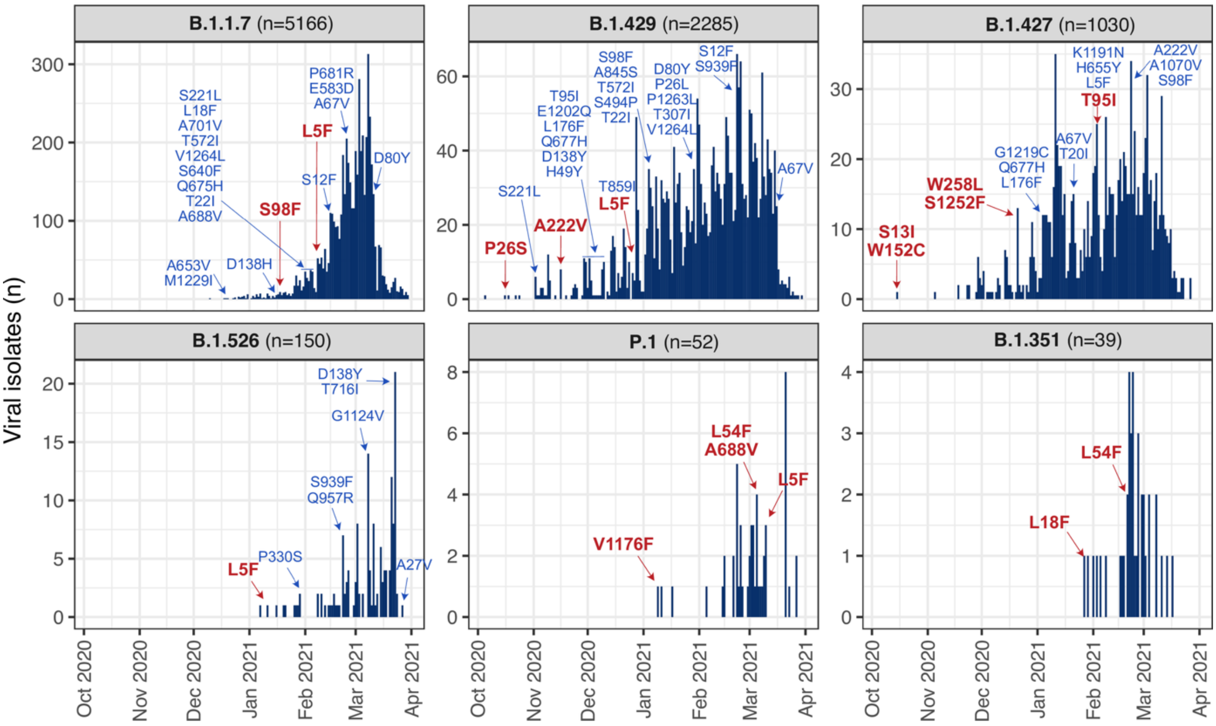


Figure 4.24 Variants of concern are further evolving by acquiring additional spike mutations. 6 VOCs were detected in the cohort of SARS-CoV-2 genomes from the end of 2020, but primarily in the first quartile of 2021. The VOCs detected by abundance were B.1.1.7 (n=5166), B.1.429 (n=2285), B.1.427 (n=1030), B.1.526 (n=150), P.1 (n=52), B.1.351 (n=39). Frequency histograms for every VOC are showing the number of isolates (n; y axes) over time (x axes). On the histograms, additional spike mutations, other than their defining ones (Figure 4.20), are annotated with labels at the time of their first appearance. Labels in bold red are showing recurrent mutations found in more than 1% of the genomes from the respective VOC thereafter, while labels in light blue are showing less frequent mutations (<1% and >0.1% of genomes). L5F is a recurrent mutation, which is now found in almost all the VOCs and was also detected as an LFSM (Figure 4.19A), “surfing” in low frequencies in 2020. Additional mutations were found as LFSMs, such as Q677H, Q677R, T859I among others. Adapted figure from Tasakis et al., 2021 under the Creative Commons License 4.0. The data and illustration were produced by myself based on suggestions by Dr. Marilyn Diaz, Prof. Laurent Verkoczy and Prof. Papavasiliou.

4.3.6 Discussion

The first case of COVID-19 was reported in Wuhan, China in December 2019 after an infection of a novel coronavirus, later termed as SARS-CoV-2 (Wang et al., 2020). SARS-CoV-2 is a positive-sense single-stranded RNA (ssRNA+) coronavirus and its rapid worldwide spread led the World Health Organization to declare COVID-19 a global pandemic on March 11th, 2020 (Cucinotta and Vanelli, 2020). Up until the third quartile of 2020, it was widely believed that SARS-CoV-2 is keeping a stable genomic profile, because its genome encodes for a non-structural protein 14 (Nsp14) guaranteeing a strict proofreading activity during RNA synthesis toward viral replication, also found in SARS-CoV-1 (Rausch et al., 2020). However, by early 2021 and even more so now, a number of variants have been reported as concerning (termed as Variants of Concern or VOCs) due to mutations in the viral Spike protein, which is recognized by the host’s receptor ACE2 initiating the infection (Sanyaolu et al., 2021). I hypothesized that the driving force behind the accumulation of mutations in the SARS-CoV-2 genome is the continuous interindividual spread of the virus through infection for two major reasons: 1) intracellular enzymatic mechanisms, such as ADARs or APOBECs deaminases, may be the reason for generating mutations in the viral genome, and 2) through successful infections, mutations may be selected toward enhancing SARS-CoV-2 for immune evasion, establishing an overall "efficient" viral profile to bypass immunity. Therefore, as infections continue to occur with time they should be accumulating progressively though selection and expansion.

I tested my hypothesis in a cohort of 62 211 fully covered and publicly available SARS-CoV-2 genomes isolated from infected individuals from the United States from early 2020 up until March 2021 (see 3.4.1). First, I found that mutations in the viral genome are indeed gradually accumulating over time as I hypothesized (Figure 4.15A) and, second, the intrinsic pattern of mutations in my cohort (Figure 4.15B) pinpoints to primarily APOBEC and ADAR deaminase activity in the viral genome, as others also found (Giorgio et al., 2020; Klimczak et al., 2020). The vast majority of mutations across the SARS-CoV-2 genome were C-to-U(T) which are generally attributed to the activity of APOBEC3 deaminases, which is mechanistically possible since they can target ssRNA (Poulain et al., 2020). This broadens even more the horizons of the functions of the APOBEC3 subfamily, especially considering that they were previously considered as antivirals (Stavrou and Ross, 2015). A-to-G or T-to-C base changes in the viral genome have been attributed ADAR activity on dsRNA intermediates of SARS-CoV-2 during viral replication (Giorgio et al., 2020), which should be fundamentally possible by the primarily cytoplasmic and interferon-inducible ADAR1-p150 isoform as also suggested for other viruses (Doria et al., 2009; Lamers et al., 2019). However, it is not impossible that other RNA modifications in the SARS-CoV-2 genome have given rise to mutations. For example, the G-to-U(T) change, which is also prevalent in the intrinsic pattern (Figure 4.15B), may be due to methylated $m^{22}G$ or m^1G guanosines according to a high throughput method for identifying RNA modifications through cDNA sequencing (Ryvkin et al., 2013). Intracellular mechanisms of RNA modification and editing can therefore be crucial for SARS-CoV-2 genome evolution as infections continue to occur and the virus is aggregating mutations progressively.

A first step to explore with my dataset how the genome of SARS-CoV-2 has evolved, was to evaluate the predominant mutations in the viral genome. I observed that the distribution of the various Single Nucleotide Variations (SNVs) in aggregate per Open Reading Frame (ORF) of the SARS-CoV-2 genome (Figure 4.15C) was different from the intrinsic pattern of mutations (Figure 4.15B). This was due to the fact that specific mutations were enriched in the different ORFs. I therefore screened the SARS-CoV-2 genome and I found that in 2020 there were 14 predominant mutations (in at least 10% of the sequences/genomes of the cohort), summarized in Table 4.1 and Figure 4.16A. Most of these mutations were due to C-to-U(T) or A-to-G/T-to-C base changes and among the top ones I found the mutation D614G (>80% frequency) in the Spike protein. D614G was the first mutation thoroughly explored as it was noted as concerning, due to association with increased infectivity and severity of COVID-19 (Hou et al., 2020). But other mutations I also found as predominant, were in key components of the virus, such as the P323L (~82% frequency) in the RNA-dependent RNA polymerase or the N129D in the proof-reading exonuclease Nsp14 (~11% frequency). Due to discrepancies in the mutation frequencies, I considered that there may be different timing patterns between the different mutations, as per when they appeared in the cohort or when they were brought to prominence. I indeed found that the 14 predominant mutations in 2020 (Figure 4.16A and Table 4.1) were grouped into four distinct different patterns (A-D, Figure 4.16B) with regards to their abundance during 2020. Particularly, pattern A represents a group of mutations that appeared in the second half of 2020 (ORF1a: L89F, 1b: N129D, R216C, 3a: G172V, N: P199L, N: P67S), pattern B entails mutations that were abundantly present throughout 2020 (ORF1a: T85I, 1b: P323L, S:D614G, 3a: Q57H), while pattern C mutations that were abundant early in 2020 (ORF 1b: P540L, Y541C, 8:L84S) and disappeared as soon as the mutations of pattern A appeared, and last, one mutation (ORF8: S24L) which follows a unique pattern D, appearing to “surf” throughout the year 2020. These findings indicate that there were at least three major SARS-CoV-2 genomic variants circulating the United States in 2020, according to my findings.

Noticing that the different predominant mutations in 2020 were appearing in groups, I decided to further explore the genome evolution of SARS-CoV-2 with a signature approach of co-existing mutations (for the predominant ones in 2020), inspired by the COSMIC signatures which profile cancer genomes (Alexandrov et al., 2020, 2013). I first compiled a “dictionary” of signatures, which I constructed by calling all the different combinations of mutations existing in my cohort, accounting for overall 48 putative signatures (s1-s48; Figure 4.17), different than the non-variant genome (s0). I focused for the downstream analyses on the signatures that profiled at least 0.1% of the genomes in my cohort, which was 15 variant signatures and the non-variant s0. I performed a time-scaled phylogenetic analysis employing multiple sequence alignments of the first genomes profiled with my cohort for each signature (and annotated the last from each in the time scale), including the reference genome of SARS-CoV-2 (Wuhan-Hu-1), which was the viral isolate from the first infected individual (Wu et al., 2020). As shown in Figure 4.18, my signatures approach provided in-depth resolution in studying how the SARS-CoV-2 genome evolved throughout 2020 and up until very early 2021. Signatures defined with the mutations of pattern C (s1, s12 and s7) along with the non-variant s0, were the ones that appeared first, but disappeared as soon as signatures related with the mutations of pattern B arrived, likely through genetic “drift” for selection (Slatkin and Excoffier, 2012). From this clade of signatures, s6 and s22 remained predominant till the end of 2020 and very early 2021. Of note, is the mutation S:D614G, which clearly provides a selection advantage through genomic fitness, as suggested by others (Plante et al., 2020), considering that a signature which lost this mutation (s8) faded out rapidly in 2020. A number of serial acquisition of mutations led to a mutational “burst”, which likely occurred in June 2020 with the mutations I described above following pattern A, and gave rise to three signatures (s42, s41 and s48) from which s48 remained predominant toward the end of 2020 and early 2021. It is not unlikely that this mutational burst may be associated with the mutation 1b:N129D in the exonuclease Nsp14, which could potentially downgrade the proof-reading activity of Nsp14 allowing more mutations to occur.

Overall, I found three signatures (s6, s22 and s48) likely compiling three distinct variants of SARS-CoV-2 circulating in 2020 and early 2021 in the United States.

The signatures approach I discussed above to describe the genome evolution of SARS-CoV-2 in the United States, may also be beneficial for predicting mutations currently in low frequency which may come to prominence upon positive selection. I focused on the Spike protein, which is key for the SARS-CoV-2 infection process (Luan et al., 2020) and I called all the low-frequency spike mutations (LFSM) detected per quarter (Q1-Q4) of 2020 (Figure 4.19). In Q1, I detected only 2 LFSMs (L5F and V438A), while there was a gradual accumulation of LFSMs throughout the year ending up in Q4 of 2020 with a pool of mutations, which may not appear now as problematic but they may in fact be if positively selected. From 2020 to 2021, many of the LFSMs I detected were brought to prominence (>10% of genomes) in 2021 alone. In Figure 4.19B, I summarize predominant mutations (overall 34, also shown in Figure 4.21) detected in 2021, of which 22 were new compared to 2020 (highlighted in red) and some spike mutations were seen as LFSMs in Q4 2021 (noted with an asterisk). However, this is partly due to the arrival and rapid prominence of the different VOCs in the United States as of early 2021 (Darby and Hiscox, 2021). I profiled all the sequences in my cohort with the PANGO lineages to detect VOCs (Rambaut et al., 2020) The VOCs I found in the cohort I investigated (summarized in Figure 4.20 with their defining Spike mutations) were predominantly B.1.1.7 (n=5166), B.1.429 (n=2285) and B.1.427 (n=1030), but I also detected other less abundant VOCs (B.1.526, P.1 and B.1.351). Most of the newly acquired mutations in 2021 were found in the predominant VOCs I detected, but not all (i.e. 1a:M1788I). Therefore, I repeated the signature generation and profiling pipeline to call new signatures for 2021 (Figure 4.21) and I also projected the signatures of 2020 (Figure 4.17) to the genomes isolated in 2021 to evaluate the genome evolution through its mutational profile. First, I found that the signature s48 from 2020 was still abundant (5838 viral isolates) and in fact the most abundant in 2021. From the newly inferred signatures, I called 348 non-variant putative signatures in 2021 (starting with "i", summarized in Figure 4.21). When I overlaid the variant PANGO lineages with the signatures of both 2020 and 2021 only for the sequences isolated in 2021 (Figure 4.22), I realized that the second most abundant signature i342 was in fact B.1.1.7, but the same VOC lineage was present in my dataset, though less frequently, with more signatures (i300-i348). The same case was for the lineage B.1.2, which is in fact a "core" lineage since the beginning of 2020, and majorly entails the signatures s48, i286 and i264. Although the three first months of 2021, which compiles my dataset, is a short time to draw real-time phylogeny, it is evident that the viral genome is further evolving. And that includes both the VOCs and the B.1.2 lineage, which by the time of my analysis was not identified as problematic. All in all, with the appearance of the VOCs the mutational profile of SARS-CoV-2 has shifted dramatically across the United States throughout the pandemic (Figure 4.23). The three major variants my analysis suggested by the end of 2020 (s6, s22 and s48) "drifted" for selection when the VOCs appeared with only s48 remaining, likely being still "fit" to compete the positive selection of VOCs through fitness in the population.

The last aspect I explored was the further evolution of the VOCs, as the diverse profiling of signatures per lineage suggested (Figure 4.22). As I show in the Figure 4.24, the different VOCs I detected acquired new mutations in their Spike protein, other than their defining ones (Figure 4.20). Some of the mutations acquired were more abundant (in more than 1% of the genomes, highlighted in red) than others (>0.1% of genomes in light blue). A number of mutations were previously detected as LFSM in 2020 (Figure 4.19A) or in other VOCs (Figure 4.20), such as L5F, V1176F, Q677H, E1202Q, T859I, S13I or W152C. Of note, is the mutation L5F which I found as an LFSM in small frequencies throughout 2020 and is now recurrently found in a many VOCs, such as B.1.1.7, B.1.429, B.1.526 and P.1. Additionally, the mutations S13I and W152C are defining mutations for B.1.429, but they are also occasionally found in B.1.427. Last, it should be noted that there is likely a strong selection for phenylalanines (F) in the Spike protein, considering that most of the more abundant and recurrent mutations lead to that amino acid replacement (i.e. L5F, L54F, L18F, V1176F, S98F among others). Therefore, my findings indicate that VOCs are also further evolving, by acquiring new mutations, naturally undergoing selection through infection. Although it was suggested that the SARS-CoV-2 genome may recombine (Gallaher, 2020), I could not detect evidence of recombination in the cohort I investigated. However, what I believe that is a likely scenario is that APOBECs or ADARs are behind these recurrent mutations, considering that they have preferential deamination motifs (Eggington et al., 2011; Poulain et al., 2020). This can also be the reason why same mutations in the SARS-CoV-2 genome have been convergently evolved in different parts of the world (Martin et al., 2021; Zhou et al., 2021), and furthermore why infected individuals also acquire new mutations (some of which recurrent and found in my data) in the different genomic copies of their viral load during persistent infections (Kemp et al., 2021).

All in all, my findings underscore important aspects in the genome evolution of SARS-CoV-2, a major driving force of which is the RNA deaminases. Here, I presented and discussed data from an evolutionary perspective which give crucial insights for my bigger question in this dissertation, which is the genomic damage by RNA deaminases.

5. Conclusions and Future Perspectives

The family of Adenosine deaminases acting on RNA (ADAR) comprises of 3 enzymes, two of which, ADAR1 and ADAR2, have demonstrated deaminase activity; both these enzymes deaminate Adenosines-to-Inosines (A-to-I), where Inosines are recognized as Guanosines (G), which is also known as RNA editing. (Nishikura, 2010). A number of functions have been described for ADAR1. It has been demonstrated that it deaminates endogenous dsRNAs, so as to block cellular response to them as non-self dsRNAs which would stimulate components of the RIG-I-Like Receptor pathway (RLP), such as MAVS or MDA5 and induce interferon (IFN) type-I response (Liddicoat et al., 2015). Although the last highlights the importance of ADAR1 in immunity, allowing the host to trigger antiviral response, pro-viral roles have been suggested as well for this enzyme, as it can block RLP or PKR (Protein Kinase R), which also recognizes dsRNA (Lamers et al., 2019). The 3'UTR of the EIF2AK2 transcript, which encodes PKR, is actually a target of deamination by ADAR1, as previously suggested (Toth et al., 2009), which I validated as presented in chapter 4.1.3. This is perhaps a way for ADAR1 to downregulate PKR, through interfering with translation of its transcript.

However, the impact of ADAR1 in crucial homeostatic mechanisms does not stop there. It has been demonstrated that ADAR1 is several human tumors leading to consequently high A-to-I RNA editing activity, with few exceptions such as types of kidney cancer (Han et al., 2015). Elevated A-to-I RNA editing in tumors, promotes transcriptomic heterogeneity, which likely impacts proteomic heterogeneity as well (Paz-Yaacov et al., 2015). Multiple Myeloma (MM) is no exception; ADAR1 is overexpressed in Multiple Myeloma, either through copy-number gain (1q21 amplification) or IFN induction, as Dr. Laganà found (Tasakis et al., 2020), similarly shown in breast cancer as well (Fumagalli et al., 2015). MM patients with high A-to-I RNA editing activity, consequence of the ADAR1 overexpression, show poor disease outcomes, as previously shown (Lazzari et al., 2017; Teoh et al., 2018), regardless of whether they have the 1q21 gain or not (Tasakis et al., 2020). But here, I explored a different possibility for ADAR1. Recent findings show that ADARs can deaminate DNA within DNA/RNA hybrids *in vitro* (Zheng et al., 2017), which appears to be a crucial function of ADAR1 in maintaining genomic stability by mutating DNA within R-loops of the telomeres (Shiromoto et al., 2021). However, R-loops are formed genome-wide co-transcriptionally (Chen et al., 2019), which raises the possibility that ADAR1 may globally mutate genomic DNA. I hypothesized that, since RNA editing can also be co-transcriptional (Laurencikiene et al., 2006), ADAR1 may "lose touch" with the target-transcript and access R-loops in the cognate locus, formed between the template strand and the nascent RNA, so as to edit RNA and mutate DNA. Therefore, ADAR1-dependent mutations should be found in genes, whose transcripts are (highly) edited (Figure 4.1). I tested this hypothesis in 23 MM patients who have matched RNA-seq and WES data from two timepoints of the disease (pre- and post-relapse) and I obtained correlative data of RNA editing events pre-relapse and unique DNA mutation events post-relapse, likely selected, in the vicinity of the RNA-editing sites (see 4.1.2). The top candidate was the *EIF2AK2* gene, which encodes for PKR, which besides its crucial antiviral role, is involved with p53, NFkB or apoptotic pathways among others and may impact cancer development and progression (Gal-Ben-Ari et al., 2019). Furthermore, not all MM patients retained their global RNA editing activity post-relapse; patients who presented decreased levels of A-to-I RNA editing (measured with the Alu Editing Index; (Roth et al., 2019)), showed an enrichment of acquiring new mutations possibly through ADAR1 activity, which suggests that they may have achieved the same functional outcome, compared to the other group of patients, but now through "fixing" into the genome a permanent DNA mutation (see 4.1.4).

These data, are the first to correlate ADAR-dependent gDNA mutagenesis, in a fashion that DNA mutation may not be the primary function of the enzyme, but it may arise as a collateral genomic damage by an enzyme that aberrantly edits RNA *in situ* (Aim 1). Therefore, ADAR-dependent DNA mutations will be mostly found in genes whose transcripts are edited. If this is true, then I would have described a novel mechanism that cancers may take advantage of, to expand their mutational spectra. To prove that, I employed a series of experiments presented in chapter 4.2, using site-directed mRNA editing tools (principles described in 1.3) to induce A-to-I editing in certain transcripts and look for DNA mutations in their cognate genes. First, I employed Ramos B-cells, which lose expression of IgM through somatic hypermutation in the V_H region by AID (Sale and Neuberger, 1998). Ramos AID-/- B-cells do not mutate their V_H and, therefore, homogeneously present IgM$^+$ cell populations (see 4.2.2). I therefore recruited the endogenous ADAR1, which is the major A-to-I deaminase expressed in these cells (see 4.2.3), with gRNAs (designed as in (Qu et al., 2019)) in pairs targeting both the coding (positive-sense) and template (negative-sense) strands of the locus. I observed abundant loss of IgM in cells with bi-stranded targeting accompanied by minimally specific DNA mutation signal, but no RNA editing was detected prior to that in the transcript of the V_H. After troubleshooting I realized that the loss of function experiment is prone to specific artifacts that may be related to editing, but lead to undesired outcomes (see 4.2.4). Therefore, I took an alternative approach which allowed me to first report RNA editing and then look for DNA mutations, only in the subset of cells that reported high rates of RNA editing through a gain of function approach. For this, I relied on a HEK cell line that expresses an mCherry/eGFP cassette, in which the eGFP gene is inactivated through a premature stop codon (UAG) and, thus, the cells are not eGFP-fluorescent (originally described in Montiel-Gonzalez et al., 2013). I targeted the UAG stop codon with a gRNA, employing the λN-ADAR tool (Montiel-Gonzalez et al., 2013) and isolated 25 000 cells which were editing the UAG stop codon (U**A**G>U**G**G, and therefore activating eGFP). After 5 weeks, I obtained 1 eGFP-activated clone, which reported an A-to-G base change within the originally targeted UAG stop codon (see 4.2.5). This was a first indication that ADAR *may* mutate in a rate of 1 in 25 000. In a repeat, I upscaled the priming population of eGFP-activated cells via RNA editing and I found overall 7 clones that contained eGFP-activated cells, which would adjust the mutation rate to 1 in 50 000. However, because the A-to-G base change was detected from cDNA amplicons of the cassette, I tried an alternative tool, RESTORE, which recruits the endogenous

ADAR1 (Merkle et al., 2019), and in a priming population of 25 000 eGFP-activated cells via site-directed mRNA editing I detected 1 clone which was purely eGFP+ due to 100% A-to-G base change on-target from the gDNA amplicons. With this finding I reported a validated ADAR1-dependent mutation in a rate of 1 in 25 000 (see 4.2.6), also achieving the 'Aim 1' of my dissertation. Replicates of this experiment are currently ongoing to validate the ADAR1-dependent mutation rate. In a similar fashion, and especially because the Ramos AID-/- system remains important for the reasons above (also discussed in chapter 4.2.4 and 4.2.7), a Ramos AID-/- IgM$^-$ cell line due to a premature UAG stop codon in the V_H gene is being engineered, so as to additionally introduce RNA editing bias, which will be reported as IgM gain.

Finally, I explored the concept of genomic alteration by RNA deaminases under the perspective of evolution, which was the 'Aim 3' of my dissertation. To do that, I leveraged the fact that RNA deaminases from both ADAR and APOBEC families are known to mutate viral genomes (Samuel, 2012; Stavrou and Ross, 2015), which also appears to be the case for SARS-CoV-2, the coronavirus responsible for COVID-19 (Giorgio et al., 2020; Klimczak et al., 2020). RNA editing of SARS-CoV-2 is host dependent (Giorgio et al., 2020), which suggests that mutations by RNA deaminases can potentially be introduced from host-to-host, allowing the gradual accumulation of mutations throughout the viral genome relatively rapidly. A related phenomenon in cancer biology, would be the accumulation of selected mutations through clonal evolution (Greaves and Maley, 2012), rules that ADAR-dependent mutations in cancer genomes, as I discussed throughout this dissertation, will likely follow. To address mutational expansion by RNA deaminases on SARS-CoV-2, I employed a publicly available dataset of fully-covered SARS-CoV-2 genomes (see 3.4.1) during the first year of the pandemic from the United States, where lockdown or related regulations were not as strict as in other countries at the time. I found that the viral genome throughout the year was indeed gradually accumulating mutations (Figure 4.15), many of which co-occurred as "signatures" on the genome (Figures 4.17, 4.18, 4.21), leading to a substantially different genomic profile toward the first quarter of 2021 (Figures 4.18 and Figure 4.23). The intrinsic pattern of mutations (including synonymous and nonsynonymous changes) ranked C-to-T(U), A-to-G or T(U)-to-C changes amongst the most abundant SNVs (Figure 4.15B), and the majority of the predominant mutations were of those base changes (Table 4.1), which is very likely through APOBEC and ADAR activity on the viral genome, as also suggested by others (Giorgio et al., 2020; Miladi et al., 2020; Poulain et al., 2020; Simmonds, 2020). Furthermore, it is suggested that certain mutations in SARS-CoV-2 genome, may have convergently emerged from different parts of the world (van Dorp et al., 2020; Zhou et al., 2021), which is something that RNA deaminases could very likely be responsible of, since they are known to have preferential deamination motifs (Chen and MacCarthy, 2017). This, may in fact be "good news" as the viral evolution might "exhaust" available deamination motifs in the near future, which may slow down the diversification of the viral genomes and, thus, attenuate the emergence of variants of concern (McCormick et al., 2021), which are also under evolution (Figure 4.24), and hopefully terminate the pandemic faster.

To conclude, RNA deaminases play a prominent role in health and disease, regulating several components of pathways crucial to homeostasis. Novel roles and mechanistic functions of these enzymes are constantly expanding their potential. Here, I focused on ADARs and their emerging versatile role of not only being RNA editors, but also potentially DNA mutators. I focused on Multiple Myeloma, a cancer which is sensitive to ADAR1 activity, and I showed the potential impact of genome-wide ADAR1-mediated mutagenesis in disease progression, starting from patient data. I hypothesize that this phenomenon is not limited to Multiple Myeloma, but rather is exploited by most tumors (all of which express high levels of ADAR1) to expand their mutational spectra and therefore adapt. I further employed a series of proof-of-concept experiments involving site-directed mRNA editing tools, to explore the possibility that ADAR-mediated RNA editing may drive specific DNA mutation. These experiments have addressed a number of open questions in the field and have given important answers to the concept of ADAR-mediated mutagenesis, but have also provided a better insight for genome editing tools with ADARs, which are currently in development.

6. Appendix A

Table A1. List of primers used for amplicons generation. Primers are summarized by name, Sequence in the 5’→3’ orientation and the melting temperature (ºC) for each (Tm).

Name	Sequence (5’→ 3’)	Tm (ºC)
Eif2ak2-Fw	TCCAAATCAAATTAACCCCATAAGAGCCAC	61
Eif2ak2-Rv	AGAGGAGTTGGCAACTAATTGGATGTGGGG	66
pENTR-Fw	TTTTTTCTAGACCCAGCTTTCTTGTA	57
pENTR-Rv	GGTGTTTCGTCCTTTCCACA	56
qADAR1-Fw	CACTTCCAGTGCGGAGTAGC	64
qADAR1-Rv	CCCTGCCGCGGATTCATT	58
qADAR2-Fw	CTGACACGCTCTTCAATGGTT	62
qADAR2-Rv	GGCGCAGTTCGTTCAAGAT	58
qActb-Fw	TGGAGAAAATCTGGCACCACACC	70
qActb-Rv	GATGGGCACAGTGTGGGTGACCC	76
qGAPDH-Fw	GAAGGTGAAGGTCGGAGTC	60
qGAPDH-Rv	GAAGATGGTGATGGGATTTC	58
Mavs-Fw	TACCCTGCCTGGCCTCAAACTATTA	74
Mavs-Rv	ACTTCATGCTGTCTGGGAGCAA	66
Vh-cDNA-Fw	TGAAACACCTGTGGTTCTTCCT	58
Vh-cDNA-Rv	GGGAATTCTCACAGGAGACGA	57
Vh-gDNA-Fw	CCCCAAGCTTCCCAGGTGCAGCTACAGCAG	71
Vh-gDNA-Rv	GCGGTACCTGAGGAGACGGTGACC	66
pmCherry-Fw	CGCCTACAACGTCAACATCAAGC	70
peGFP-Rv	GGACTGGGTGCTCAGATAATGGTT	72

7. Appendix B

Table A2. List of gRNAs. gRNAs are summarized by name, sequence (5'→3') and target-transcript. Underlined parts of the sequence represent BoxB loops, underline parts in red represent GluR2 loops and the C in bold stands for the A:C mismatch for the targeted adenosine (see 1.3.1). Asterisks (*) in the name note that the gRNA was used for generating a knock-out of the target with the respective gRNA.

Name	Sequence (5'→ 3')	Target
gRNA-01-Fw*	caccGCTAGAGGAAGCCAAAGCCA	*ADAR1*(exon 3)
gRNA-01-Rv*	aaacTGGCTTTGGCTTCCTCTAGC	*ADAR1*(exon 3)
gRNA-02-Fw*	caccGGACAGGAGACGGAATTCGC	*ADAR1*(exon 4)
gRNA-02-Rv*	aaacGCGAATTCCGTCTCCTGTCC	*ADAR1*(exon 4)
gRNA-NT-Fw*	caccGTATTACTGATATTGGT	None
gRNA-NT-Rv*	aaacACCAATATCAGTAATAC	None
arRNA1	tggcggatccagctccagtagtaaccactgaaggacccacca**C**aaacaccgcaggtgagggacagggtctccgaaggcttc	V_H
arRNA2	gccccttccctgggggctggcggatccagctccagtag**C**aaccactgaaggacccaccataaacaccgcaggtgagggaca	V_H
arRNA3	agacgtccataccgtacctcccgtctgtgccaggactcgccc**C**agtaataactctcgcacagtaatacacagccgtgt	V_H
arRNA4	gaagccttcggagaccctgtccctcacctgcggtgtttatggtgggtccttcagtggttactactggagctggatccgcca	V_H
arRNA5	acacggctgtgtattactgtgcgagagttattactagggcgagtcctggcacagacgggaggtacggtatggacgtct	V_H
arRNA(-) or gCtr	caggagggcgggccctgaaaaagggccatgggatgcccatcgaagatgagggtgagggcccctgaaaaagggcccgggggcgg	None
gGFP	tcagggtagtggccctgaaaaagggccaagtgttggc**C**atggaacaggtagttttcggccctgaaaaagggcctagtgcaaat	eGFP
gGFP_R	ggtgaatagtataacaatatgctaaatgttgttatagtatccacctagtgacaagtgttggc**C**atggaacaggtagttt	eGFP

8. Appendix C

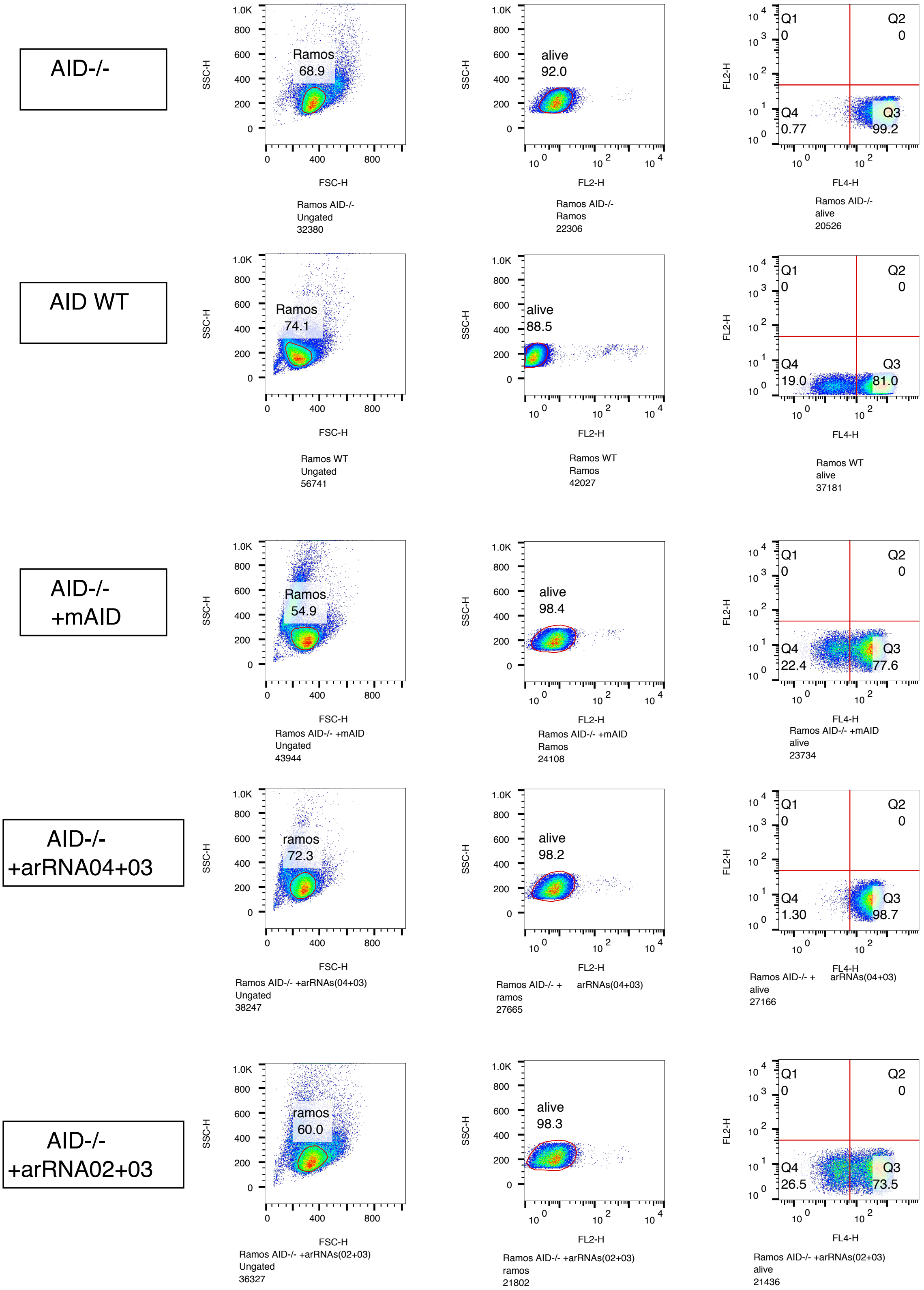
Table A3. List of plasmids. Plasmids are summarized by name, the product they express and under which promoter, what antibiotic resistance they have and if they have any other characteristics. The Addgene codes or other product number identifiers are provided in methods of the Aim 2, chapter 3.3.

Name	Product	Promoter	Resistance	Other characteristics
gRNA-pl	gRNA	U6	Kanamycin	
Crispr-pl	gRNA, pSpCas9	U6, CBh	Ampicillin	NLS, GFP
mAID-cDNA-pl	mAID	CMV (pcDNA3.1)	Ampicillin	
4λN-ADAR-pl	4λN-ADAR	CMV	Ampicillin	

9. Appendix D

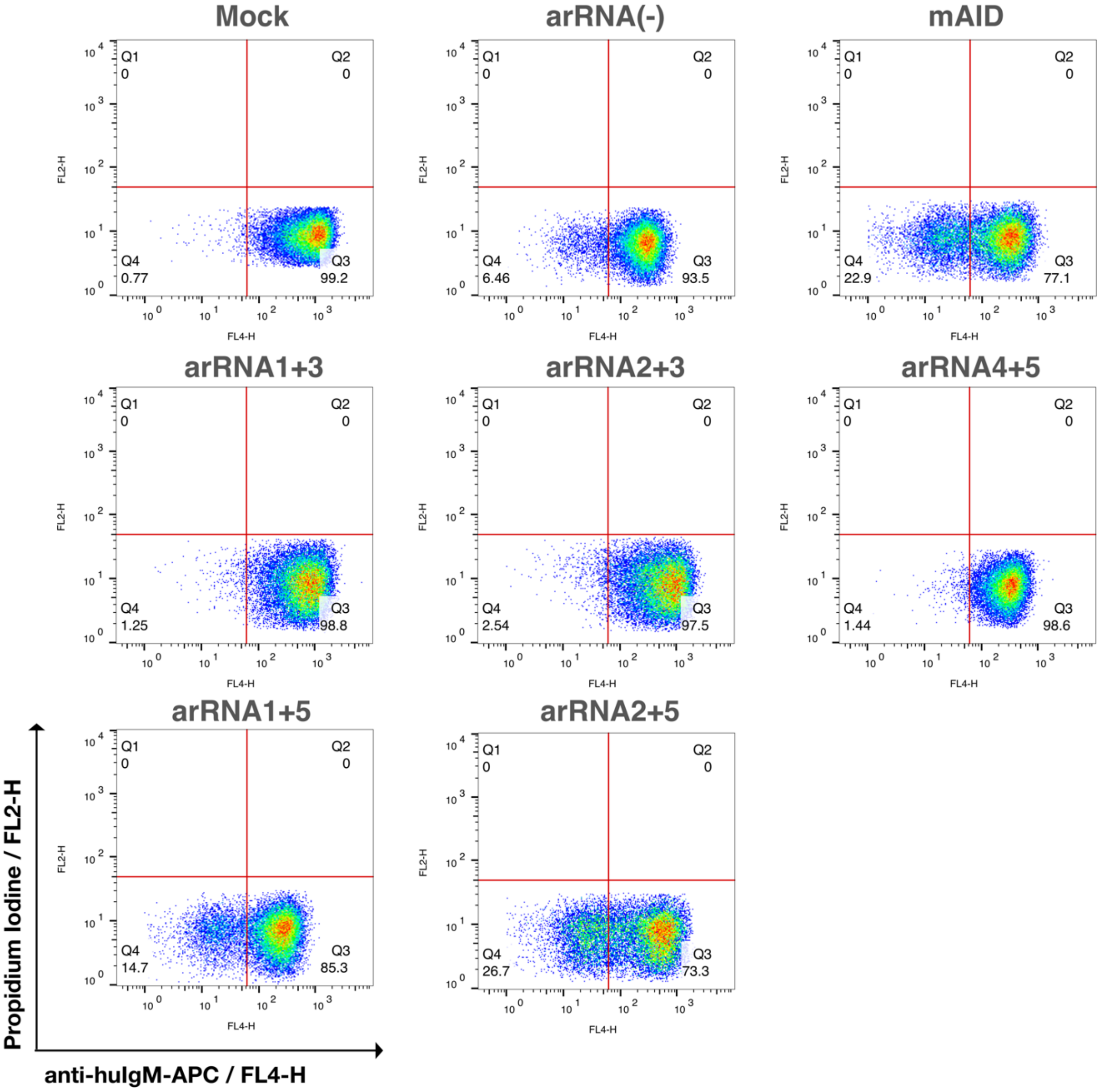
Part 1 - Gating strategy of Ramos WT and AID-/- cells.

Related information to chapters 3.3.5, 4.2.2 and 4.2.4

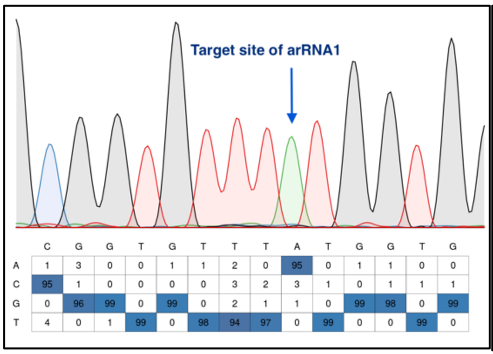


Part 2 - Loss of IgM after targeting the V$_H$ region with vector-expressing gRNAs

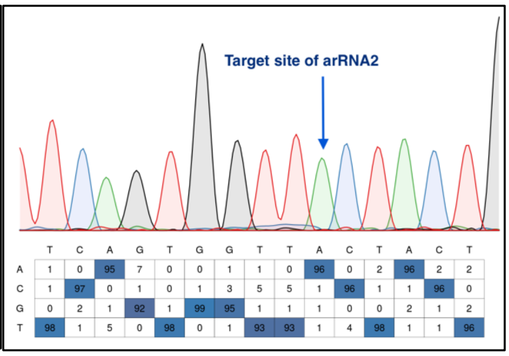
Related information to 3.3.5, and 4.2.4 and Figure 4.9



Part 3 – Chromatogram of arRNA(-) VH gDNA amplicon at the region of target arRNA1 and arRNA2



	C	G	G	T	G	T	T	T	A	T	G	G	T	G
A	1	3	0	0	1	1	2	0	95	0	1	1	0	0
C	95	1	0	0	0	0	3	2	3	1	0	1	1	1
G	0	96	99	0	99	0	2	1	1	0	99	98	0	99
T	4	0	1	99	0	98	94	97	0	99	0	0	99	0

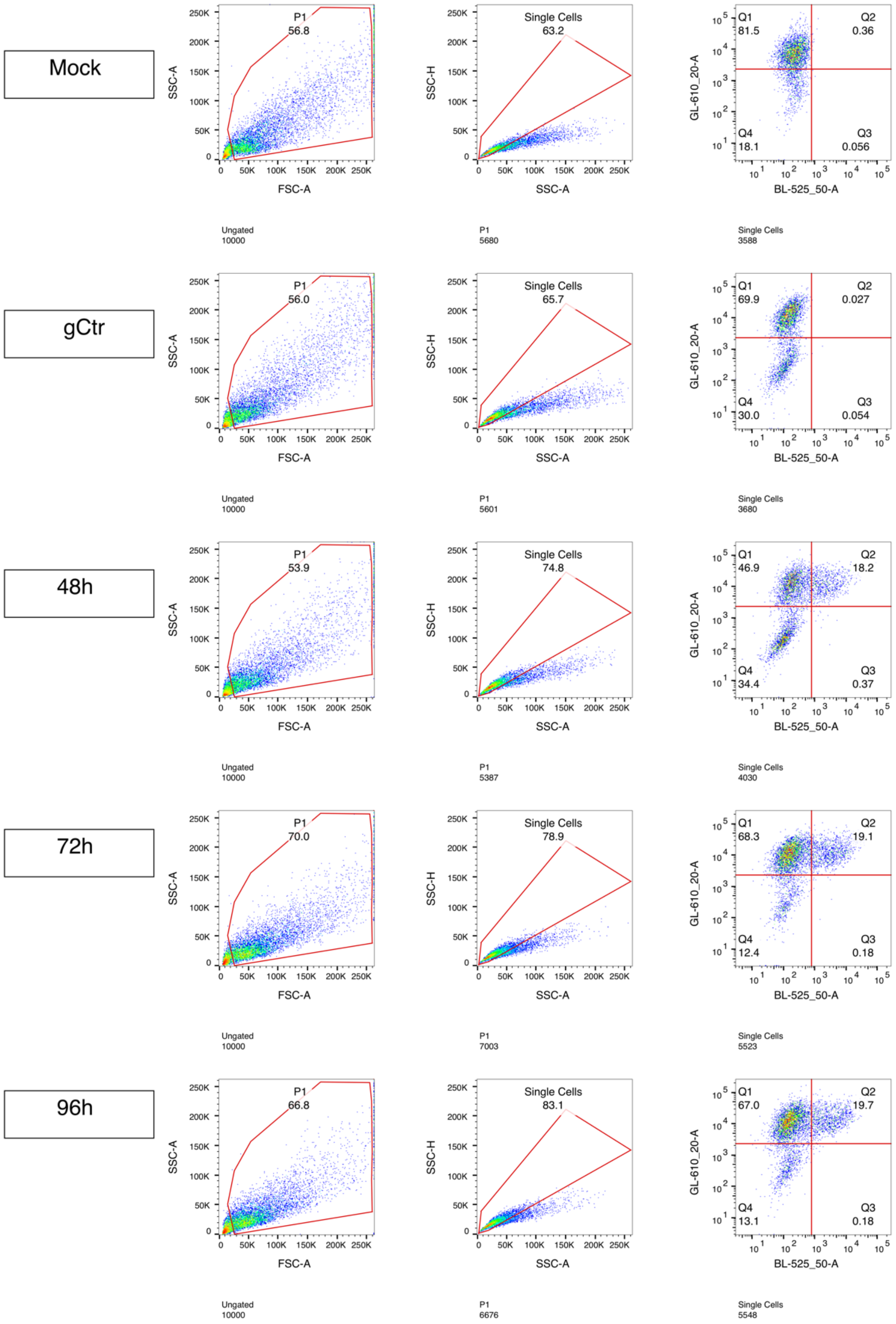


	T	C	A	G	T	G	G	T	T	A	C	T	A	C	T
A	1	0	95	7	0	0	1	1	0	96	0	2	96	2	2
C	1	97	0	1	0	1	3	5	5	1	96	1	1	96	0
G	0	2	1	92	1	99	95	1	1	1	0	0	2	1	2
T	98	1	5	0	98	0	1	93	93	1	4	98	1	1	96

The chromatogram on the left from Tasakis et al., 2020 and it is reused under the Creative Commons License 4.0. I produced the data and the illustration.

Part 4 – Gating strategy for HEK293T-W58X cells

Related information to chapters 4.2.5 and 4.2.6



List of publications

A summary of all the publications I was involved in as a PhD student supervised by Prof. Dr. Nina Papavasiliou is provided here below. The list includes peer-reviewed publications, preprints, reviews and book chapters I co-wrote through the findings from my main doctoral projects presented in this dissertation, but also through collaborations relevant to my doctoral research.

Research articles

Published in peer-reviewed journals

Tasakis, R.N., Samaras, G., Jamison, A., Lee, M., Paulus, A., Whitehouse, G., Verkoczy, L., Papavasiliou, F.N., Diaz, M., 2021. SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. PLOS ONE 16, e0255169. https://doi.org/10.1371/journal.pone.0255169

Kluesner, M., **Tasakis, R.N.**, Lerner, T., Arnold, A., Wüst, S., Binder, M., Webber, B.R., Moriarity, B.S., Pecori, R., 2021. MultiEditR: The first tool for detection and quantification of multiple RNA editing sites from Sanger sequencing demonstrates comparable fidelity to RNA-seq. Mol. Ther. - Nucleic Acids. https://doi.org/10.1016/j.omtn.2021.07.008

Stroppel, A.S., Latifi, N., Hanswillemenke, A., **Tasakis, R.N.**, Papavasiliou, F.N., Stafforst, T., 2021. Harnessing self-labeling enzymes for selective and concurrent A-to-I and C-to-U RNA base editing. Nucleic Acids Research. https://doi.org/10.1093/nar/gkab541

Preprints

Tasakis, R.N., Laganà, A., Stamkopoulou, D., Melnekoff, D.T., Nedumaran, P., Leshchenko, V., Pecori, R., Parekh, S., Papavasiliou, F.N., 2020. ADAR1 can drive Multiple Myeloma progression by acting both as an RNA editor of specific transcripts and as a DNA mutator of their cognate genes. bioRxiv 2020.02.11.943845. https://doi.org/10.1101/2020.02.11.943845

Under revision (unpublished)

Pecori, R., Ren, W., Wang, X., Berglund, M., Li, W., **Tasakis, R.N.**, Di Giorgio, S., Ye, X., Arnold, A., Wüst, S., Selvasaravanan, K.D., Fuell, F., Stafforst, T., Amini, R., Enblad, G., Sander, B., Wahlin, B., Zhang, H., Binder, M., Papavasiliou, F.N., Pan-Hammarström, Q. RNA-editing-initiated MAVS signaling is a key epitranscriptomic alteration in human B cell lymphoma. *In preparation*

Ren, M., Sidiropoulou, E., **Tasakis, R.N.**, Donato, E., Gonzalez Menendez, I., Busse, C., Luck, T.J., Dolnik, A., Bullinger, L., Trumpp, A., Quintanilla-Martinez, L., Kreuz, M., Hübschmann, D., Chapuy, B., Siebert, R., Papavasiliou, F.N., Sander, S. AID inactivation promotes APOBEC-mediated mutagenesis in germinal center lymphoma. *In preparation*

Reviews

Tasakis, R.N., Papavasiliou, F.N., Shaknovich, R, 2019. RNA Editors and DNA Mutators: Cancer Heterogeneity Through Sequence Diversification. OBM Genetics. http://dx.doi.org/10.21926/obm.genet.1902072

Book chapters

Lerner, T., Kluesner, M., **Tasakis, R.N.**, Moriarity, B.S., Papavasiliou, F.N., Pecori, R., 2021. C-to-U RNA Editing: From Computational Detection to Experimental Validation, in: Picardi, E., Pesole, G. (Eds.), RNA Editing: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 51–67. https://doi.org/10.1007/978-1-0716-0787-9_4

Casati, B., Stamkopoulou, D., **Tasakis, R.N.**, Pecori, R., 2021. ADAR-Mediated RNA Editing and Its Therapeutic Potentials, in: Jurga, S., Barciszewski, J. (Eds.), Epitranscriptomics, RNA Technologies. Springer International Publishing, Cham, pp. 471–503. https://doi.org/10.1007/978-3-030-71612-7_18

References

Abudayyeh, O.O., Gootenberg, J.S., Franklin, B., Koob, J., Kellner, M.J., Ladha, A., Joung, J., Kirchgatterer, P., Cox, D.B.T., Zhang, F., 2019. A cytosine deaminase for programmable single-base RNA editing. Science 365, 382–386. https://doi.org/10.1126/science.aax7063

Aguilera, A., García-Muse, T., 2012. R loops: from transcription byproducts to threats to genome stability. Mol. Cell 46, 115–124. https://doi.org/10.1016/j.molcel.2012.04.009

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., Islam, S.M.A., Lopez-Bigas, N., Klimczak, L.J., McPherson, J.R., Morganella, S., Sabarinathan, R., Wheeler, D.A., Mustonen, V., Getz, G., Rozen, S.G., Stratton, M.R., 2020. The repertoire of mutational signatures in human cancer. Nature 578, 94–101. https://doi.org/10.1038/s41586-020-1943-3

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., Boyault, S., Burkhardt, B., Butler, A.P., Caldas, C., Davies, H.R., Desmedt, C., Eils, R., Eyfjörd, J.E., Foekens, J.A., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D.T.W., Jones, D., Knappskog, S., Kool, M., Lakhani, S.R., López-Otín, C., Martin, S., Munshi, N.C., Nakamura, H., Northcott, P.A., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J.V., Puente, X.S., Raine, K., Ramakrishna, M., Richardson, A.L., Richter, J., Rosenstiel, P., Schlesner, M., Schumacher, T.N., Span, P.N., Teague, J.W., Totoki, Y., Tutt, A.N.J., Valdés-Mas, R., van Buuren, M.M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L.R., Zucman-Rossi, J., Andrew Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S.M., Siebert, R., Campo, E., Shibata, T., Pfister, S.M., Campbell, P.J., Stratton, M.R., 2013. Signatures of mutational processes in human cancer. Nature 500, 415–421. https://doi.org/10.1038/nature12477

Alexanian, R., Dimopoulos, M., 1994. The Treatment of Multiple Myeloma. N. Engl. J. Med. 330, 484–489. https://doi.org/10.1056/NEJM199402173300709

Al-Qaisi, T.S., Su, Y.-C., Roffler, S.R., 2018. Transient AID expression for in situ mutagenesis with improved cellular fitness. Sci. Rep. 8, 9413. https://doi.org/10.1038/s41598-018-27717-2

Amon, J.D., Koshland, D., 2016. RNase H enables efficient repair of R-loop induced DNA damage. eLife 5, e20533. https://doi.org/10.7554/eLife.20533

Anderson, K.C., Alsina, M., Bensinger, W., Biermann, J.S., Chanan-Khan, A., Cohen, A.D., Devine, S., Djulbegovic, B., Gasparetto, C., Huff, C.A., Jagasia, M., Medeiros, B.C., Meredith, R., Raje, N., Schriber, J., Singhal, S., Somlo, G., Stockerl-Goldstein, K., Tricot, G., Vose, J.M., Weber, D., Yahalom, J., Yunus, F., 2009. Multiple Myeloma. J. Natl. Compr. Canc. Netw. 7, 908–942. https://doi.org/10.6004/jnccn.2009.0061

Athanasiadis, A., Rich, A., Maas, S., 2004. Widespread A-to-I RNA Editing of Alu-Containing mRNAs in the Human Transcriptome. PLOS Biol. 2, e391. https://doi.org/10.1371/journal.pbio.0020391

Auwera, G.A.V. der, Carneiro, M.O., Hartl, C., Poplin, R., Angel, G. del, Levy-Moonshine, A., Jordan, T., Shakir, K., Roazen, D., Thibault, J., Banks, E., Garimella, K.V., Altshuler, D., Gabriel, S., DePristo, M.A., 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr. Protoc. Bioinforma. 43, 11.10.1-11.10.33. https://doi.org/10.1002/0471250953.bi1110s43

Balogh, E., Chandler, J.C., Varga, M., Tahoun, M., Menyhárd, D.K., Schay, G., Goncalves, T., Hamar, R., Légrádi, R., Szekeres, Á., Gribouval, O., Kleta, R., Stanescu, H., Bockenhauer, D., Kerti, A., Williams, H., Kinsler, V., Di, W.-L., Curtis, D., Kolatsi-Joannou, M., Hammid, H., Szőcs, A., Perczel, K., Maka, E., Toldi, G., Sava, F., Arrondel, C., Kardos, M., Fintha, A., Hossain, A., D'Arco, F., Kaliakatsos, M., Koeglmeier, J., Mifsud, W., Moosajee, M., Faro, A., Jávorszky, E., Rudas, G., Saied, M.H., Marzouk, S., Kelen, K., Götze, J., Reusz, G., Tulassay, T., Dragon, F., Mollet, G., Motameny, S., Thiele, H., Dorval, G., Nürnberg, P., Perczel, A., Szabó, A.J., Long, D.A., Tomita, K., Antignac, C., Waters, A.M., Tory, K., 2020. Pseudouridylation defect due to DKC1 and NOP10 mutations causes nephrotic syndrome with cataracts, hearing impairment, and enterocolitis. Proc. Natl. Acad. Sci. 117, 15137–15147. https://doi.org/10.1073/pnas.2002328117

Bazak, L., Levanon, E.Y., Eisenberg, E., 2014. Genome-wide analysis of Alu editability. Nucleic Acids Res. 42, 6876–6884. https://doi.org/10.1093/nar/gku414

Beck, B., Blanpain, C., 2013. Unravelling cancer stem cell potential. Nat. Rev. Cancer 13, 727–738. https://doi.org/10.1038/nrc3597

Blanc, V., Davidson, N.O., 2003. C-to-U RNA Editing: Mechanisms Leading to Genetic Diversity*. J. Biol. Chem. 278, 1395–1398. https://doi.org/10.1074/jbc.R200024200

Blanc, V., Henderson, J.O., Newberry, R.D., Xie, Y., Cho, S.-J., Newberry, E.P., Kennedy, S., Rubin, D.C., Wang, H.L., Luo, J., Davidson, N.O., 2007. Deletion of the AU-Rich RNA Binding Protein Apobec-1 Reduces Intestinal Tumor Burden in Apcmin Mice. Cancer Res. 67, 8565–8573. https://doi.org/10.1158/0008-5472.CAN-07-1593

Blanc, V., Xie, Y., Kennedy, S., Riordan, J.D., Rubin, D.C., Madison, B.B., Mills, J.C., Nadeau, J.H., Davidson, N.O., 2019. Apobec1 complementation factor (A1CF) and RBM47 interact in tissue-specific regulation of C to U RNA editing in mouse intestine and liver. RNA 25, 70–81. https://doi.org/10.1261/rna.068395.118

BLAST® Command Line Applications User Manual, 2008. . National Center for Biotechnology Information (US).

Bodenhofer, U., Bonatesta, E., Horejš-Kainrath, C., Hochreiter, S., 2015. msa: an R package for multiple sequence alignment. Bioinformatics 31, 3997–3999. https://doi.org/10.1093/bioinformatics/btv494

Bohnsack, K.E., Höbartner, C., Bohnsack, M.T., 2019. Eukaryotic 5-methylcytosine (m5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. Genes 10. https://doi.org/10.3390/genes10020102

Brusa, R., Zimmermann, F., Koh, D.S., Feldmeyer, D., Gass, P., Seeburg, P.H., Sprengel, R., 1995. Early-onset epilepsy and postnatal lethality associated with an editing-deficient GluR-B allele in mice. Science 270, 1677–1680. https://doi.org/10.1126/science.270.5242.1677

Budke, B., Kuzminov, A., 2006. Hypoxanthine Incorporation Is Nonmutagenic in Escherichia coli. J. Bacteriol. 188, 6553–6560. https://doi.org/10.1128/JB.00447-06

Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., Gilbert, W.V., 2014. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature 515, 143–146. https://doi.org/10.1038/nature13802

Casati, B., Stamkopoulou, D., Tasakis, R.N., Pecori, R., 2021. ADAR-Mediated RNA Editing and Its Therapeutic Potentials, in: Jurga, S., Barciszewski, J. (Eds.), Epitranscriptomics, RNA Technologies. Springer International Publishing, Cham, pp. 471–503. https://doi.org/10.1007/978-3-030-71612-7_18

Chalmers, Z.R., Connelly, C.F., Fabrizio, D., Gay, L., Ali, S.M., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D.S., Roels, S., White, J., Otto, G.A., Ross, J.S., Garraway, L., Miller, V.A., Stephens, P.J., Frampton, G.M., 2017. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. Genome Med. 9, 34. https://doi.org/10.1186/s13073-017-0424-2

Chan, J.F.-W., Kok, K.-H., Zhu, Z., Chu, H., To, K.K.-W., Yuan, S., Yuen, K.-Y., 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg. Microbes Infect. 9, 221–236. https://doi.org/10.1080/22221751.2020.1719902

Chen, C.X., Cho, D.S., Wang, Q., Lai, F., Carter, K.C., Nishikura, K., 2000. A third member of the RNA-specific adenosine deaminase gene family, ADAR3, contains both single- and double-stranded RNA binding domains. RNA 6, 755–767.

Chen, J., MacCarthy, T., 2017. The preferred nucleotide contexts of the AID/APOBEC cytidine deaminases have differential effects when mutating retrotransposon and virus sequences compared to host genes. PLoS Comput. Biol. 13. https://doi.org/10.1371/journal.pcbi.1005471

Chen, J., Miller, B.F., Furano, A.V., 2014. Repair of naturally occurring mismatches can induce mutations in flanking DNA. eLife 3, e02001. https://doi.org/10.7554/eLife.02001

Chen, J.-Y., Zhang, X., Fu, X.-D., Chen, L., 2019. R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. Nat. Protoc. 14, 1661–1685. https://doi.org/10.1038/s41596-019-0154-6

Chen, L., Li, Y., Lin, C.H., Chan, T.H.M., Chow, R.K.K., Song, Y., Liu, M., Yuan, Y.-F., Fu, L., Kong, K.L., Qi, L., Li, Y., Zhang, N., Tong, A.H.Y., Kwong, D.L.-W., Man, K., Lo, C.M., Lok, S., Tenen, D.G., Guan, X.-Y., 2013. Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. Nat. Med. 19, 209–216. https://doi.org/10.1038/nm.3043

Chester, A., Somasekaram, A., Tzimina, M., Jarmuz, A., Gisbourne, J., O'Keefe, R., Scott, J., Navaratnam, N., 2003. The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. EMBO J. 22, 3971–3982. https://doi.org/10.1093/emboj/cdg369

Chiu, Y.-L., Greene, W.C., 2008. The APOBEC3 Cytidine Deaminases: An Innate Defensive Network Opposing Exogenous Retroviruses and Endogenous Retroelements. Annu. Rev. Immunol. 26, 317–353. https://doi.org/10.1146/annurev.immunol.26.021607.090350

Chng, W.J., Glebov, O., Bergsagel, P.L., Kuehl, W.M., 2007. Genetic events in the pathogenesis of multiple myeloma. Best Pract. Res. Clin. Haematol. 20, 571–596. https://doi.org/10.1016/j.beha.2007.08.004

Cho, D.-S.C., Yang, W., Lee, J.T., Shiekhattar, R., Murray, J.M., Nishikura, K., 2003. Requirement of dimerization for RNA editing activity of adenosine deaminases acting on RNA. J. Biol. Chem. 278, 17093–17102. https://doi.org/10.1074/jbc.M213127200

Chung, H., Calis, J.J.A., Wu, X., Sun, T., Yu, Y., Sarbanes, S.L., Dao Thi, V.L., Shilvock, A.R., Hoffmann, H.-H., Rosenberg, B.R., Rice, C.M., 2018. Human ADAR1 Prevents Endogenous RNA from Triggering Translational Shutdown. Cell 172, 811-824.e14. https://doi.org/10.1016/j.cell.2017.12.038

Conticello, S.G., Langlois, M., Yang, Z., Neuberger, M.S., 2007. DNA Deamination in Immunity: AID in the Context of Its APOBEC Relatives, in: Advances in Immunology, AID for Immunoglobulin Diversity. Academic Press, pp. 37–73. https://doi.org/10.1016/S0065-2776(06)94002-4

Conticello, S.G., Thomas, C.J.F., Petersen-Mahrt, S.K., Neuberger, M.S., 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases. Mol. Biol. Evol. 22, 367–377. https://doi.org/10.1093/molbev/msi026

Cook, A.J.L., Raftery, J.M., Lau, K.K.E., Jessup, A., Harris, R.S., Takeda, S., Jolly, C.J., 2007. DNA-Dependent Protein Kinase Inhibits AID-Induced Antibody Gene Conversion. PLOS Biol. 5, e80. https://doi.org/10.1371/journal.pbio.0050080

Corre, J., Cleynen, A., Robiou du Pont, S., Buisson, L., Bolli, N., Attal, M., Munshi, N., Avet-Loiseau, H., 2018. Multiple myeloma clonal evolution in homogeneously treated patients. Leukemia 32, 2636–2647. https://doi.org/10.1038/s41375-018-0153-6

Cowling, V.H., 2019. CAPAM: The mRNA Cap Adenosine N6-Methyltransferase. Trends Biochem. Sci. 44, 183–185. https://doi.org/10.1016/j.tibs.2019.01.002

Cox, D.B.T., Gootenberg, J.S., Abudayyeh, O.O., Franklin, B., Kellner, M.J., Joung, J., Zhang, F., 2017. RNA editing with CRISPR-Cas13. Science 358, 1019–1027. https://doi.org/10.1126/science.aaq0180

Cucinotta, D., Vanelli, M., 2020. WHO Declares COVID-19 a Pandemic. Acta Bio Medica Atenei Parm. 91, 157–160. https://doi.org/10.23750/abm.v91i1.9397

Danovi, D., Meulmeester, E., Pasini, D., Migliorini, D., Capra, M., Frenk, R., Graaf, P. de, Francoz, S., Gasparini, P., Gobbi, A., Helin, K., Pelicci, P.G., Jochemsen, A.G., Marine, J.-C., 2004. Amplification of Mdmx (or Mdm4) Directly Contributes to Tumor Formation by Inhibiting p53 Tumor Suppressor Activity. Mol. Cell. Biol. 24, 5835–5843. https://doi.org/10.1128/MCB.24.13.5835-5843.2004

D'Antonio, M., Tamayo, P., Mesirov, J.P., Frazer, K.A., 2016. Kataegis Expression Signature in Breast Cancer Is Associated with Late Onset, Better Prognosis, and Higher HER2 Levels. Cell Rep. 16, 672–683. https://doi.org/10.1016/j.celrep.2016.06.026

Darby, A.C., Hiscox, J.A., 2021. Covid-19: variants and vaccination. BMJ 372, n771. https://doi.org/10.1136/bmj.n771

Davies, N.G., Abbott, S., Barnard, R.C., Jarvis, C.I., Kucharski, A.J., Munday, J.D., Pearson, C.A.B., Russell, T.W., Tully, D.C., Washburne, A.D., Wenseleers, T., Gimma, A., Waites, W., Wong, K.L.M., Zandvoort, K. van, Silverman, J.D., Group1‡, C.C.-19 W., Consortium‡, C.-19 G.U. (COG-U., Diaz-Ordaz, K., Keogh, R., Eggo, R.M., Funk, S., Jit, M., Atkins, K.E., Edmunds, W.J., 2021. Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England. Science 372. https://doi.org/10.1126/science.abg3055

Dehghanifard, A., Kaviani, S., Abroun, S., Mehdizadeh, M., Saiedi, S., Maali, A., Ghaffari, S., Azad, M., 2018. Various Signaling Pathways in Multiple Myeloma Cells and Effects of Treatment on These Pathways. Clin. Lymphoma Myeloma Leuk. 18, 311–320. https://doi.org/10.1016/j.clml.2018.03.007

Delaunay, S., Frye, M., 2019. RNA modifications regulating cell fate in cancer. Nat. Cell Biol. 21, 552–559. https://doi.org/10.1038/s41556-019-0319-0

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., McKenna, A., Fennell, T.J., Kernytsky, A.M., Sivachenko, A.Y., Cibulskis, K., Gabriel, S.B., Altshuler, D., Daly, M.J., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43, 491–498. https://doi.org/10.1038/ng.806

Dickerson, S.K., Market, E., Besmer, E., Papavasiliou, F.N., 2003. AID Mediates Hypermutation by Deaminating Single Stranded DNA. J. Exp. Med. 197, 1291–1296. https://doi.org/10.1084/jem.20030481

Diroma, M.A., Ciaccia, L., Pesole, G., Picardi, E., 2019. Elucidating the editome: bioinformatics approaches for RNA editing detection. Brief. Bioinform. 20, 436–447. https://doi.org/10.1093/bib/bbx129

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinforma. Oxf. Engl. 29, 15–21. https://doi.org/10.1093/bioinformatics/bts635

Dominguez, D., Freese, P., Alexis, M.S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N.J., Van Nostrand, E.L., Pratt, G.A., Yeo, G.W., Graveley, B.R., Burge, C.B., 2018. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. Mol. Cell 70, 854-867.e9. https://doi.org/10.1016/j.molcel.2018.05.001

Doria, M., Neri, F., Gallo, A., Farace, M.G., Michienzi, A., 2009. Editing of HIV-1 RNA by the double-stranded RNA deaminase ADAR1 stimulates viral infection. Nucleic Acids Res. 37, 5848–5858. https://doi.org/10.1093/nar/gkp604

Eggington, J.M., Greene, T., Bass, B.L., 2011. Predicting sites of ADAR editing in double-stranded RNA. Nat. Commun. 2, 319. https://doi.org/10.1038/ncomms1324

Ekdahl, Y., Farahani, H.S., Behm, M., Lagergren, J., Öhman, M., 2012. A-to-I editing of microRNAs in the mammalian brain increases during development. Genome Res. 22, 1477–1487. https://doi.org/10.1101/gr.131912.111

El Yacoubi, B., Bailly, M., de Crécy-Lagard, V., 2012. Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs. Annu. Rev. Genet. 46, 69–95. https://doi.org/10.1146/annurev-genet-110711-155641

Fakhri, B., Vij, R., 2016. Clonal Evolution in Multiple Myeloma. Clin. Lymphoma Myeloma Leuk. 16 Suppl, S130-134. https://doi.org/10.1016/j.clml.2016.02.025

Fisher, R., Pusztai, L., Swanton, C., 2013. Cancer heterogeneity: implications for targeted therapeutics. Br. J. Cancer 108, 479–485. https://doi.org/10.1038/bjc.2012.581

Fossat, N., Tourle, K., Radziewic, T., Barratt, K., Liebhold, D., Studdert, J.B., Power, M., Jones, V., Loebel, D.A.F., Tam, P.P.L., 2014. C to U RNA editing mediated by APOBEC1 requires RNA-binding protein RBM47. EMBO Rep. 15, 903–910. https://doi.org/10.15252/embr.201438450

Frye, M., Harada, B.T., Behm, M., He, C., 2018. RNA modifications modulate gene expression during development. Science 361, 1346–1349. https://doi.org/10.1126/science.aau1646

Fu, L., Qin, Y.-R., Ming, X.-Y., Zuo, X.-B., Diao, Y.-W., Zhang, L.-Y., Ai, J., Liu, B.-L., Huang, T.-X., Cao, T.-T., Tan, B.-B., Xiang, D., Zeng, C.-M., Gong, J., Zhang, Q., Dong, S.-S., Chen, J., Liu, H., Wu, J.-L., Qi, R.Z., Xie, D., Wang, L.-D., Guan, X.-Y., 2017. RNA editing of SLC22A3 drives early tumor invasion and metastasis in familial esophageal cancer. Proc. Natl. Acad. Sci. 114, E4631–E4640. https://doi.org/10.1073/pnas.1703178114

Fumagalli, D., Gacquer, D., Rothé, F., Lefort, A., Libert, F., Brown, D., Kheddoumi, N., Shlien, A., Konopka, T., Salgado, R., Larsimont, D., Polyak, K., Willard-Gallo, K., Desmedt, C., Piccart, M., Abramowicz, M., Campbell, P.J., Sotiriou, C., Detours, V., 2015. Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. Cell Rep. 13, 277–289. https://doi.org/10.1016/j.celrep.2015.09.032

Gal-Ben-Ari, S., Barrera, I., Ehrlich, M., Rosenblum, K., 2019. PKR: A Kinase to Remember. Front. Mol. Neurosci. 11. https://doi.org/10.3389/fnmol.2018.00480

Gallaher, W.R., 2020. A palindromic RNA sequence as a common breakpoint contributor to copy-choice recombination in SARS-COV-2. Arch. Virol. 165, 2341–2348. https://doi.org/10.1007/s00705-020-04750-z

George, C.X., Li, Z., Okonski, K.M., Toth, A.M., Wang, Y., Samuel, C.E., 2009. Tipping the balance: antagonism of PKR kinase and ADAR1 deaminase functions by virus gene products. J. Interferon Cytokine Res. Off. J. Int. Soc. Interferon Cytokine Res. 29, 477–487. https://doi.org/10.1089/jir.2009.0065

George, C.X., Samuel, C.E., 1999. 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. 96, 4621–4626.

George, C.X., Wagner, M.V., Samuel, C.E., 2005. Expression of Interferon-inducible RNA Adenosine Deaminase ADAR1 during Pathogen Infection and Mouse Embryo Development Involves Tissue-selective Promoter Utilization and Alternative Splicing*. J. Biol. Chem. 280, 15020–15028. https://doi.org/10.1074/jbc.M500476200

Giorgio, S.D., Martignano, F., Torcia, M.G., Mattiuz, G., Conticello, S.G., 2020. Evidence for host-dependent RNA editing in the transcriptome of SARS-CoV-2. Sci. Adv. 6, eabb5813. https://doi.org/10.1126/sciadv.abb5813

Greaves, M., Maley, C.C., 2012. Clonal evolution in cancer. Nature 481, 306–313. https://doi.org/10.1038/nature10762

Han, L., Diao, L., Yu, S., Xu, X., Li, Jie, Zhang, R., Yang, Y., Werner, H.M.J., Eterovic, A.K., Yuan, Y., Li, Jun, Nair, N., Minelli, R., Tsang, Y.H., Cheung, L.W.T., Jeong, K.J., Roszik, J., Ju, Z., Woodman, S.E., Lu, Y., Scott, K.L., Li, J.B., Mills, G.B., Liang, H., 2015. The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. Cancer Cell 28, 515–528. https://doi.org/10.1016/j.ccell.2015.08.013

Hannon, G.J., 2002. RNA interference. Nature 418, 244–251. https://doi.org/10.1038/418244a

Harjanto, D., Papamarkou, T., Oates, C.J., Rayon-Estrada, V., Papavasiliou, F.N., Papavasiliou, A., 2016. RNA editing generates cellular subsets with diverse sequence within populations. Nat. Commun. 7, 12145. https://doi.org/10.1038/ncomms12145

Harris, R.S., Petersen-Mahrt, S.K., Neuberger, M.S., 2002. RNA Editing Enzyme APOBEC1 and Some of Its Homologs Can Act as DNA Mutators. Mol. Cell 10, 1247–1253. https://doi.org/10.1016/S1097-2765(02)00742-6

Hartner, J.C., Schmittwolf, C., Kispert, A., Müller, A.M., Higuchi, M., Seeburg, P.H., 2004. Liver disintegration in the mouse embryo caused by deficiency in the RNA-editing enzyme ADAR1. J. Biol. Chem. 279, 4894–4902. https://doi.org/10.1074/jbc.M311347200

Hartner, J.C., Walkley, C.R., Lu, J., Orkin, S.H., 2009. ADAR1 is essential for the maintenance of hematopoiesis and suppression of interferon signaling. Nat. Immunol. 10, 109–115. https://doi.org/10.1038/ni.1680

Hassan, S.S., Choudhury, P.P., Uversky, V.N., Dayhoff, G.W., Aljabali, A.A.A., Uhal, B.D., Lundstrom, K., Rezaei, N., Seyran, M., Pizzol, D., Adadi, P., Lal, A., Soares, A., El-Aziz, T.M.A., Kandimalla, R., Tambuwala, M., Azad, G.K., Sherchan, S.P., Baetas-da-Cruz, W., Takayama, K., Serrano-Aroca, Á., Chauhan, G., Palu, G., Brufsky, A.M., 2020. Variability of Accessory Proteins Rules the SARS-CoV-2 Pathogenicity. bioRxiv 2020.11.06.372227. https://doi.org/10.1101/2020.11.06.372227

Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., Seeburg, P.H., 2000. Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature 406, 78–81. https://doi.org/10.1038/35017558

Hoernes, T.P., Erlacher, M.D., 2017. Translating the epitranscriptome. WIREs RNA 8, e1375. https://doi.org/10.1002/wrna.1375

Hong, H., An, O., Chan, T.H.M., Ng, V.H.E., Kwok, H.S., Lin, J.S., Qi, L., Han, J., Tay, D.J.T., Tang, S.J., Yang, H., Song, Y., Bellido Molias, F., Tenen, D.G., Chen, L., 2018. Bidirectional regulation of adenosine-to-inosine (A-to-I) RNA editing by DEAH box helicase 9 (DHX9) in cancer. Nucleic Acids Res. 46, 7953–7969. https://doi.org/10.1093/nar/gky396

Hou, Y.J., Chiba, S., Halfmann, P., Ehre, C., Kuroda, M., Dinnon, K.H., Leist, S.R., Schäfer, A., Nakajima, N., Takahashi, K., Lee, R.E., Mascenik, T.M., Graham, R., Edwards, C.E., Tse, L.V., Okuda, K., Markmann, A.J., Bartelt, L., de Silva, A., Margolis, D.M., Boucher, R.C., Randell, S.H., Suzuki, T., Gralinski, L.E., Kawaoka, Y., Baric, R.S., 2020. SARS-CoV-2 D614G variant exhibits efficient replication ex vivo and transmission in vivo. Science 370, 1464–1468. https://doi.org/10.1126/science.abe8499

Hu, B., Guo, H., Zhou, P., Shi, Z.-L., 2021. Characteristics of SARS-CoV-2 and COVID-19. Nat. Rev. Microbiol. 19, 141–154. https://doi.org/10.1038/s41579-020-00459-7

Huang, Xinxin, Lv, J., Li, Y., Mao, S., Li, Z., Jing, Z., Sun, Y., Zhang, X., Shen, S., Wang, X., Di, M., Ge, J., Huang, Xingxu, Zuo, E., Chi, T., 2020. Programmable C-to-U RNA editing using the human APOBEC3A deaminase. EMBO J. 39, e104741. https://doi.org/10.15252/embj.2020104741

Inoue, H., Nojima, H., Okayama, H., 1990. High efficiency transformation of Escherichia coli with plasmids. Gene 96, 23–28. https://doi.org/10.1016/0378-1119(90)90336-P

Iorio, F., Garcia-Alonso, L., Brammeld, J.S., Martincorena, I., Wille, D.R., McDermott, U., Saez-Rodriguez, J., 2018. Pathway-based dissection of the genomic heterogeneity of cancer hallmarks' acquisition with SLAPenrich. Sci. Rep. 8, 6713. https://doi.org/10.1038/s41598-018-25076-6

Jalili, P., Bowen, D., Langenbucher, A., Park, S., Aguirre, K., Corcoran, R.B., Fleischman, A.G., Lawrence, M.S., Zou, L., Buisson, R., 2020. Quantification of ongoing APOBEC3A activity in tumor cells by monitoring RNA editing at hotspots. Nat. Commun. 11, 2971. https://doi.org/10.1038/s41467-020-16802-8

Janku, F., 2014. Tumor heterogeneity in the clinic: is it a real problem? Ther. Adv. Med. Oncol. 6, 43–51. https://doi.org/10.1177/1758834013517414

Jarvis, M.C., Ebrahimi, D., Temiz, N.A., Harris, R.S., 2018. Mutation Signatures Including APOBEC in Cancer Cell Lines. JNCI Cancer Spectr. 2. https://doi.org/10.1093/jncics/pky002

Katrekar, D., Yen, J., Xiang, Y., Saha, A., Meluzzi, D., Savva, Y., Mali, P., 2021. Robust RNA editing via recruitment of endogenous ADARs using circular guide RNAs. bioRxiv 2021.01.12.426286. https://doi.org/10.1101/2021.01.12.426286

Kawahara, Y., Kwak, S., Sun, H., Ito, K., Hashida, H., Aizawa, H., Jeong, S.-Y., Kanazawa, I., 2003. Human spinal motoneurons express low relative abundance of GluR2 mRNA: an implication for excitotoxicity in ALS. J. Neurochem. 85, 680–689. https://doi.org/10.1046/j.1471-4159.2003.01703.x

Keegan, L.P., Gallo, A., O'Connell, M.A., 2001. The many roles of an RNA editor. Nat. Rev. Genet. 2, 869–878. https://doi.org/10.1038/35098584

Kellner, M.J., Koob, J.G., Gootenberg, J.S., Abudayyeh, O.O., Zhang, F., 2019. SHERLOCK: nucleic acid detection with CRISPR nucleases. Nat. Protoc. 14, 2986–3012. https://doi.org/10.1038/s41596-019-0210-2

Kemp, S.A., Collier, D.A., Datir, R.P., Ferreira, I.A.T.M., Gayed, S., Jahun, A., Hosmillo, M., Rees-Spear, C., Mlcochova, P., Lumb, I.U., Roberts, D.J., Chandra, A., Temperton, N., Sharrocks, K., Blane, E., Modis, Y., Leigh, K.E., Briggs, J.A.G., Gils, M.J. van, Smith, K.G.C., Bradley, J.R., Smith, C., Doffinger, R., Ceron-Gutierrez, L., Barcenas-Morales, G., Pollock, D.D., Goldstein, R.A., Smielewska, A., Skittrall, J.P., Gouliouris, T., Goodfellow, I.G., Gkrania-Klotsas, E., Illingworth, C.J.R., McCoy, L.E., Gupta, R.K., 2021. SARS-CoV-2 evolution during treatment of chronic infection. Nature 592, 277–282. https://doi.org/10.1038/s41586-021-03291-y

Kim, D., Lee, J.-Y., Yang, J.-S., Kim, J.W., Kim, V.N., Chang, H., 2020. The Architecture of SARS-CoV-2 Transcriptome. Cell 181, 914-921.e10. https://doi.org/10.1016/j.cell.2020.04.011

Kim, D.D.Y., Kim, T.T.Y., Walsh, T., Kobayashi, Y., Matise, T.C., Buyske, S., Gabriel, A., 2004. Widespread RNA Editing of Embedded Alu Elements in the Human Transcriptome. Genome Res. 14, 1719–1725. https://doi.org/10.1101/gr.2855504

Kim, S., Scheffler, K., Halpern, A.L., Bekritsky, M.A., Noh, E., Källberg, M., Chen, X., Kim, Y., Beyter, D., Krusche, P., Saunders, C.T., 2018. Strelka2: fast and accurate calling of germline and somatic variants. Nat. Methods 15, 591–594. https://doi.org/10.1038/s41592-018-0051-x

Kiran, A.M., O'Mahony, J.J., Sanjeev, K., Baranov, P.V., 2013. Darned in 2013: inclusion of model organisms and linking with Wikipedia. Nucleic Acids Res. 41, D258-261. https://doi.org/10.1093/nar/gks961

Klimczak, L.J., Randall, T.A., Saini, N., Li, J.-L., Gordenin, D.A., 2020. Similarity between mutation spectra in hypermutated genomes of rubella virus and in SARS-CoV-2 genomes accumulated during the COVID-19 pandemic. PLOS ONE 15, e0237689. https://doi.org/10.1371/journal.pone.0237689

Kluesner, M., Tasakis, R.N., Lerner, T., Arnold, A., Wüst, S., Binder, M., Webber, B.R., Moriarity, B.S., Pecori, R., 2021. MultiEditR: The first tool for detection and quantification of multiple RNA editing sites from Sanger sequencing demonstrates comparable fidelity to RNA-seq. Mol. Ther. - Nucleic Acids. https://doi.org/10.1016/j.omtn.2021.07.008

Kluesner, M.G., Nedveck, D.A., Lahr, W.S., Garbe, J.R., Abrahante, J.E., Webber, B.R., Moriarity, B.S., 2018. EditR: A Method to Quantify Base Editing from Sanger Sequencing. CRISPR J. 1, 239–250. https://doi.org/10.1089/crispr.2018.0014

Knudson, A.G., 1971. Mutation and Cancer: Statistical Study of Retinoblastoma. Proc. Natl. Acad. Sci. 68, 820–823. https://doi.org/10.1073/pnas.68.4.820

Kono, M., Suganuma, M., Akiyama, M., Ito, Y., Ujiie, H., Morimoto, K., 2014. Novel ADAR1 mutations including a single amino acid deletion in the deaminase domain underlie dyschromatosis symmetrica hereditaria in Japanese families. Int. J. Dermatol. 53, e194-196. https://doi.org/10.1111/j.1365-4632.2012.05765.x

Lagana, A., Melnekoff, D., Beno, I., Leshchenko, V., Perumal, D., Keats, J.J., DeRome, M., Yesil, J., Auclair, D., Madduri, D., Chari, A., Cho, H.J., Barlogie, B., Jagannath, S., Dudley, J., Parekh, S., 2017. Clonal Evolution in Newly Diagnosed Multiple Myeloma Patients: A Follow-up Study from the Mmrf Commpass Genomics Project. Blood 130, 325. https://doi.org/10.1182/blood.V130.Suppl_1.325.325

Laha, S., Chakraborty, J., Das, S., Manna, S.K., Biswas, S., Chatterjee, R., 2020. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. Infect. Genet. Evol. 85, 104445. https://doi.org/10.1016/j.meegid.2020.104445

Lai, Z., Markovets, A., Ahdesmaki, M., Chapman, B., Hofmann, O., McEwen, R., Johnson, J., Dougherty, B., Barrett, J.C., Dry, J.R., 2016. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 44, e108–e108. https://doi.org/10.1093/nar/gkw227

Lamers, M.M., van den Hoogen, B.G., Haagmans, B.L., 2019. ADAR1: "Editor-in-Chief" of Cytoplasmic Innate Immunity. Front. Immunol. 10. https://doi.org/10.3389/fimmu.2019.01763

Lan, J., Rajan, N., Bizet, M., Penning, A., Singh, N.K., Guallar, D., Calonne, E., Li Greci, A., Bonvin, E., Deplus, R., Hsu, P.J., Nachtergaele, S., Ma, C., Song, R., Fuentes-Iglesias, A., Hassabi, B., Putmans, P., Mies, F., Menschaert, G., Wong, J.J.L., Wang, J., Fidalgo, M., Yuan, B., Fuks, F., 2020. Functional role of Tet-mediated RNA hydroxymethylcytosine in mouse ES cells and during differentiation. Nat. Commun. 11, 4956. https://doi.org/10.1038/s41467-020-18729-6

Laurencikiene, J., Källman, A.M., Fong, N., Bentley, D.L., Öhman, M., 2006. RNA editing and alternative splicing: the importance of co-transcriptional coordination. EMBO Rep. 7, 303–307. https://doi.org/10.1038/sj.embor.7400621

Lazzari, E., Mondala, P.K., Santos, N.D., Miller, A.C., Pineda, G., Jiang, Q., Leu, H., Ali, S.A., Ganesan, A.-P., Wu, C.N., Costello, C., Minden, M., Chiaramonte, R., Stewart, A.K., Crews, L.A., Jamieson, C.H.M., 2017. Alu -dependent RNA editing of GLI1 promotes malignant regeneration in multiple myeloma. Nat. Commun. 8, 1922. https://doi.org/10.1038/s41467-017-01890-w

Lecossier, D., Bouchonnet, F., Clavel, F., Hance, A.J., 2003. Hypermutation of HIV-1 DNA in the Absence of the Vif Protein. Science 300, 1112–1112. https://doi.org/10.1126/science.1083338

Lerner, T., Kluesner, M., Tasakis, R.N., Moriarity, B.S., Papavasiliou, F.N., Pecori, R., 2021. C-to-U RNA Editing: From Computational Detection to Experimental Validation, in: Picardi, E., Pesole, G. (Eds.), RNA Editing: Methods and Protocols, Methods in Molecular Biology. Springer US, New York, NY, pp. 51–67. https://doi.org/10.1007/978-1-0716-0787-9_4

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., Olshansky, M., Rechavi, G., Jantsch, M.F., 2004. Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. 22, 1001–1005. https://doi.org/10.1038/nbt996

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, J.B., Levanon, E.Y., Yoon, J.-K., Aach, J., Xie, B., LeProust, E., Zhang, K., Gao, Y., Church, G.M., 2009. Genome-Wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing. Science 324, 1210–1213. https://doi.org/10.1126/science.1170995

Li, T., Yang, X., Li, W., Song, J., Li, Z., Zhu, X., Wu, X., Liu, Y., 2021. ADAR1 Stimulation by IFN-α Downregulates the Expression of MAVS via RNA Editing to Regulate the Anti-HBV Response. Mol. Ther. 29, 1335–1348. https://doi.org/10.1016/j.ymthe.2020.11.031

Libin, P.J.K., Deforche, K., Abecasis, A.B., Theys, K., 2019. VIRULIGN: fast codon-correct alignment and annotation of viral genomes. Bioinformatics 35, 1763–1765. https://doi.org/10.1093/bioinformatics/bty851

Licht, K., Kapoor, U., Amman, F., Picardi, E., Martin, D., Bajad, P., Jantsch, M.F., 2019. A high resolution A-to-I editing map in the mouse identifies editing events controlled by pre-mRNA splicing. Genome Res. 29, 1453–1463. https://doi.org/10.1101/gr.242636.118

Liddicoat, B.J., Chalk, A.M., Walkley, C.R., 2016. ADAR1, inosine and the immune sensing system: distinguishing self from non-self. WIREs RNA 7, 157–172. https://doi.org/10.1002/wrna.1322

Liddicoat, B.J., Piskol, R., Chalk, A.M., Ramaswami, G., Higuchi, M., Hartner, J.C., Li, J.B., Seeburg, P.H., Walkley, C.R., 2015. RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. Science 349, 1115–1120. https://doi.org/10.1126/science.aac7049

Liu, M.-C., Liao, W.-Y., Buckley, K.M., Yang, S.Y., Rast, J.P., Fugmann, S.D., 2018. AID/APOBEC-like cytidine deaminases are ancient innate immune mediators in invertebrates. Nat. Commun. 9, 1948. https://doi.org/10.1038/s41467-018-04273-x

Livneh, I., Moshitch-Moshkovitz, S., Amariglio, N., Rechavi, G., Dominissini, D., 2020. The m 6 A epitranscriptome: transcriptome plasticity in brain development and function. Nat. Rev. Neurosci. 21, 36–51. https://doi.org/10.1038/s41583-019-0244-z

Lo Giudice, C., Silvestris, D.A., Roth, S.H., Eisenberg, E., Pesole, G., Gallo, A., Picardi, E., 2020a. Quantifying RNA Editing in Deep Transcriptome Datasets. Front. Genet. 11, 194. https://doi.org/10.3389/fgene.2020.00194

Lo Giudice, C., Tangaro, M.A., Pesole, G., Picardi, E., 2020b. Investigating RNA editing in deep transcriptome datasets with REDItools and REDIportal. Nat. Protoc. 15, 1098–1131. https://doi.org/10.1038/s41596-019-0279-7

Longerich, S., Basu, U., Alt, F., Storb, U., 2006. AID in somatic hypermutation and class switch recombination. Curr. Opin. Immunol., Lymphocyte development / Tumour immunology 18, 164–174. https://doi.org/10.1016/j.coi.2006.01.008

Luan, J., Lu, Y., Jin, X., Zhang, L., 2020. Spike protein recognition of mammalian ACE2 predicts the host range and an optimized ACE2 for SARS-CoV-2 infection. Biochem. Biophys. Res. Commun. 526, 165–169. https://doi.org/10.1016/j.bbrc.2020.03.047

Maas, S., Patt, S., Schrey, M., Rich, A., 2001. Underediting of glutamate receptor GluR-B mRNA in malignant gliomas. Proc. Natl. Acad. Sci. 98, 14687–14692.

Machhi, J., Herskovitz, J., Senan, A.M., Dutta, D., Nath, B., Oleynikov, M.D., Blomberg, W.R., Meigs, D.D., Hasan, M., Patel, M., Kline, P., Chang, R.C.-C., Chang, L., Gendelman, H.E., Kevadiya, B.D., 2020. The Natural History, Pathobiology, and Clinical Manifestations of SARS-CoV-2 Infections. J. Neuroimmune Pharmacol. 15, 359–386. https://doi.org/10.1007/s11481-020-09944-5

Maggi, F., Novazzi, F., Genoni, A., Baj, A., Spezia, P.G., Focosi, D., Zago, C., Colombo, A., Cassani, G., Pasciuta, R., Tamborini, A., Rossi, A., Prestia, M., Capuano, R., Azzi, L., Donadini, A., Catanoso, G., Grossi, P.A., Maffioli, L., Bonelli, G., 2021. Imported SARS-CoV-2 Variant P.1 in Traveler Returning from Brazil to Italy. Emerg. Infect. Dis. 27, 1249–1251. https://doi.org/10.3201/eid2704.210183

Malig, M., Hartono, S.R., Giafaglione, J.M., Sanz, L.A., Chedin, F., 2020. Ultra-deep Coverage Single-molecule R-loop Footprinting Reveals Principles of R-loop Formation. J. Mol. Biol. 432, 2271–2288. https://doi.org/10.1016/j.jmb.2020.02.014

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., Trono, D., 2003. Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. Nature 424, 99–103. https://doi.org/10.1038/nature01709

Mannion, N.M., Greenwood, S.M., Young, R., Cox, S., Brindle, J., Read, D., Nellåker, C., Vesely, C., Ponting, C.P., McLaughlin, P.J., Jantsch, M.F., Dorin, J., Adams, I.R., Scadden, A.D.J., Ohman, M., Keegan, L.P., O'Connell, M.A., 2014. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. Cell Rep. 9, 1482–1494. https://doi.org/10.1016/j.celrep.2014.10.041

Marceca, G.P., Distefano, R., Tomasello, L., Lagana, A., Russo, F., Calore, F., Romano, G., Bagnoli, M., Gasparini, P., Ferro, A., Acunzo, M., Ma, Q., Croce, C.M., Nigita, G., 2021. MiREDiBase, a manually curated database of validated and putative editing events in microRNAs. Sci. Data 8, 199. https://doi.org/10.1038/s41597-021-00979-8

Marin, M., Rose, K.M., Kozak, S.L., Kabat, D., 2003. HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. Nat. Med. 9, 1398–1403. https://doi.org/10.1038/nm946

Martin, D.P., Weaver, S., Tegally, H., San, E.J., Shank, S.D., Wilkinson, E., Lucaci, A.G., Giandhari, J., Naidoo, S., Pillay, Y., Singh, L., Lessells, R.J., NGS-SA, COVID-19 Genomics UK (COG-UK), Gupta, R.K., Wertheim, J.O., Nekturenko, A., Murrell, B., Harkins, G.W., Lemey, P., MacLean, O.A., Robertson, D.L., de Oliveira, T., Kosakovsky Pond, S.L., 2021. The emergence and ongoing convergent evolution of the N501Y lineages coincides with a major global shift in the SARS-CoV-2 selective landscape. MedRxiv Prepr. Serv. Health Sci. 2021.02.23.21252268. https://doi.org/10.1101/2021.02.23.21252268

Martincorena, I., Campbell, P.J., 2015. Somatic mutation in cancer and normal cells. Science 349, 1483–1489. https://doi.org/10.1126/science.aab4082

Maura, F., Petljak, M., Lionetti, M., Cifola, I., Liang, W., Pinatel, E., Alexandrov, L.B., Fullam, A., Martincorena, I., Dawson, K.J., Angelopoulos, N., Samur, M.K., Szalat, R., Zamora, J., Tarpey, P., Davies, H., Corradini, P., Anderson, K.C., Minvielle, S., Neri, A., Avet-Loiseau, H., Keats, J., Campbell, P.J., Munshi, N.C., Bolli, N., 2018. Biological and prognostic impact of APOBEC-induced mutations in the spectrum of plasma cell dyscrasias and multiple myeloma cell lines. Leukemia 32, 1043–1047. https://doi.org/10.1038/leu.2017.345

McCallum, M., Bassi, J., Marco, A.D., Chen, A., Walls, A.C., Iulio, J.D., Tortorici, M.A., Navarro, M.-J., Silacci-Fregni, C., Saliba, C., Sprouse, K.R., Agostini, M., Pinto, D., Culap, K., Bianchi, S., Jaconi, S., Cameroni, E., Bowen, J.E., Tilles, S.W., Pizzuto, M.S., Guastalla, S.B., Bona, G., Pellanda, A.F., Garzoni, C., Voorhis, W.C.V., Rosen, L.E., Snell, G., Telenti, A., Virgin, H.W., Piccoli, L., Corti, D., Veesler, D., 2021. SARS-CoV-2 immune evasion by the B.1.427/B.1.429 variant of concern. Science 373, 648–654. https://doi.org/10.1126/science.abi7994

McCormick, K.D., Jacobs, J.L., Mellors, J.W., 2021. The emerging plasticity of SARS-CoV-2. Science 371, 1306–1308. https://doi.org/10.1126/science.abg4493

Merkle, T., Merz, S., Reautschnig, P., Blaha, A., Li, Q., Vogel, P., Wettengel, J., Li, J.B., Stafforst, T., 2019. Precise RNA editing by recruiting endogenous ADARs with antisense oligonucleotides. Nat. Biotechnol. 37, 133–138. https://doi.org/10.1038/s41587-019-0013-6

Miladi, M., Fuchs, J., Maier, W., Weigang, S., Pedrosa, N.D. i, Weiss, L., Lother, A., Nekrutenko, A., Ruzsics, Z., Panning, M., Kochs, G., Gilsbach, R., Grüning, B., 2020. The landscape of SARS-CoV-2 RNA modifications. bioRxiv 2020.07.18.204362. https://doi.org/10.1101/2020.07.18.204362

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37, 1530–1534. https://doi.org/10.1093/molbev/msaa015

Montiel-Gonzalez, M.F., Diaz Quiroz, J.F., Rosenthal, J.J.C., 2019. Current strategies for Site-Directed RNA Editing using ADARs. Methods, Mining the Epitranscriptome: Detection of RNA editing and RNA modifications 156, 16–24. https://doi.org/10.1016/j.ymeth.2018.11.016

Montiel-Gonzalez, M.F., Vallecillo-Viejo, I., Yudowski, G.A., Rosenthal, J.J.C., 2013. Correction of mutations within the cystic fibrosis transmembrane conductance regulator by site-directed RNA editing. Proc. Natl. Acad. Sci. 110, 18285–18290. https://doi.org/10.1073/pnas.1306243110

Montiel-González, M.F., Vallecillo-Viejo, I.C., Rosenthal, J.J.C., 2016. An efficient system for selectively altering genetic information within mRNAs. Nucleic Acids Res. 44, e157–e157. https://doi.org/10.1093/nar/gkw738

Morgan, G.J., Walker, B.A., Davies, F.E., 2012. The genetic architecture of multiple myeloma. Nat. Rev. Cancer 12, 335–348. https://doi.org/10.1038/nrc3257

Morse, D.P., Bass, B.L., 1999. Long RNA hairpins that contain inosine are present in Caenorhabditis elegans poly(A)+ RNA. Proc. Natl. Acad. Sci. 96, 6048–6053. https://doi.org/10.1073/pnas.96.11.6048

Mourier, T., Sadykov, M., Carr, M.J., Gonzalez, G., Hall, W.W., Pain, A., 2021. Host-directed editing of the SARS-CoV-2 genome. Biochem. Biophys. Res. Commun., COVID-19 538, 35–39. https://doi.org/10.1016/j.bbrc.2020.10.092

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., Honjo, T., 2000. Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. Cell 102, 553–563. https://doi.org/10.1016/S0092-8674(00)00078-7

Muramatsu, M., Sankaranand, V.S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N.O., Honjo, T., 1999. Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells *. J. Biol. Chem. 274, 18470–18476. https://doi.org/10.1074/jbc.274.26.18470

Muto, T., Muramatsu, M., Taniwaki, M., Kinoshita, K., Honjo, T., 2000. Isolation, Tissue Distribution, and Chromosomal Localization of the Human Activation-Induced Cytidine Deaminase (AID) Gene. Genomics 68, 85–88. https://doi.org/10.1006/geno.2000.6268

Namba, M., Ohtsuki, T., Mori, M., Togawa, A., Wada, H., Sugihara, T., Yawata, Y., Kimoto, T., 1989. Establishment of Five Human Myeloma Cell Lines. In Vitro Cell. Dev. Biol. 25, 723–729.

NCBI SARS-CoV-2 Resources [WWW Document], n.d. URL https://www.ncbi.nlm.nih.gov/sars-cov-2/ (accessed 5.10.21).

Nemec, P., Zemanova, Z., Greslikova, H., Michalova, K., Filkova, H., Tajtlova, J., Kralova, D., Kupska, R., Smetana, J., Krejci, M., Pour, L., Zahradova, L., Sandecka, V., Adam, Z., Buchler, T., Spicka, I., Gregora, E., Kuglik, P., Hajek, R., 2010. Gain of 1q21 Is an Unfavorable Genetic Prognostic Factor for Multiple Myeloma Patients Treated with High-Dose Chemotherapy. Biol. Blood Marrow Transplant. 16, 548–554. https://doi.org/10.1016/j.bbmt.2009.11.025

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., Menzies, A., Martin, S., Leung, K., Chen, L., Leroy, C., Ramakrishna, M., Rance, R., Lau, K.W., Mudie, L.J., Varela, I., McBride, D.J., Bignell, G.R., Cooke, S.L., Shlien, A., Gamble, J., Whitmore, I., Maddison, M., Tarpey, P.S., Davies, H.R., Papaemmanuil, E., Stephens, P.J., McLaren, S., Butler, A.P., Teague, J.W., Jönsson, G., Garber, J.E., Silver, D., Miron, P., Fatima, A., Boyault, S., Langerød, A., Tutt, A., Martens, J.W.M., Aparicio, S.A.J.R., Borg, Å., Salomon, A.V., Thomas, G., Børresen-Dale, A.-L., Richardson, A.L., Neuberger, M.S., Futreal, P.A., Campbell, P.J., Stratton, M.R., 2012. Mutational Processes Molding the Genomes of 21 Breast Cancers. Cell 149, 979–993. https://doi.org/10.1016/j.cell.2012.04.024

Nishikura, K., 2016. A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. Mol. Cell Biol. 17, 83–96. https://doi.org/10.1038/nrm.2015.4

Nishikura, K., 2010. Functions and regulation of RNA editing by ADAR deaminases. Annu. Rev. Biochem. 79, 321–349. https://doi.org/10.1146/annurev-biochem-060208-105251

Oakes, E., Anderson, A., Cohen-Gadol, A., Hundley, H.A., 2017. Adenosine Deaminase That Acts on RNA 3 (ADAR3) Binding to Glutamate Receptor Subunit B Pre-mRNA Inhibits RNA Editing in Glioblastoma*. J. Biol. Chem. 292, 4326–4335. https://doi.org/10.1074/jbc.M117.779868

Ogando, N.S., Zevenhoven-Dobbe, J.C., Meer, Y. van der, Bredenbeek, P.J., Posthuma, C.C., Snijder, E.J., 2020. The Enzymatic Activity of the nsp14 Exoribonuclease Is Critical for Replication of MERS-CoV and SARS-CoV-2. J. Virol. 94. https://doi.org/10.1128/JVI.01246-20

Ohlson, J., Pedersen, J.S., Haussler, D., Öhman, M., 2007. Editing modifies the GABAA receptor subunit α3. RNA 13, 698–703. https://doi.org/10.1261/rna.349107

O'Toole, Á., Hill, V., Pybus, O.G., Watts, A., Bogoch, I.I., Khan, K., Messina, J.P., The COVID-19 Genomics UK (COG-UK) consortium, Network for Genomic Surveillance in South Africa (NGS-SA), Brazil-UK CADDE Genomic Network, Tegally, H., Lessells, R.R., Giandhari, J., Pillay, S., Tumedi, K.A., Nyepetsi, G., Kebabonye, M., Matsheka, M., Mine, M., Tokajian, S., Hassan, H., Salloum, T., Merhi, G., Koweyes, J., Geoghegan, J.L., de Ligt, J., Ren, X., Storey, M., Freed, N.E., Pattabiraman, C., Prasad, P., Desai, A.S., Vasanthapuram, R., Schulz, T.F., Steinbrück, L., Stadler, T., Swiss Viollier Sequencing Consortium, Parisi, A., Bianco, A., García de Viedma, D., Buenestado-Serrano, S., Borges, V., Isidro, J., Duarte, S., Gomes, J.P., Zuckerman, N.S., Mandelboim, M., Mor, O., Seemann, T., Arnott, A., Draper, J., Gall, M., Rawlinson, W., Deveson, I., Schlebusch, S., McMahon, J., Leong, L., Lim, C.K., Chironna, M., Loconsole, D., Bal, A., Josset, L., Holmes, E., St. George, K., Lasek-Nesselquist, E., Sikkema, R.S., Oude Munnink, B., Koopmans, M., Brytting, M., Sudha rani, V., Pavani, S., Smura, T., Heim, A., Kurkela, S., Umair, M., Salman, M., Bartolini, B., Rueca, M., Drosten, C., Wolff, T., Silander, O., Eggink, D., Reusken, C., Vennema, H., Park, A., Carrington, C., Sahadeo, N., Carr, M., Gonzalez, G., SEARCH Alliance San Diego, National Virus Reference Laboratory, SeqCOVID-Spain, Danish Covid-19 Genome Consortium (DCGC), Communicable Diseases Genomic Network (CDGN), Dutch National SARS-CoV-2 surveillance program, Division of Emerging Infectious Diseases (KDCA), de Oliveira, T., Faria, N., Rambaut, A., Kraemer, M.U.G., 2021. Tracking the international spread of SARS-CoV-2 lineages B.1.1.7 and B.1.351/501Y-V2. Wellcome Open Res. 6, 121. https://doi.org/10.12688/wellcomeopenres.16661.1

Pang, B., McFaline, J.L., Burgis, N.E., Dong, M., Taghizadeh, K., Sullivan, M.R., Elmquist, C.E., Cunningham, R.P., Dedon, P.C., 2012. Defects in purine nucleotide metabolism lead to substantial incorporation of xanthine and hypoxanthine into DNA and RNA. Proc. Natl. Acad. Sci. 109, 2319–2324. https://doi.org/10.1073/pnas.1118455109

Pater, A.A., Bosmeny, M.S., Barkau, C.L., Ovington, K.N., Chilamkurthy, R., Parasrampuria, M., Eddington, S.B., Yinusa, A.O., White, A.A., Metz, P.E., Sylvain, R.J., Hebert, M.M., Benzinger, S.W., Sinha, K., Gagnon, K.T., 2021. Emergence and Evolution of a Prevalent New SARS-CoV-2 Variant in the United States. bioRxiv 2021.01.11.426287. https://doi.org/10.1101/2021.01.11.426287

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H.T., Danan-Gotthold, M., Knisbacher, B.A., Eisenberg, E., Levanon, E.Y., 2015. Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. Cell Rep. 13, 267–276. https://doi.org/10.1016/j.celrep.2015.08.080

Peischl, S., Dupanloup, I., Bosshard, L., Excoffier, L., 2016. Genetic surfing in human populations: from genes to genomes. Curr. Opin. Genet. Dev., Genetics of human origin 41, 53–61. https://doi.org/10.1016/j.gde.2016.08.003

Petljak, M., Alexandrov, L.B., Brammeld, J.S., Price, S., Wedge, D.C., Grossmann, S., Dawson, K.J., Ju, Y.S., Iorio, F., Tubio, J.M.C., Koh, C.C., Georgakopoulos-Soares, I., Rodríguez–Martín, B., Otlu, B., O'Meara, S., Butler, A.P., Menzies, A., Bhosle, S.G., Raine, K., Jones, D.R., Teague, J.W., Beal, K., Latimer, C., O'Neill, L., Zamora, J., Anderson, E., Patel, N., Maddison, M., Ng, B.L., Graham, J., Garnett, M.J., McDermott, U., Nik-Zainal, S., Campbell, P.J., Stratton, M.R., 2019. Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. Cell 176, 1282-1294.e20. https://doi.org/10.1016/j.cell.2019.02.012

Petljak, M., Maciejowski, J., 2020. Molecular origins of APOBEC-associated mutations in cancer. DNA Repair 94, 102905. https://doi.org/10.1016/j.dnarep.2020.102905

Picardi, E., D'Erchia, A.M., Lo Giudice, C., Pesole, G., 2017. REDIportal: a comprehensive database of A-to-I RNA editing events in humans. Nucleic Acids Res. 45, D750–D757. https://doi.org/10.1093/nar/gkw767

Picardi, E., D'Erchia, A.M., Montalvo, A., Pesole, G., 2015a. Using REDItools to Detect RNA Editing Events in NGS Datasets. Curr. Protoc. Bioinforma. 49, 12.12.1-12.12.15. https://doi.org/10.1002/0471250953.bi1212s49

Picardi, E., Manzari, C., Mastropasqua, F., Aiello, I., D'Erchia, A.M., Pesole, G., 2015b. Profiling RNA editing in human tissues: towards the inosinome Atlas. Sci. Rep. 5, 14941. https://doi.org/10.1038/srep14941

Picardi, E., Pesole, G., 2013. REDItools: high-throughput RNA editing detection made easy. Bioinforma. Oxf. Engl. 29, 1813–1814. https://doi.org/10.1093/bioinformatics/btt287

Planas, D., Bruel, T., Grzelak, L., Guivel-Benhassine, F., Staropoli, I., Porrot, F., Planchais, C., Buchrieser, J., Rajah, M.M., Bishop, E., Albert, M., Donati, F., Prot, M., Behillil, S., Enouf, V., Maquart, M., Smati-Lafarge, M., Varon, E., Schortgen, F., Yahyaoui, L., Gonzalez, M., De Sèze, J., Péré, H., Veyer, D., Sève, A., Simon-Lorière, E., Fafi-Kremer, S., Stefic, K., Mouquet, H., Hocqueloux, L., van der Werf, S., Prazuck, T., Schwartz, O., 2021. Sensitivity of infectious SARS-CoV-2 B.1.1.7 and B.1.351 variants to neutralizing antibodies. Nat. Med. 27, 917–924. https://doi.org/10.1038/s41591-021-01318-5

Plante, J.A., Liu, Y., Liu, J., Xia, H., Johnson, B.A., Lokugamage, K.G., Zhang, X., Muruato, A.E., Zou, J., Fontes-Garfias, C.R., Mirchandani, D., Scharton, D., Bilello, J.P., Ku, Z., An, Z., Kalveram, B., Freiberg, A.N., Menachery, V.D., Xie, X., Plante, K.S., Weaver, S.C., Shi, P.-Y., 2020.

Spike mutation D614G alters SARS-CoV-2 fitness. Nature 1–6. https://doi.org/10.1038/s41586-020-2895-3

Porath, H.T., Carmi, S., Levanon, E.Y., 2014. A genome-wide map of hyper-edited RNA reveals numerous new sites. Nat. Commun. 5, 4726. https://doi.org/10.1038/ncomms5726

Poulain, F., Lejeune, N., Willemart, K., Gillet, N.A., 2020. Footprint of the host restriction factors APOBEC3 on the genome of human viruses. PLOS Pathog. 16, e1008718. https://doi.org/10.1371/journal.ppat.1008718

Poulsen, H., Nilsson, J., Damgaard, C.K., Egebjerg, J., Kjems, J., 2001. CRM1 Mediates the Export of ADAR1 through a Nuclear Export Signal within the Z-DNA Binding Domain. Mol. Cell. Biol. 21, 7862–7871. https://doi.org/10.1128/MCB.21.22.7862-7871.2001

Qu, L., Yi, Z., Zhu, S., Wang, C., Cao, Z., Zhou, Z., Yuan, P., Yu, Y., Tian, F., Liu, Z., Bao, Y., Zhao, Y., Wei, W., 2019. Programmable RNA editing by recruiting endogenous ADAR using engineered RNAs. Nat. Biotechnol. 37, 1059–1069. https://doi.org/10.1038/s41587-019-0178-z

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Ramaswami, G., Li, J.B., 2014. RADAR: a rigorously annotated database of A-to-I RNA editing. Nucleic Acids Res. 42, D109-113. https://doi.org/10.1093/nar/gkt996

Rambaut, A., Holmes, E.C., O'Toole, Á., Hill, V., McCrone, J.T., Ruis, C., du Plessis, L., Pybus, O.G., 2020. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat. Microbiol. 5, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5

Ramiro, A.R., Stavropoulos, P., Jankovic, M., Nussenzweig, M.C., 2003. Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. Nat. Immunol. 4, 452–456. https://doi.org/10.1038/ni920

Ramos, A.H., Lichtenstein, L., Gupta, M., Lawrence, M.S., Pugh, T.J., Saksena, G., Meyerson, M., Getz, G., 2015. Oncotator: Cancer Variant Annotation Tool. Hum. Mutat. 36, E2423–E2429. https://doi.org/10.1002/humu.22771

Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., Zhang, F., 2013. Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8, 2281–2308. https://doi.org/10.1038/nprot.2013.143

Rausch, J.W., Capoferri, A.A., Katusiime, M.G., Patro, S.C., Kearney, M.F., 2020. Low genetic diversity may be an Achilles heel of SARS-CoV-2. Proc. Natl. Acad. Sci. 117, 24614–24616. https://doi.org/10.1073/pnas.2017726117

Reardon, S., 2020. Step aside CRISPR, RNA editing is taking off. Nature 578, 24–27. https://doi.org/10.1038/d41586-020-00272-5

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., Jenkinson, E.M., Bacino, C.A., Battini, R., Bertini, E., Brogan, P.A., Brueton, L.A., Carpanelli, M., De Laet, C., de Lonlay, P., del Toro, M., Desguerre, I., Fazzi, E., Garcia-Cazorla, À., Heiberg, A., Kawaguchi, M., Kumar, R., Lin, J.-P.S.-M., Lourenco, C.M., Male, A.M., Marques, W., Mignot, C., Olivieri, I., Orcesi, S., Prabhakar, P., Rasmussen, M., Robinson, R.A., Rozenberg, F., Schmidt, J.L., Steindl, K., Tan, T.Y., van der Merwe, W.G., Vanderver, A., Vassallo, G., Wakeling, E.L., Wassmer, E., Whittaker, E., Livingston, J.H., Lebon, P., Suzuki, T., McLaughlin, P.J., Keegan, L.P., O'Connell, M.A., Lovell, S.C., Crow, Y.J., 2012. Mutations in ADAR1 cause Aicardi-Goutières syndrome associated with a type I interferon signature. Nat. Genet. 44, 1243–1248. https://doi.org/10.1038/ng.2414

Rosenberg, B.R., Hamilton, C.E., Mwangi, M.M., Dewell, S., Papavasiliou, F.N., 2011. Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs. Nat. Struct. Mol. Biol. 18, 230–236. https://doi.org/10.1038/nsmb.1975

Roth, S.H., Levanon, E.Y., Eisenberg, E., 2019. Genome-wide quantification of ADAR adenosine-to-inosine RNA editing activity. Nat. Methods 16, 1131–1138. https://doi.org/10.1038/s41592-019-0610-9

Roundtree, I.A., Evans, M.E., Pan, T., He, C., 2017. Dynamic RNA Modifications in Gene Expression Regulation. Cell 169, 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045

Ryvkin, P., Leung, Y.Y., Silverman, I.M., Childress, M., Valladares, O., Dragomir, I., Gregory, B.D., Wang, L.-S., 2013. HAMR: high-throughput annotation of modified ribonucleotides. RNA 19, 1684–1692. https://doi.org/10.1261/rna.036806.112

Safra, M., Sas-Chen, A., Nir, R., Winkler, R., Nachshon, A., Bar-Yaacov, D., Erlacher, M., Rossmanith, W., Stern-Ginossar, N., Schwartz, S., 2017. The m 1 A landscape on cytosolic and mitochondrial mRNA at single-base resolution. Nature 551, 251–255. https://doi.org/10.1038/nature24456

Sale, J.E., Neuberger, M.S., 1998. TdT-Accessible Breaks Are Scattered over the Immunoglobulin V Domain in a Constitutively Hypermutating B Cell Line. Immunity 9, 859–869. https://doi.org/10.1016/S1074-7613(00)80651-2

Salter, J.D., Bennett, R.P., Smith, H.C., 2016. The APOBEC Protein Family: United by Structure, Divergent in Function. Trends Biochem. Sci. 41, 578–594. https://doi.org/10.1016/j.tibs.2016.05.001

Samuel, C.E., 2012. ADARs: Viruses and Innate Immunity, in: Samuel, C.E. (Ed.), Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing, Current Topics in Microbiology and Immunology. Springer, Berlin, Heidelberg, pp. 163–195. https://doi.org/10.1007/82_2011_148

Sanyaolu, A., Okorie, C., Marinkovic, A., Haider, N., Abbasi, A.F., Jaferi, U., Prakash, S., Balendra, V., 2021. The emerging SARS-CoV-2 variants of concern. Ther. Adv. Infect. Dis. 8, 20499361211024372. https://doi.org/10.1177/20499361211024372

Sanz, L.A., Hartono, S.R., Lim, Y.W., Steyaert, S., Rajpurkar, A., Ginno, P.A., Xu, X., Chédin, F., 2016. Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. Mol. Cell 63, 167–178. https://doi.org/10.1016/j.molcel.2016.05.032

Saraconi, G., Severi, F., Sala, C., Mattiuz, G., Conticello, S.G., 2014. The RNA editing enzyme APOBEC1 induces somatic mutations and a compatible mutational signature is present in esophageal adenocarcinomas. Genome Biol. 15, 417. https://doi.org/10.1186/s13059-014-0417-z

Schaffer, A.A., Kopel, E., Hendel, A., Picardi, E., Levanon, E.Y., Eisenberg, E., 2020. The cell line A-to-I RNA editing catalogue. Nucleic Acids Res. 48, 5849–5858. https://doi.org/10.1093/nar/gkaa305

Schwartz, S., 2018. m1A within cytoplasmic mRNAs at single nucleotide resolution: a reconciled transcriptome-wide map. RNA 24, 1427–1436. https://doi.org/10.1261/rna.067348.118

Sendinc, E., Valle-Garcia, D., Dhall, A., Chen, H., Henriques, T., Navarrete-Perea, J., Sheng, W., Gygi, S.P., Adelman, K., Shi, Y., 2019. PCIF1 Catalyzes m6Am mRNA Methylation to Regulate Gene Expression. Mol. Cell 75, 620-630.e9. https://doi.org/10.1016/j.molcel.2019.05.030

Sharma, S., Patnaik, S.K., Taggart, R.T., Kannisto, E.D., Enriquez, S.M., Gollnick, P., Baysal, B.E., 2015. APOBEC3A cytidine deaminase induces RNA editing in monocytes and macrophages. Nat. Commun. 6, 6881. https://doi.org/10.1038/ncomms7881

Sheehy, A.M., Gaddis, N.C., Choi, J.D., Malim, M.H., 2002. Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. Nature 418, 646–650. https://doi.org/10.1038/nature00939

Shiromoto, Y., Sakurai, M., Minakuchi, M., Ariyoshi, K., Nishikura, K., 2021. ADAR1 RNA editing enzyme regulates R-loop formation and genome stability at telomeres in cancer cells. Nat. Commun. 12, 1654. https://doi.org/10.1038/s41467-021-21921-x

Siegel, R.L., Miller, K.D., Fuchs, H.E., Jemal, A., 2021. Cancer Statistics, 2021. CA. Cancer J. Clin. 71, 7–33. https://doi.org/10.3322/caac.21654

Simmonds, P., 2020. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5, e00408-20, /msphere/5/3/mSphere408-20.atom. https://doi.org/10.1128/mSphere.00408-20

Slatkin, M., Excoffier, L., 2012. Serial Founder Effects During Range Expansion: A Spatial Analog of Genetic Drift. Genetics 191, 171–181. https://doi.org/10.1534/genetics.112.139022

Sloan, K.E., Warda, A.S., Sharma, S., Entian, K.-D., Lafontaine, D.L.J., Bohnsack, M.T., 2017. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. RNA Biol. 14, 1138–1152. https://doi.org/10.1080/15476286.2016.1259781

Slotkin, W., Nishikura, K., 2013. Adenosine-to-inosine RNA editing and human disease. Genome Med. 5, 105. https://doi.org/10.1186/gm508

Snyder, E.M., McCarty, C., Mehalow, A., Svenson, K.L., Murray, S.A., Korstanje, R., Braun, R.E., 2017. APOBEC1 complementation factor (A1CF) is dispensable for C-to-U RNA editing in vivo. RNA 23, 457–465. https://doi.org/10.1261/rna.058818.116

Sohail, A., Klapacz, J., Samaranayake, M., Ullah, A., Bhagwat, A.S., 2003. Human activation-induced cytidine deaminase causes transcription-dependent, strand-biased C to U deaminations. Nucleic Acids Res. 31, 2990–2994. https://doi.org/10.1093/nar/gkg464

Song, Y., Yang, W., Fu, Q., Wu, L., Zhao, X., Zhang, Y., Zhang, R., 2020. irCLASH reveals RNA substrates recognized by human ADARs. Nat. Struct. Mol. Biol. 27, 351–362. https://doi.org/10.1038/s41594-020-0398-4

Stafforst, T., Schneider, M.F., 2012. An RNA–Deaminase Conjugate Selectively Repairs Point Mutations. Angew. Chem. Int. Ed. 51, 11166–11169. https://doi.org/10.1002/anie.201206489

Starrett, G.J., Luengas, E.M., McCann, J.L., Ebrahimi, D., Temiz, N.A., Love, R.P., Feng, Y., Adolph, M.B., Chelico, L., Law, E.K., Carpenter, M.A., Harris, R.S., 2016. The DNA cytosine deaminase APOBEC3H haplotype I likely contributes to breast and lung cancer mutagenesis. Nat. Commun. 7, 12918. https://doi.org/10.1038/ncomms12918

Stavrou, S., Ross, S.R., 2015. APOBEC3 Proteins in Viral Immunity. J. Immunol. 195, 4565–4570. https://doi.org/10.4049/jimmunol.1501504

Stefl, R., Oberstrass, F.C., Hood, J.L., Jourdan, M., Zimmermann, M., Skrisovska, L., Maris, C., Peng, L., Hofr, C., Emeson, R.B., Allain, F.H.-T., 2010. The Solution Structure of the ADAR2 dsRBM-RNA Complex Reveals a Sequence-Specific Readout of the Minor Groove. Cell 143, 225–237. https://doi.org/10.1016/j.cell.2010.09.026

Stolz, R., Sulthana, S., Hartono, S.R., Malig, M., Benham, C.J., Chedin, F., 2019. Interplay between DNA sequence and negative superhelicity drives R-loop structures. Proc. Natl. Acad. Sci. 116, 6260–6269. https://doi.org/10.1073/pnas.1819476116

Tasakis, R.N., Laganà, A., Stamkopoulou, D., Melnekoff, D.T., Nedumaran, P., Leshchenko, V., Pecori, R., Parekh, S., Papavasiliou, F.N., 2020. ADAR1 can drive Multiple Myeloma progression by acting both as an RNA editor of specific transcripts and as a DNA mutator of their cognate genes. bioRxiv 2020.02.11.943845. https://doi.org/10.1101/2020.02.11.943845

Tasakis, R.N., Papavasiliou, F.N., Shaknovich, R., 2019. RNA Editors and DNA Mutators: Cancer Heterogeneity Through Sequence Diversification. OBM Genet. 3, 17.

Tasakis, R.N., Samaras, G., Jamison, A., Lee, M., Paulus, A., Whitehouse, G., Verkoczy, L., Papavasiliou, F.N., Diaz, M., 2021. SARS-CoV-2 variant evolution in the United States: High accumulation of viral mutations over time likely through serial Founder Events and mutational bursts. PLOS ONE 16, e0255169. https://doi.org/10.1371/journal.pone.0255169

Teng, B., Burant, C.F., Davidson, N.O., 1993. Molecular cloning of an apolipoprotein B messenger RNA editing protein. Science 260, 1816–1819. https://doi.org/10.1126/science.8511591

Teoh, P.J., An, O., Chung, T.-H., Chooi, J.Y., Toh, S.H.M., Fan, S., Wang, W., Koh, B.T.H., Fullwood, M.J., Ooi, M.G., de Mel, S., Soekojo, C.Y., Chen, L., Ng, S.B., Yang, H., Chng, W.J., 2018. Aberrant hyperediting of the myeloma transcriptome by ADAR1 confers oncogenicity and is a marker of poor prognosis. Blood 132, 1304–1317. https://doi.org/10.1182/blood-2018-02-832576

Toth, A.M., Li, Z., Cattaneo, R., Samuel, C.E., 2009. RNA-specific Adenosine Deaminase ADAR1 Suppresses Measles Virus-induced Apoptosis and Activation of Protein Kinase PKR*. J. Biol. Chem. 284, 29350–29356. https://doi.org/10.1074/jbc.M109.045146

Toyoshima, Y., Nemoto, K., Matsumoto, S., Nakamura, Y., Kiyotani, K., 2020. SARS-CoV-2 genomic variations associated with mortality rate of COVID-19. J. Hum. Genet. 65, 1075–1082. https://doi.org/10.1038/s10038-020-0808-9

Tzou, P.L., Tao, K., Nouhin, J., Rhee, S.-Y., Hu, B.D., Pai, S., Parkin, N., Shafer, R.W., 2020. Coronavirus Antiviral Research Database (CoV-RDB): An Online Database Designed to Facilitate Comparisons between Candidate Anti-Coronavirus Compounds. Viruses 12, 1006. https://doi.org/10.3390/v12091006

Upton, D.C., Unniraman, S., 2011. Assessing Somatic Hypermutation in Ramos B Cells after Overexpression or Knockdown of Specific Genes. J. Vis. Exp. JoVE. https://doi.org/10.3791/3573

Vallecillo-Viejo, I.C., Liscovitch-Brauer, N., Montiel-Gonzalez, M.F., Eisenberg, E., Rosenthal, J.J.C., 2017. Abundant off-target edits from site-directed RNA editing can be reduced by nuclear localization of the editing enzyme. RNA Biol. 15, 104–114. https://doi.org/10.1080/15476286.2017.1387711

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., Ortiz, A.T., Balloux, F., 2020. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. 83, 104351. https://doi.org/10.1016/j.meegid.2020.104351

Vogel, P., Moschref, M., Li, Q., Merkle, T., Selvasaravanan, K.D., Li, J.B., Stafforst, T., 2018. Efficient and precise editing of endogenous transcripts with SNAP-tagged ADARs. Nat. Methods 15, 535–538. https://doi.org/10.1038/s41592-018-0017-z

Vogel, P., Schneider, M.F., Wettengel, J., Stafforst, T., 2014. Improving site-directed RNA editing in vitro and in cell culture by chemical modification of the guideRNA. Angew. Chem. Int. Ed Engl. 53, 6267–6271. https://doi.org/10.1002/anie.201402634

Vogel, P., Stafforst, T., 2019. Critical review on engineering deaminases for site-directed RNA editing. Curr. Opin. Biotechnol., Analytical Biotechnology 55, 74–80. https://doi.org/10.1016/j.copbio.2018.08.006

Walker, B.A., Wardell, C.P., Melchor, L., Brioli, A., Johnson, D.C., Kaiser, M.F., Mirabella, F., Lopez-Corral, L., Humphray, S., Murray, L., Ross, M., Bentley, D., Gutiérrez, N.C., Garcia-Sanz, R., San Miguel, J., Davies, F.E., Gonzalez, D., Morgan, G.J., 2014. Intraclonal heterogeneity is a critical early event in the development of myeloma and precedes the development of clinical symptoms. Leukemia 28, 384–390. https://doi.org/10.1038/leu.2013.199

Wang, C., Horby, P.W., Hayden, F.G., Gao, G.F., 2020. A novel coronavirus outbreak of global health concern. The Lancet 395, 470–473. https://doi.org/10.1016/S0140-6736(20)30185-9

Wang, I.X., So, E., Devlin, J.L., Zhao, Y., Wu, M., Cheung, V.G., 2013. ADAR Regulates RNA Editing, Transcript Stability, and Gene Expression. Cell Rep. 5, 849–860. https://doi.org/10.1016/j.celrep.2013.10.002

Wang, L.-G., Lam, T.T.-Y., Xu, S., Dai, Z., Zhou, L., Feng, T., Guo, P., Dunn, C.W., Jones, B.R., Bradley, T., Zhu, H., Guan, Y., Jiang, Y., Yu, G., 2020. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol. Biol. Evol. 37, 599–603. https://doi.org/10.1093/molbev/msz240

Wang, P., Casner, R.G., Nair, M.S., Wang, M., Yu, J., Cerutti, G., Liu, L., Kwong, P.D., Huang, Y., Shapiro, L., Ho, D.D., 2021. Increased resistance of SARS-CoV-2 variant P.1 to antibody neutralization. Cell Host Microbe 29, 747-751.e4. https://doi.org/10.1016/j.chom.2021.04.007

Wang, Q., Miyakoda, M., Yang, W., Khillan, J., Stachura, D.L., Weiss, M.J., Nishikura, K., 2004. Stress-induced apoptosis associated with null mutation of ADAR1 RNA editing deaminase gene. J. Biol. Chem. 279, 4952–4961. https://doi.org/10.1074/jbc.M310162200

Wang, R., Chen, J., Gao, K., Hozumi, Y., Yin, C., Wei, G.-W., 2021. Analysis of SARS-CoV-2 mutations in the United States suggests presence of four substrains and novel variants. Commun. Biol. 4, 1–14. https://doi.org/10.1038/s42003-021-01754-6

Wang, R., Hozumi, Y., Yin, C., Wei, G.-W., 2020a. Decoding SARS-CoV-2 Transmission and Evolution and Ramifications for COVID-19 Diagnosis, Vaccine, and Medicine. J. Chem. Inf. Model. 60, 5853–5865. https://doi.org/10.1021/acs.jcim.0c00501

Wang, R., Hozumi, Y., Zheng, Y.-H., Yin, C., Wei, G.-W., 2020b. Host Immune Response Driving SARS-CoV-2 Evolution. Viruses 12. https://doi.org/10.3390/v12101095

West, A.P., Wertheim, J.O., Wang, J.C., Vasylyeva, T.I., Havens, J.L., Chowdhury, M.A., Gonzalez, E., Fang, C.E., Lonardo, S.S.D., Hughes, S., Rakeman, J.L., Lee, H.H., Barnes, C.O., Gnanapragasam, P.N.P., Yang, Z., Gaebler, C., Caskey, M., Nussenzweig, M.C., Keeffe, J.R., Bjorkman, P.J., 2021. Detection and characterization of the SARS-CoV-2 lineage B.1.526 in New York. https://doi.org/10.1101/2021.02.14.431043

Wettengel, J., Reautschnig, P., Geisler, S., Kahle, P.J., Stafforst, T., 2017. Harnessing human ADAR2 for RNA repair – Recoding a PINK1 mutation rescues mitophagy. Nucleic Acids Res. 45, 2797–2808. https://doi.org/10.1093/nar/gkw911

WHO Coronavirus (COVID-19) Dashboard [WWW Document], n.d. URL https://covid19.who.int (accessed 7.5.21).

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T.L., Miller, E., Bache, S.M., Müller, K., Ooms, J., Robinson, D., Seidel, D.P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686

Woolf, T.M., Chase, J.M., Stinchcomb, D.T., 1995. Toward the therapeutic editing of mutated RNA sequences. Proc. Natl. Acad. Sci. U. S. A. 92, 8298–8302. https://doi.org/10.1073/pnas.92.18.8298

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., Hu, Y., Tao, Z.-W., Tian, J.-H., Pei, Y.-Y., Yuan, M.-L., Zhang, Y.-L., Dai, F.-H., Liu, Y., Wang, Q.-M., Zheng, J.-J., Xu, L., Holmes, E.C., Zhang, Y.-Z., 2020. A new coronavirus associated with human respiratory disease in China. Nature 579, 265–269. https://doi.org/10.1038/s41586-020-2008-3

Wu, T.D., Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinforma. Oxf. Engl. 21, 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Yan, Q., Shields, E.J., Bonasio, R., Sarma, K., 2019. Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach. Cell Rep. 29, 1369-1380.e5. https://doi.org/10.1016/j.celrep.2019.09.052

Yi, C., Sun, X., Ye, J., Ding, L., Liu, M., Yang, Z., Lu, X., Zhang, Y., Ma, L., Gu, W., Qu, A., Xu, J., Shi, Z., Ling, Z., Sun, B., 2020. Key residues of the receptor binding motif in the spike protein of SARS-CoV-2 that interact with ACE2 and neutralizing antibodies. Cell. Mol. Immunol. 17, 621–630. https://doi.org/10.1038/s41423-020-0458-z

Yin, X., Riva, L., Pu, Y., Martin-Sancho, L., Kanamune, J., Yamamoto, Y., Sakai, K., Gotoh, S., Miorin, L., De Jesus, P.D., Yang, C.-C., Herbert, K.M., Yoh, S., Hultquist, J.F., García-Sastre, A., Chanda, S.K., 2021. MDA5 Governs the Innate Immune Response to SARS-CoV-2 in Lung Epithelial Cells. Cell Rep. 34, 108628. https://doi.org/10.1016/j.celrep.2020.108628

Yoon, C.-H., Lee, E.-S., Lim, D.-S., Bae, Y.-S., 2009. PKR, a p53 target gene, plays a crucial role in the tumor-suppressor function of p53. Proc. Natl. Acad. Sci. 106, 7852–7857. https://doi.org/10.1073/pnas.0812148106

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y., 2017. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Zheng, C., Ikewaki, K., Walsh, B.W., Sacks, F.M., 2006. Metabolism of apoB lipoproteins of intestinal and hepatic origin during constant feeding of small amounts of fat. J. Lipid Res. 47, 1771–1779. https://doi.org/10.1194/jlr.M500528-JLR200

Zheng, Y., Lorenzo, C., Beal, P.A., 2017. DNA editing in DNA/RNA hybrids by adenosine deaminases that act on RNA. Nucleic Acids Res. 45, 3369–3377. https://doi.org/10.1093/nar/gkx050

Zhou, H.-Y., Ji, C.-Y., Fan, H., Han, N., Li, X.-F., Wu, A., Qin, C.-F., 2021. Convergent evolution of SARS-CoV-2 in human and animals. Protein Cell. https://doi.org/10.1007/s13238-021-00847-6

Zinshteyn, B., Nishikura, K., 2009. Adenosine-to-inosine RNA editing. WIREs Syst. Biol. Med. 1, 202–209. https://doi.org/10.1002/wsbm.10