Probing RNA binding specificities of AID/APOBEC proteins by iCLIP

Brenda Valeiras

MRC Laboratory of Molecular Biology
University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Darwin College

October 2018
To my grandparents, for their love and teachings on the value of education and hard work.

Especially to my abuela Lidia, who one day envisioned that this was possible.
Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except where specifically indicated in the preface. This dissertation contains fewer than 60,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Brenda Valeiras
October 2018
Summary

Probing RNA binding specificities of AID/APOBEC proteins by iCLIP

Brenda Valeiras

The AID/APOBEC protein family comprises a group of cytosine deaminases found in vertebrates that are capable of modifying cytosine to uracil in the context of RNA or single-stranded DNA. They exert diverse valuable physiological functions including antibody diversification and restriction of viral infection. However, off-target mutations have also been shown to contribute to cancer development, making it crucial to better understand the interactions and mechanisms that regulate AID/APOBEC activity and editing site fidelity.

In this regard, a new focus on RNA as a putative regulator of AID/APOBECs has recently emerged. Regardless of whether it is used or not as a substrate for deamination, most members of the family have been shown to retain the ability to bind RNA, emphasizing a potential regulatory role for this interaction. However, little is known about AID/APOBECs RNA binding specificity. A promiscuous binding has been suggested in some cases while in vitro evidence for other members of the family indicate a certain level of specificity. Therefore, to thoroughly unravel the AID/APOBECs RNA binding specificity, in my doctoral research I applied cross-linking and immunoprecipitation (iCLIP), an unbiased technique that allows identification of protein-bound RNAs with nucleotide resolution in living cells.

As a first approach, I adapted the technique for its use in yeast and probed the RNA binding of AID and APOBEC3G, revealing different degrees of preference for small structured RNAs and recognition of particular sites within them. I then expanded the analysis to mammalian cells (HEK293T) and evaluated an extended set of APOBECs finding that, even in the presence of a broader and more complex pool of RNAs, small RNAs were still significantly bound by some members of the family. Furthermore, the comparative analysis of AID, APOBEC1, APOBEC3G, APOBEC3A and APOBEC3B iCLIP data obtained in my research, revealed shared and individual preferences for certain RNAs, suggesting a degree of binding specificity among APOBECs.

In summary, my thesis outlines for the first time a comprehensive analysis of the RNA binding specificity of different AID/APOBECs in vivo, including the description of novel interactions with nucleotide resolution. The results obtained are of great value and open the field for further investigation of the specific meaning and validation of each preferential binding, providing new insights into understanding the role of AID/APOBEC interaction with RNA.
Preface

In this thesis, I describe my doctoral research performed at the MRC Laboratory of Molecular Biology, Cambridge, UK, between October 2014 and October 2018.

The computational work was performed in collaboration with Dr. Igor Ruiz de los Mozos, from Dr. Jerne Ule group, Francis Crick Institute, London, UK, who kindly uploaded the sequencing data into the web server iCount, established by Dr. Tomaž Curk. Hit per protein matrix and heatmaps presented in Chapter 5 as well as general statistical advice were performed in collaboration with Meltem Gürel, from Cancer Research UK, Cambridge, UK.
Acknowledgements

Firstly, I would like to thank my supervisor, Cristina Rada, for giving me the possibility to work on this challenging project and always be willing to help and support my decisions. I am also grateful for the input from past members of the group Ben Taylor and Yee Ling Wu.

I would like to express my especial gratitude to Dr. Ina Huppertz and Dr. Igor Ruiz De Los Mozos for so kindly guiding me into the iCLIP world, sharing their inexhaustible methodological knowledge. Likewise, I want to thank Dr. Alastair Crisp for his help with data analysis and Dr. Andrew Mckenzie for his valuable input, support and advice. For their companionship and fun times in and outside work, I wish to thank my amazing LMB friends Cara, Marina, Davide, Hayat, Benedicte, Michal, Thierry and Migle.

I am also very grateful for the opportunity given to me by the Darwin Trust of Edinburgh with the César Milstein Studentship that allowed me not only to do my PhD but also to meet Celia Milstein, who I admire deeply and want to thank for her always kind support, care, and advice.

My good friend Meltem Gürel, I cannot thank her enough not only for her enormous contributions to my thesis, long hours and help with data analysis but mostly for her unconditional invaluable friendship that made me feel accompanied throughout my PhD and made it more fun as well. Her, together with Elaine, Robbin, Matías and Coni made my life in Cambridge a much more enjoyable experience.

For being by my side, for his incredible patience and loving support through my ups and downs, I especially want to thank Simon, who cheered me up when I was running out of steam.

Finally, I would have not been able to do any of this work without the inestimable support and encouragement from my amazing family and friends back home. Their visits and constant presence even from a distance gave me the confidence and drive needed to overcome the difficulties and shared with me the happiness of every achievement along the way. Especial thanks to my inspiring parents, Roberto and Graciela, and my sister Bárbara for always being there backing up every step of my career even when it meant going to the other side of the country or the world. Likewise, to my aunts, cousins and little ones: Blume, Ali, Lo, Lu, Lei, Cin, Isa, Valen and Ame. My chosen family: Jesi, mi Concu, Silvi, Gas, Pola, Marian, Mode, Pantu, Lu and Pichi. To all of them, my greatest gratitude and love. Thank you!
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3’UTR</td>
<td>3’ untranslated region</td>
</tr>
<tr>
<td>5’UTR</td>
<td>5’ untranslated region</td>
</tr>
<tr>
<td>AID</td>
<td>Activation induced deaminase</td>
</tr>
<tr>
<td>A1</td>
<td>APOBEC 1</td>
</tr>
<tr>
<td>A2</td>
<td>APOBEC 2</td>
</tr>
<tr>
<td>A3</td>
<td>APOBEC3</td>
</tr>
<tr>
<td>A3A</td>
<td>APOBEC3A</td>
</tr>
<tr>
<td>A3B</td>
<td>APOBEC3B</td>
</tr>
<tr>
<td>A3C</td>
<td>APOBEC3C</td>
</tr>
<tr>
<td>A3F</td>
<td>APOBEC3F</td>
</tr>
<tr>
<td>A3G</td>
<td>APOBEC3G</td>
</tr>
<tr>
<td>bp</td>
<td>Base pair</td>
</tr>
<tr>
<td>CTD</td>
<td>C-terminal domain</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>CDA</td>
<td>Cytidine deaminase</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>FDR</td>
<td>False discovery rate</td>
</tr>
<tr>
<td>HEK</td>
<td>Human embryonic kidney</td>
</tr>
<tr>
<td>iCLIP</td>
<td>Individual-nucleotide resolution CLIP</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>miRNA</td>
<td>MicroRNA</td>
</tr>
<tr>
<td>miscRNA</td>
<td>Miscellaneous RNA</td>
</tr>
<tr>
<td>NTD</td>
<td>N-terminal domain</td>
</tr>
<tr>
<td>nt</td>
<td>Nucleotides</td>
</tr>
<tr>
<td>ORF</td>
<td>Open reading frame</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase chain reaction</td>
</tr>
<tr>
<td>RNase</td>
<td>Ribonuclease</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>ssDNA</td>
<td>Single stranded DNA</td>
</tr>
<tr>
<td>snRNA</td>
<td>Small nuclear RNA</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>------------------------------------------------</td>
</tr>
<tr>
<td>SDS-PAGE</td>
<td>Sodium dodecyl sulfate polyacrylamide gel</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>UV</td>
<td>Ultraviolet</td>
</tr>
<tr>
<td>CLIP</td>
<td>UV crosslinking and immunoprecipitation</td>
</tr>
</tbody>
</table>
# Table of contents

Declaration \hspace{1cm} v

Summary \hspace{1cm} vii

Preface \hspace{1cm} ix

Acknowledgements \hspace{1cm} xi

List of abbreviations \hspace{1cm} xiii

## 1 Introduction

1.1 The AID/APOBEC family \hspace{1cm} 1

1.2 Multifaceted functions of AID/APOBEC proteins \hspace{1cm} 3

1.2.1 AID \hspace{1cm} 3

1.2.2 APOBEC 1 \hspace{1cm} 4

1.2.3 APOBEC3s \hspace{1cm} 5

1.2.4 APOBEC 2 and APOBEC 4 \hspace{1cm} 7

1.3 AID/APOBECs in Cancer \hspace{1cm} 8

1.4 Multilevel regulation of AID/APOBEC deaminases \hspace{1cm} 10

1.4.1 AID/APOBECs controlled expression and subcellular localization \hspace{1cm} 10

1.4.2 Cofactors and substrate nucleic acid recognition \hspace{1cm} 11

1.4.3 RNA binding role in AID/APOBEC regulation \hspace{1cm} 12

1.5 High-throughput sequencing methods to study protein-RNA interactions \hspace{1cm} 13

1.5.1 Clip-based methods \hspace{1cm} 13

1.6 Aim of the thesis \hspace{1cm} 16

## 2 Materials and Methods

2.1 Materials \hspace{1cm} 17

2.1.1 Antibodies \hspace{1cm} 17

2.1.2 Plasmids \hspace{1cm} 17

2.1.3 Buffers \hspace{1cm} 17

2.1.4 RT primers for the iCLIP protocol \hspace{1cm} 19

2.2 Basic Molecular biology methods \hspace{1cm} 20

2.2.1 Polymerase Chain reaction (PCR) \hspace{1cm} 20

2.2.2 PCR purification and visualization \hspace{1cm} 20
2.3 Yeast and cell maintenance

2.3.1 Cell Culture

2.3.2 Transient transfection

2.3.3 Yeast Culture

2.3.4 Yeast transformation

2.4 iCLIP

2.4.1 Input material for yeast adapted iCLIP

2.4.2 Input material for HEK293 iCLIP

2.4.3 General iCLIP protocol

2.5 Protein-based techniques

2.5.1 Preparation of protein lysate

2.5.2 Protein Concentration measurement

2.5.3 Western Blot analysis

2.6 Fluorescence Microscopy

2.7 Computational methods

2.7.1 Sequencing and Mapping

2.7.2 Addition of tRNAs

2.7.3 Identification of significant iCLIP crosslink sites

2.7.4 Heatmaps

3 Analysis of RNA binding preferences of AID and the catalytic domain of APOBEC3G in yeast by iCLIP

3.1 iCLIP optimization in yeast and library construction for AID and cA3G

3.1.1 Adjusting UV-C crosslinking

3.1.2 Partial RNAse digestion

3.1.3 Library preparation for AID and cA3G

3.2 AID and cA3G preferentially bind yeast tRNAs in vivo

3.3 AID and cA3G bind different sites within tRNAs

3.4 Discussion

4 Screening the RNA binding of AID/APOBEC deaminases in HEK 293 cells by iCLIP

4.1 Localization of overexpressed tagged deaminases in HEK293

4.2 iCLIP optimization for AID/APOBEC deaminases in HEK293

4.3 AID/APOBEC deaminases library preparation
5 AID and APOBEC proteins distinct RNA binding specificities revealed by iCLIP

5.1 AID/APOBEC deaminases have distinct preferences across different RNA types

5.2 Global analysis of AID/APOBEC deaminases differentially bound RNAs

5.3 AID specifically binds tRNAs in HEK293 cells

5.4 AID preferentially binds miRNAs in HEK293 living cells

6 Discussion

7 References

8 Appendix
Chapter 1

Introduction

1. The AID/APOBEC family

Activation-induced deaminase (AID) and apolipoprotein B mRNA-editing catalytic polypeptide-like (APOBEC) proteins constitute a family of cytosine deaminases (CDA) that share the ability to catalyse the deamination of cytidine to uridine in the context of RNA or single-stranded DNA (ssDNA). In humans there are 11 different AID/APOBEC cytidine deaminases: AID, APOBEC1, APOBEC2, APOBEC3 (A-H) and APOBEC4 (reviewed in [1]). Each member is selective in the sequence context of the deaminated cytosine, with the two preceding nucleotides identifying the signature of individual deaminases [2] (Figure 1.1).

AID/APOBEC proteins constitute a subset of a larger family of zinc-dependent deaminases involved in purine and pyrimidine metabolism, including adenosine deaminase active on RNA (ADARs) [3] and adenosine deaminase active on tRNA [4], responsible for A to I editing. They all share a conserved zinc-coordinating deaminase domain (ZDD) with a H[AV]E-x[24-36]-PCxxC motif (X representing any aminoacid) where the histidine and cysteine residues coordinate the zinc ion, whereas the glutamic acid acts as a proton donor during the catalytic deamination reaction. Within the AID/APOBEC family, AID, APOBEC1 (A1), APOBEC2(A2), APOBEC3A (A3A), APOBEC3C (A3C), and APOBEC3H (A3H) have a single ZDD domain, whereas APOBEC3B (A3B), APOBEC3DE (A3DE), APOBEC3F (A3F), and APOBEC3G (A3G) have two, N-and C-terminal domains (NTD and CTD) [5] (Figure 1.1).
Chapter 1. Introduction

Figure 1.1. The AID/APOBEC family.
Bar diagrams of human AID/APOBEC proteins containing one or two conserved zinc coordinating deaminase domains (ZDD) (in yellow). Numbers correspond indicate the length of the aminoacid sequence. The (*) next to the ZDD for APOBEC4 indicates that it is divergent from the consensus ZDD. Cellular localization: C = cytoplasmic, N = nuclear, N/C = even distribution and n/C = predominantly cytoplasmic with traffic to nucleus. Most frequent edited C sequence context, W= A or T, R= A or G, Y= T or C. (?) in all cases denotes unknown or mixed results depending on the system as in the case of APOBEC2 deaminase activity (see text in section 1.2.4). Adapted from [1].

The core structure of AID/APOBEC cytidine deaminases domains consists of five β-strands and six α-helices. The β-strands are organized into a hydrophobic β-sheet core, and the α-helices are positioned around this core. Subtle differences in the length of secondary structural elements and loop regions, deletions/insertions of residues, and specific residues near the active site are likely the primary discriminators for sequence preference, substrate binding affinity, catalytic rate and overall function among the family members (reviewed in [1]) (Figure 1.2).
Figure 1.2. Structure of AID/APOBEC cytidine deaminases. APOBEC3A NMR solution structure (PDB ID 2M65) as a representative AID/APOBEC structure depicting the canonical cytidine deaminase fold common among all family members consisting of five-stranded b-sheet flanked by six a-helices. b-strands are shown in grey, while a-helices are shown in aquamarine, and intervening loops are coloured green. The catalytic zinc ion is depicted as a purple sphere. Sidechains of the zinc-coordinating residues are depicted in orange while sidechains of the catalytic glutamic acid are depicted in red. (taken from [6])

1.2 Multifaceted functions of AID/APOBEC proteins

Even though at a cursory level AID/APOBEC deaminases might appear very similar due to a shared basic structural homology and catalytic activity, different members of the family have very diverse functions and substrate selectivity.

1.2.1 AID

AID has a crucial function in the adaptive humoral immune system, being responsible for antibody diversity. It is primarily expressed in mature B- cells and within secondary lymphoid organs such as the spleen and lymph nodes [7]. Upon B lymphocyte activation, the cytidine deaminase activity of AID introduces numerous dC to dU transitions within the transcribed regions of the immunoglobulin locus. These mutations enable antibody diversification through the generation of new antibody variable regions (somatic
hypermutation) and heavy chain switching (class switch recombination) [8,9]. In fact, in humans, genetic deficiency of AID leads to Type-2 Hyper-IgM Syndrome (HIGM2), an immunodeficiency characterized by high serum levels of IgM and lack of any other immunoglobulin isotype [10]. This phenotype was also found in AID knockout mice [8], demonstrating AID essential role in normal immunoglobulin maturation.

However, accumulating evidence suggests that AID function might not be limited to antibody diversification. A non-canonical role in antiviral activity has also been proposed, based on the observation that malignant transformation by Abelson murine leukaemia virus is greatly enhanced in AID-deficient mice [11]. It was also described that AID has a direct impact on Kaposi’s sarcoma-associated herpesvirus (KSHV) viral fitness by inhibiting lytic reactivation and by reducing infectivity of virions [12]. Furthermore, HBV infectivity in primary human hepatocytes was decreased by ectopic AID expression induced by many proinflammatory cytokines via NF-κB signaling. Interleukins have also been shown to enhance AID expression in B cells outside the germinal centers and human colonic epithelial cells [13,14]. Thus, AID expression is part of a broader inflammatory response following infection.

1.2.2 APOBEC 1

Apolipoprotein B editing complex 1 (APOBEC1) was the first member to be discovered as the posttranscriptional editor of apolipoprotein B (apoB) mRNA, where it specifically deaminates cytosine 6666 to uracil, producing a premature stop codon and resulting in a shorter protein product with a distinct function in the transport of lipids within the small intestine [15-17]. APOBEC1 expression is most abundant in mammalian small intestine and liver, correlating with its function [18]. Specific targeting to the RNA editing site is achieved through its interaction with ACF (APOBEC1 complementation factor), an RNA binding protein that recognizes a cis-acting 11 nucleotide AU rich mooring sequence 3’ of the edited site [19,20]. Recently RBM47 has been described as a novel interactor of APOBEC1 and A1CF, being expressed in tissues where C to U RNA editing occurs. Furthermore, RBM47 can substitute for A1CF, being necessary and sufficient for APOBEC1-mediated editing of ApoB mRNA [21].
A broader role for A1 in the regulation of mRNAs was first suggested by the discovery of A1-dependent editing of the tumour suppressor NF1 mRNA [22]. Later, a transcriptome-wide RNA sequencing screen comparing wild-type and A1-deficient mice revealed mooring sequence-dependent editing of cytidines to uridines within the 3’ untranslated region (3’UTR) of 32 mRNAs [23]. Besides, independent of editing activity, A1 binding to AU-rich mooring and 3’ UTR sequences may itself regulate mRNA stability, suggesting additional roles for APOBEC1 beyond its function in ApoB regulation [24].

Furthermore, under appropriate conditions, APOBEC 1 can also deaminate deoxycytidine in ssDNA as shown in E. coli mutator assays [25,26]. Unlike RNA editing, ssDNA editing by A1 did not require a cofactor. Supporting APOBEC1 DNA editing potential, its expression in neurons showed a protective function against herpes simplex virus involving ssDNA deamination of the viral genome [27]. Also, further evidence suggests A1 might also participate in the restriction of other viral infections and endogenous retroviral elements [28-30]. Overall, APOBEC1 is far a more versatile enzyme than initially thought, with roles in mRNA regulation, ssDNA editing and viral restriction.

1.2.3 APOBEC3s

In humans there are seven different APOBEC3 cytidine deaminases (A–C, DE, and F–H) mostly recognized by their role in innate immunity, restricting viral infection and endogenous retroelement transposition. They are preferentially expressed in lymphoid cells but can also be induced in other tissues by a variety of cellular signals, such as interferon [31].

The first member to be assigned a function in antiviral immunity was A3G by demonstration of its ability to restrict HIV infectivity [32]. In the primary described mechanism, A3G was shown to be incorporated into HIV virion particles through binding to both viral and cellular RNAs [33]. Deamination of cytosines on the nascent DNA strand during reverse transcription of viral RNA leads to either non-infectious virions or to the
degradation of hypermutated viral cDNA [34,35]. Furthermore, several reports have suggested a significant extent of restriction through alternative deaminase-independent mechanism that directly block or hinders reverse transcription priming and elongation [36,37]. The overall importance of A3G in host cell defence is accentuated by the HIV encoded accessory protein known as viral infectivity factor (Vif), whose primary function is to suppress the antiviral activity of A3G, mediating its ubiquitination and proteasomal degradation [38-41]. Along with A3G, other A3 deaminases such as A3D, A3F, and A3H, provide different levels of anti-HIV activity and are also counteracted by Vif [42].

In addition to HIV, A3 enzyme-induced mutations are also implicated in inhibition of a broad range of RNA viruses including other retroviruses such as human T-cell leukaemia virus type-1 (HTLV-1) and human foamy virus (FV), restricted mostly by A3G [43-46]

Furthermore, A3 proteins are also associated with restriction of DNA viruses. Hepatitis B Virus (HBV) genome is mutated at different levels by A3A, A3B, A3C, A3F and A3G [47]. Cytokine-mediated upregulation of A3A and A3B particularly results in the degradation of hepatocytes persistent HBV covalently closed circular nuclear DNA without apparent damage to the host genomic DNA [48]. A3C, A3H, and A3B deaminate human papillomavirus (HPV) genomes [49] while A3C has also been shown to mutate herpes viruses such as herpes simplex-1 (HSV-1) and EBV viruses [50]. Conversely, A3A alone has been shown to inhibit the parvovirus, adeno-associated virus type 2 (AAV 2) in a deaminase-independent manner [51,52].

Not surprisingly, APOBEC3s are also main inhibitors of endogenous retrotransposons which are essentially retroviruses integrated into the host genome with the ability to copy themselves into random new locations. They do so by an RNA intermediate and subsequent reverse transcription, providing an opportunity for APOBEC3 restriction. Most APOBEC3 proteins can abrogate the retrotransposition of autonomous long interspersed nuclear elements (LINE) and nonautonomous short interspersed nuclear elements (SINE) at varying rates. Inhibition is mostly through hypermutation of retroelement ssDNA or deaminase-independent mechanisms such as A3G sequestering of SINE RNAs as large ribonucleoprotein complexes in the cytoplasm [53,54]. Even long
terminal repeat (LTR) based elements, which are non-functional in humans, show A3F and A3G hypermutation footprints [55], supporting the notion that preventing propagation of these intracellular mobile elements corresponds to a primary function of APOBEC3 enzymes as an ancestral defence mechanism against retroelements.

Moreover, even though most A3 proteins use ssDNA as a preferred substrate for deamination, fairly recently A3A and A3G have been shown to edit different mRNAs, some of which are associated with viral pathogenesis [56,57]. This finding does not only extend the role of A3 proteins in innate immunity but also opens the possibility of yet-to-be identified new functions.

1.2.4 APOBEC 2 and APOBEC 4

APOBEC2 and APOBEC4 deaminase activity and physiological functions remain poorly understood. Even though APOBEC2 crystal structure was the first one to become available and used for homology modelling of other family members, it seems to be functionally quite different [58]. A2 has been shown to be essential in muscle development and to be involved in embryogenesis [59,60]. Interestingly, even though no role in viral restriction has been assigned to APOBEC2, its expression is enhanced by pro-inflammatory cytokines such as TNFα and IL-1β suggesting it could be involved in intrinsic immunity as other APOBECs [61]. However, APOBEC2 deaminase activity is still under debate. While it proved to be non-mutagenic in early yeast and bacteria mutator assays [26,62], more recent evidence suggests that APOBEC2 DNA mutagenic activity targets specific tumor suppressor genes [63]. APOBEC4 (A4) has no ascribed function, and its sequence is distinctly divergent from other APOBEC genes with significant alteration in the presumptive deaminase domain. As APOBEC2, APOBEC4 is non-mutagenic when expressed in yeast and bacteria [62]. But in both cases, specific cofactors might be needed for the mutagenic activity, or they may function through deaminase-independent mechanisms. Further studies will aid on the elucidation of the function and mutating potential of these less understood members of the APOBEC deaminase family.
Overall, AID/APOBEC cytidine deaminases have crucial and diverse physiological roles in immunity and gene expression control. However, the advantages of these beneficial activities are counterbalanced by the risk of exposing the host genome to active mutagenesis that can lead to unwanted point mutations and translocations potentially contributing to cancer development.

1.3 AID/APOBECs in Cancer

The first evidence for a link between AID/APOBEC deaminases and cancer came from transgenic animals, where overexpression of APOBEC1 (A1) in mice induced hepatocellular carcinomas [64,65], most likely as a side effect of its then proved ability to mutate DNA [25,26]. Moreover, soon after AID had been discovered as the key enzyme for SHM and CSR, it was shown to be responsible for the c-Myc/IgH translocations found in Burkitt’s lymphomas, possibly as a result of “mistargeted CSR” [66,67]. Besides, B cell lymphomas have been found to harbour mutations with SHM features in several non-Ig genes and sub-clonal heterogeneity at the Ig locus, demonstrating off-target DNA damage induced by AID as well as ongoing activity during disease progression [68,69]. Deregulated AID has also been associated with various other types of cancer (reviewed in [70]).

Furthermore, whole genome sequencing approaches have uncovered the contribution of other APOBECs to carcinogenic DNA damage. APOBEC3s mutation signature was found in multiple cancer types [71-73]. A study focusing on extensive sequencing across many breast cancer samples identified the accumulation of clustered APOBEC3A and 3B mutation signatures (termed kataegis, after the Greek kataegisa meaning ‘thunder shower’) [74]. Similar mutation showers were recapitulated by overexpression of A3B or A3A in yeast, supporting the role of this deaminases in characteristic tumour mutagenesis [72,75]. Moreover, the extent of dC>dT transitions in primary breast tumors were found to be correlated with elevated A3B expression [73], and both A3A and A3B are thought to be responsible for many of the non-clustered signature mutations observed not only in breast cancers but other tumour types [76]. Other
APOBEC3 proteins were also found to be highly expressed in various cancers (reviewed in [70]).

The reason for the extreme activation of this mutational process in cancer is unknown. A potential driver in some cases could be inflammation. Cytokines produced during viral infections trigger the expression of AID/APOBEC proteins which can cause collateral genome-wide mutations. In this regard, HCV, which is known to be a significant cause of hepatocellular carcinoma, has been reported to trigger AID expression in hepatocytes [14]. Furthermore, stable transfection of normal breast epithelial cells with HPV was demonstrated to cause APOBEC3B mRNA overexpression leading to a significant increase in DNA breaks [77]. Therefore, AID/APOBEC expression could be a critical link between oncogenic viral infection and malignant transformation.

In short, AID/APOBEC proteins can influence cancer development and progression, this threat along with the variable activities of each member of the family underscores the importance of thoroughly understanding how these deaminases are generally regulated and what are the potential situations where their regulation can go wrong.

1.4 Multilevel regulation of AID/APOBEC deaminases

1.4.1 AID/APOBECs controlled expression and subcellular localization

AID/APOBEC first level of regulation seems to be their primarily confined expression in tissues and cells where they exert their most relevant physiological functions (described in section 1.2 and sum in table 1). However, most members have been shown to be potentially ectopically expressed upon different stimuli such as viral infection or oncogenic processes. Hence further levels of regulation are crucial to prevent undesired genomic deamination.

Subcellular regulation allows cells to compartmentalize potentially genotoxic proteins such as AID/APOBEC deaminases. Both APOBEC1 and AID are actively imported into the nucleus to exert their functions on ApoB mRNA and Ig locus, respectively. AID, import is
accomplished through a noncanonical nuclear localization signal (NLS) and may also be influenced by two additional interacting proteins, CTNNBL1, and GANP [78-80]. Similarly, APOBEC1 enters the nucleus by binding at least one other cellular protein, ACF [81,82]. However, both deaminases are actively exported through strong nuclear export signals (NES) and appear cytoplasmic at steady state [83-85]. APOBEC1 is maintained in the cytoplasm as an enzymatically inactive 60S complex, while AID is retained through interactions with eukaryotic Elongation Factor 1-alpha (eEF1A) and heat shock protein 90 (HSP90) [79,86-88].

Unlike AID and APOBEC1, none of the APOBEC3 proteins have demonstrated nucleocytoplasmic shuttling activity [89,90]. A3B is predominantly nuclear due to an NLS that is similar to that of AID, however, it does not seem to shuttle. Being the only member with steady-state nuclear localization, A3B expression is kept at low levels in most normal tissues probably as a mechanism to restrict its access to genomic DNA [91]. Conversely, A3G is restricted to the cytoplasm through a robust cytoplasmic retention signal (CRS) and RNA binding that favours its association with cytoplasmic RNA processing centres, stress granules and p-bodies [92-94]. Anti-HIV paralogs A3D and A3F also show cytoplasmic localization that may as well be linked to their RNA binding activity [95]. A3A is cytoplasmic and non-genotoxic in monocytes; however, when transfected into other cell types it localizes throughout the cell and becomes genotoxic, suggesting potential interaction with nuclear import or cytoplasmic retention cofactors in different cellular contexts [96].

Overall, the data suggests that the level of expression and subcellular localization are critical regulatory mechanisms that maintain genomic integrity and control deaminase activity. The precise mechanisms and co-factors involved appear to differ among AID/APOBEC proteins and might further explain target recognition.
1.4.2 Cofactors and substrate nucleic acid recognition

APOBEC1 interaction with ACF and RBM47 is the only well-described example of protein cofactors directly involved in target recognition. Unlike this example of RNA editing, DNA targeting dependence on a protein cofactor has not been identified for APOBEC1 and is not fully established in the case of AID or the APOBEC3. Furthermore, the fact that AID/APOBEC proteins only act on ssDNA renders both host and viral genomes mostly inaccessible, except during replication or transcription when small areas become temporarily single-stranded. Thus active transcription has been repeatedly invoked as a requirement for deamination (reviewed in [97]). In relation to this, C-to-U editing of retroviral DNA takes place after reverse transcription, when the nascent ssDNA is transiently exposed due to template RNA degradation. Current DNA targeting models propose that AID binds paused/stalled RNA polymerase II complexes (RNA Pol II) to access target DNA [98]. In turn, RNA Pol II associates with the pausing/stalling cofactors Spt5 and RNA exosome, both of which stimulate AID function in B cells [99,100]. However, studies looking to whether off-target mutations favor certain loci revealed that AID/APOBECs preferentially mutate promoters of active genes in yeast in the absence of any additional protein targeting factors [101]. Deaminases could, therefore, target viral genomes or particular genes by their ability to recognize structural features of the DNA induced by the melting of the double strand during the initiation of transcription or by binding to non-substrate nucleic acids.

1.4.3 RNA binding role in AID/APOBEC regulation

Regardless of whether they use it as a substrate for deamination or not, all AID/APOBEC proteins share the intrinsic ability to bind RNA, underpinning a regulatory role for this interaction that it is not fully understood.

Early work revealed that RNA binding could inhibit or attenuate AID/APOBECs ssDNA deaminase activity [102,103]. In vitro studies have shown AID to have very little catalytic activity unless pre-treated with RNase. Furthermore, as described before, AID/APOBEC cytoplasmic retention is partly achieved by binding to cellular RNAs that
bridge monomers together into inactive high molecular weight ribonucleoprotein complexes [95,104-106]. Conversely, recent studies have shown that APOBEC3H forms a dimer around a short RNA duplex and, despite the bound RNA, has potent cytidine deaminase activity [107,108]. In the case of A3G, RNA binding also mediates its incorporation into HIV particles, allowing subsequent restriction either by deaminase dependent mechanisms after RnaseH removal of viral RNA or deaminase-independent mechanisms by directly blocking transcription [36,109]. Moreover, AID restriction of HBV has also been shown to be dependent on its ability to bind viral RNA [110].

Thereby, evidence supports a role for RNA in AID/APOBEC regulation, however little is known about the RNA binding specificity of this family of deaminases. Recent evidence has suggested that in the case of APOBEC3G and APOBEC3F binding to RNA is not specific [111], yet different studies have also shown selectivity for particular cellular and viral RNAs [33,112-114]. In the case of AID, binding to structured switch-repeat encoded RNA has recently been proposed as a mechanism for its recruitment to the switch regions of immunoglobulin genes [115]. Previous in vitro studies from our group have also shown that AID is capable of binding structured RNAs (such as tRNAs) and preferentially targets tRNA yeast promoters for mutation, suggesting that RNA binding could favor recruitment to specific loci [101]. This data further highlights potential regulation by RNA and the need to better understand the RNA binding preferences of different AID/APOBECs.

So far, most studies that have addressed RNA binding to AID/APOBEC proteins have either been entirely in vitro, using a limited set of RNAs or focusing only on one member of the family. Therefore, my thesis work explores the RNA binding specificities of different AID/APOBECs in vivo with the help of an unbiased high-throughput sequencing based technique.
1.5 High-throughput sequencing methods to study protein-RNA interactions

1.5.1 Clip-based methods

Gaining information on RNAs bound or recognized by different proteins constitutes a powerful tool to elucidate the role and mechanisms behind RNA-protein interactions. A single protein can bind a multitude of RNAs with different degrees of specificity and be influenced by competition between targets or additional factors; therefore, the development of sensitive techniques for the discovery of this interactions is crucial.

Multiple approaches have been developed for the identification of functional RNA-protein interactions in vivo. Most of them involve pulling down RNA binding proteins and the analysis of co-precipitating RNAs by real-time PCR, microarrays or even high throughput sequencing. These strategies have proven to be useful; however, they cannot discriminate direct from indirect interactions, nor can they identify RNA binding sites. Furthermore, cell lysis can result in the dissociation of sometimes labile RNA-protein interaction and in vitro re-association, leading to false positives and potential loss of real binding patterns [116]. Alternative approaches involve the chemical crosslinking of RNA and proteins with formaldehyde. However, formaldehyde high reactivity with both DNA and RNA makes it hard to specifically isolate only one protein and its respective RNA target.

To overcome this, CLIP (crosslinking immunoprecipitation) techniques were developed, taking advantage of the ability of UV-irradiation to penetrate intact cells or tissues and induce covalent crosslinks between RNA nucleobases and proteins carboxyl groups that are in direct contact (~1 Ångstrom apart) [117]. Stabilization of the RNA-protein complexes allows more stringent purification steps, significantly reducing the signal-to-noise ratio. Furthermore, the crosslinking is done before the cell lysis occurs, which allows the endogenously occurring protein: RNA interactions to be recovered. Subsequent steps of CLIP protocol allow specific purification of the RNAs bound by the protein of interest. First, after irradiation with UV-C light, protein-bound RNAs are partially digested to obtain RNA fragments in an optimal size range. The succession of
immunoprecipitation, ligation of a 3’ adapter and SDS-PAGE allow removal of free RNA and specific extraction of RNA: protein complexes. The bulk protein is removed by proteinase K treatment, which leaves a short peptide remaining in the RNA at the crosslink position. Addition of a 5’ adaptor and the subsequent reverse transcription, prepare for the final amplification of the library by PCR. The library is then sequenced on a high-throughput sequencing platform, for this reason, the traditional CLIP protocol is often also referred to as HITS-CLIP or CLIP-seq. Sequencing data provides a means of identifying the bound RNAs, and importantly the position of protein binding. This is achieved by the fact that the reverse transcription reaction causes an increased incidence of substitutions and deletions at the crosslink site due to reading through the remaining bound peptide. Therefore, bioinformatic analysis can be used to determine the binding site of the protein of interest. Analysis of cDNA deletions in Nova CLIP demonstrated that they were located at YCAY motifs, which had been previously described by other methods, confirming that cDNA deletions in CLIP can identify protein-RNA cross-link sites [118].

A drawback of this original version of CLIP is that it strictly requires reverse transcription to proceed from a universal 3’ ligated adapter to a 5’ ligated adapter since both of them are required for PCR amplification. However, analysis has shown that over 80% of cDNAs are truncated and lack the 5’ adapter and are therefore not amplified leading to loss of valuable information. This is due to stalling of reverse transcriptase at the short polypeptide left at the UV-induced crosslink site. To address this issue, individual-nucleotide resolution CLIP (iCLIP), enables PCR amplification of truncated cDNAs increasing the overall cDNA yield and thereby identifying protein–RNA crosslink sites more comprehensively.
Figure 1.3 A schematic description of the iCLIP protocol. The protocol comprises immunoprecipitation, RNAse I treatment and ligation of the 3’ adaptor, as well as separation of the radioactively labelled RNA (step 1-8). The immunoprecipitated RNA:protein complex is subsequently digested with proteinase K to remove the bulky protein and prepare the RNA for the reverse transcription (step 9). cDNAs are generated (step 10) that truncate as soon as the reverse transcriptase encounters the remaining polypeptide. After circularization, linearization, PCR amplification and sequencing (step 12-16), the position of the crosslink is at the first position after the barcode. Taken from [119].

This is achieved by significant changes in the library preparation. After immunoprecipitation of the protein-RNA complex, a linker ligation at the 3’ end is performed. Instead of adding an extra linker to the 5’ end of the RNA, as used in CLIP, an adapter is introduced by an overhang in the primer during reverse transcription. A circularization step allows for the attachment of the linker to the cDNA and the digestion with a restriction enzyme causes the linearization of the cDNA enabling efficient amplification of truncated cDNAs (Figure 1.3). Furthermore, to increase the quantitative
power of the modified method, the reverse transcription primers contain a randomized sequence (random barcode), which enables computational filtering of artefacts caused by variable PCR amplification of the cDNAs. Overall iCLIP exploits the stall of reverse transcription at the crosslink site by making it possible through its modifications to amplify truncated cDNAs and define the position of the crosslink as the nucleotide preceding the start of the sequenced iCLIP fragment.

1.6 Aim of the thesis

Making use of the resolutive power of iCLIP technique, in my thesis work, I aim to determine and compare the RNA binding specificity of different deaminases \textit{in vivo} in an unbiased way, to aid on the understanding of the role of this interaction.
Chapter 2

Materials and Methods

2.1 Materials

2.1.1 Antibodies

Anti-Flag M2 (Sigma, F1804), anti-HA .11 (Biolegend, 901503), anti-Histone H3 (Abcam, ab1791), anti-actin (Santa cruz, C4), Anti-FLAG M2 Magnetic Beads (Sigma, M8823), Monoclonal ANTI-FLAG M2-Peroxidase (HRP) (Sigma, A8592), Anti-HA tag antibody (HRP) (Abcam, Ab1190), Goat Anti-Mouse IgG H&L (Alexa Fluor 488) (Abcam, Ab150117),

2.1.2 Plasmids

For yeast experiments cDNAs coding for human AID or the C-terminal domain of APOBEC3G (cA3G) were inserted into galactose-inducible pRS426-derived expression vectors [120]. Both deaminases were N-terminally HA-tagged and fused with a 5” SV40 nuclear localization sequence. An empty vector (EV) was used as a control.

For HEK293 experiments C-terminally FLAG full length human AID, APOBEC1, APOBEC3A, APOBEC3B or APOBEC3G cDNAs were cloned by PCR into GFP-expressing bicistronic pCIG plasmid. An EV was used as a control.

2.1.3 Buffers

All buffers were prepared by mixing stock solutions and subsequently filtered to prevent contamination.

<table>
<thead>
<tr>
<th>Buffer</th>
<th>Ingredients</th>
</tr>
</thead>
<tbody>
<tr>
<td>One step transformation buffer</td>
<td>40% PEG3350</td>
</tr>
<tr>
<td></td>
<td>0.2 M Lithium Acetate</td>
</tr>
<tr>
<td>Buffer Type</td>
<td>Composition</td>
</tr>
<tr>
<td>---------------------</td>
<td>--------------------------------------------------</td>
</tr>
</tbody>
</table>
| **Lysis buffer**    | 10 mM Tris-HCL, pH 7.4  
100 mM DTT  
50 mM Tris-HCL, pH 7.4  
100 mM NaCl  
1% Igepal CA-630  
0.1% SDS  
0.5% sodium deoxycholate  
On the day: 1X complete Protease Inhibitor Cocktail (Roche, UK) |
| **High-salt wash buffer** | 50 mM Tris-HCL, pH 7.4  
1 M NaCl  
1 mM EDTA  
1% Igepal CA-630  
0.1% SDS  
0.5% sodium deoxycholate |
| **PNK buffer**      | 20 mM Tris-HCL, pH 7.4  
10 mM MgCl₂  
0.2% Tween-20 |
| **5x PNK pH 6.5 buffer** | 350 mM Tris-HCL, pH 6.5  
50 mM MgCl₂  
5 mM DTT |
| **4x Ligation buffer** | 200 mM Tris-HCL, pH 7.8  
40 mM MgCl₂  
4 mM DTT |
| **PK buffer**       | 100 mM Tris-HCL, pH 7.4  
50 mM NaCl  
10 mM EDTA |
PK buffer + 7 M urea | 100 mM Tris-HCL, pH 7.4
| 50 mM NaCl
| 10 mM EDTA
| 7 M urea

### 2.1.4 RT primers for the iCLIP protocol

A list of RT primers used for the generation of iCLIP libraries is given below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence (IDT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rt1clip</td>
<td>NNAACCNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt2clip</td>
<td>NNACAANNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt3clip</td>
<td>NNATTGNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt4clip</td>
<td>NNAGGTNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt6clip</td>
<td>NNCCGNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt7clip</td>
<td>NNCTAANNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt8clip</td>
<td>NNCATTNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt9clip</td>
<td>NNGCCANNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt10clip</td>
<td>NNGACCNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt11clip</td>
<td>NNGGTNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt12clip</td>
<td>NNGTGNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt13clip</td>
<td>NNTCCGNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt14clip</td>
<td>NNTGCCNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
<tr>
<td>Rt15clip</td>
<td>NNTATTNNNAGATCGGAAGAGCGTCGTGatcCTGAACCCGC</td>
</tr>
</tbody>
</table>
2.2 Basic Molecular biology methods

2.2.1 Polymerase Chain reaction (PCR)

PCR reactions for cloning purposes were carried out using the Phusion master mix and according to the NEB protocol. The following PCR conditions were applied:

<table>
<thead>
<tr>
<th>Step</th>
<th>Temperature</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denaturation</td>
<td>98°C</td>
<td>2 min</td>
</tr>
<tr>
<td>35 cycles</td>
<td>98°C</td>
<td>5s</td>
</tr>
<tr>
<td></td>
<td>58-60°C</td>
<td>10s</td>
</tr>
<tr>
<td></td>
<td>72°C</td>
<td>30 s/kb</td>
</tr>
<tr>
<td>1 cycle</td>
<td>72°C</td>
<td>3 min</td>
</tr>
</tbody>
</table>

2.2.2 PCR purification and visualization

All PCR reactions were purified using QIAGEN PCR purification kit according to the manufacturer’s protocol. PCR reactions were eluted with TE buffer. 5X loading buffer was added to the DNA samples to increase the density of the mixture and allow visualization. Agarose gel electrophoresis was used to separate DNA according to size at 120 volts. DNA was stained with SYBR®Safe DNA gel stain (Invitrogen) and bands were visualised by UV light exposure. The QIAQuick Gel Extraction kit was used to extract DNA fragments of the correct size from agarose gels.
2.3 Yeast and cell maintenance

2.3.1 Cell Culture

HEK293 cells were maintained in Dulbecco’s Modified Eagle Medium (DMEM) with 10% fetal bovine serum (FBS), supplemented with 100 U/mL penicillin, 100 µg/mL streptomycin at 37 °C in a 5% CO₂ atmosphere.

2.3.2 Transient transfection

HEK293 cells were seeded to achieve 70-90% confluency before transfection in either 6 well plates (fluorescence microscopy) or 10cm dishes (RNA based experiments). Depending on the size of the transfection, 700 or 1000 µL of DMEM (6 well plate or 10cm dish respectively) were mixed with 7 or 10 µg of plasmid DNA and combined with 100 or 150 µL of Lipofectamine mixed with 600 or 850µL Opti-Mem medium respectively to obtain a 1:1 ratio of the diluted plasmid DNA and the diluted Lipofectamine. The mixture was incubated for 5 minutes and added to the respective plate in a drop-wise fashion and incubated for 24hs.

2.3.3 Yeast Culture

Yeast strain BY4741 (MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0) from Euroscarf (Frankfurt, Germany) was grown on YEPD agar plates at 30 °C.

2.3.4 Yeast transformation

Single colonies from YEPD plates were mixed with 100 µl of one step buffer and 3 µl of galactose-inducible expression vectors of either AID, cA3G or the empty vector (EV). The mix was left for 30 minutes at room temperature and then incubated for 30 minutes at 42 °C. Cells were then spin down, resuspended in water, plated in -URA YEPD agar plates for selection of transformants and grown at 30 °C.
2.4 iCLIP

All iCLIP experiments were performed according to Huppertz et al. Steps 1-3 (Figure 1.3) and input material were specifically adapted for its use in yeast as described below.

2.4.1 Input material for yeast adapted iCLIP

Yeast transformants of AID, cA3G and the EV control were grown in 5ml of YC (-) Arg (-) URA (+) GLUCOSE medium shaking at 300rpm at 30 °C overnight. To scale up and induce vector expression this starter culture was seeded into 250 ml of YC (-) Arg (-) URA (+) GALACTOSE medium and grown overnight. Cells were then pelleted at 4 °C and resuspended in PBS. Before and after induction an aliquot was obtained to check efficient expression (Figure 2.1). Resuspended cells were placed on plates on ice and crosslinked with 2.4 J/cm2 254 nm UV light (crosslinking optimization for yeast is described in more detail in chapter 3). Lysis was performed using glass beads and vortexing at 4 °C to preserve RNA integrity. Lysates were then cleared by centrifugation and subjected to partial RNAseI digestion as described in chapter 3. Further steps were done following the original protocol by Huppertz et al. [119].

![Western Blot Analysis](image)

**Figure 2.1 AID and cA3G induced expression**
Western Blot analysis of AID and cA3G expression in yeast transformants before (-) and after (+) galactose (GAL) induction. Anti-H3 was used as a loading control. The empty vector (EV) control showed no bands either before or after induction.
2.4.2 Input material for HEK293 iCLIP

Cells were transiently transfected with deaminase expression plasmids in 10 cm diameter plates (~10–12 million HEK293 cells). After 24hs cells were crosslinked with 0.15 mJ/cm² 254 nm UV light (UV-C) on ice, in first experiments a non-crosslinked plate was left as a control. After collection, cells were lysed, sonicated and the total RNA pool was partially digested with RNase I to produce fragments of an optimal size range for library preparation. RNase dilution optimization is described in detail in chapters 3 and 4.

2.4.3 General iCLIP protocol

Steps after RNase I digestion were the same regardless of the input material. A detailed protocol with all relevant steps and reagents based on Huppertz et al. is given in Appendix 1. In general, after partial RNase I digestion lysates were incubated with anti-HA or anti-Flag antibody and deaminase-RNA complexes were immunoprecipitated using magnetic beads. RNA was then dephosphorylated, and an adapter was ligated to the 3’ end while the 5’ end was radioactively labelled. RNA was visualised on a nitrocellulose membrane after SDS-PAGE. Autoradiograph was used as a first quality control and as a guide for RNA extraction from the membrane. The protein-RNA complexes were extracted and digested with proteinase K to remove the bulky protein. This procedure caused remnants of the protein to remain at the crosslinking sites, which terminated the reverse transcription reaction at this specific site. Before reverse transcription using superscript III (Invitrogen), RNA was purified and precipitated by phenol/chloroform. After reverse transcription (RT), primers were removed from the sample by size selection in a 6% TBE-UREA gel. cDNA products were extracted from three different segments of each lane corresponding to 70–80 nt, 80–100 nt and 100–150 nt fragments. cDNA products were then circularised, ligating the primer to the 5’ end of the cDNA with CircLigaseII (Epicentre Biotechnologies, CL9025K). An oligo complimentary to the cleavage site was then annealed to the circular cDNA (5’-GTTCAGGATCCACGAGCTCTCT CAAAAT-3’) for subsequent linearization by digestion with restriction enzyme BamHI. This allowed accessibility to both cDNA ends for final PCR library amplification using primers compatible with Solexa sequencing (5’-CAAGCAGAGACGGCATACGA GATCGGTCTCGG CATTCTGCTGAACGCGATCTCCTTCCGATCT-3’; 5’-AATGATACGGCGACCACGAGATCTACACTCTTTC
CCTACACGAGCTCTTCCGATCT-3’). Final libraries were quantified using qPCR-based KAPA Library Quantification kit before sequencing.

2.5 Protein-based techniques

2.5.1 Preparation of protein lysate

Cells were lysed by stringently resuspending the cell pellet in 300 µl RIPA buffer (Life Technologies) for 30 minutes on ice and regular pipetting every 10 minutes. Lysed cells were then centrifuged at 14.000xg for 5 minutes and the supernatant was transferred to a new tube.

2.5.2 Protein Concentration measurement

The protein concentration was measured using the Pierce BCA assay. The concentration of each sample was measured using three technical replicates and compared to the albumin standard protein samples ranging from 0 to 2mg/mL.

2.5.3 Western Blot analysis

Each sample was diluted to equal concentrations (20µg), mixed with NuPage LDS sample buffer (with final 10 mM DTT concentration) and heated 70 °C for 5 minutes and loaded on a 4-12% Bis-Tris pre-cast NuPage gel with a Pre-stained protein ladder. The SDS-PAGE was performed according to the manufacturer’s manual (Life Technologies) and the subsequent blotting was performed at 30V for 2 hours onto a nitrocellulose membrane. The membrane was blocked with PBS-T containing 5 w/v% milk powder at 4 °C overnight. Afterwards the membrane was incubated with the respective primary antibody in 3 w/v% BSA at room temperature for one hour or overnight in a cold room. After washing the membrane 3 times with PBS-T, secondary antibody conjugated with horseradish peroxidase was added and membrane was incubated for 1-2hs at room temperature. Finally, the membrane was washed three times and developed with ECL solution (GE Healthcare).
2.6 Fluorescence Microscopy

HEK293 cells were grown on cover slips in 6 well plates, they were then transiently transfected with AID/APOBEC expression plasmids and grown at 80% confluency. After 24hs medium was removed and cells were fixed in cold methanol for a minimum of 20 minutes. Cells were then blocked with 1% BSA in PBST (PBS+ 0.1% Tween 20) for 1hr, and subsequently incubated with anti-FLAG antibody overnight. After three washes with PBST, secondary antibody coupled with Alexa488 fluorophore was added and incubated for 1hr. Cells were washed and nuclei were stained with DAPI. Finally, coverslips were mounted on microscope slides with 20 µl of ProLong Gold antifade (Invitrogen). Cells were visualised using the inverted confocal laser scanning microscope LSM 710. For each coverslip, images were taken at a magnification of 64x, using a zoom factor of 1.0. All scanned images had a resolution of 1024 x 1024 at a pinhole setting of one Airy unite, 8X line-averaging mode, and a pixel depth of 16 bits. Images were further analysed with FIJI.

2.7 Computational methods

2.7.1 Sequencing and mapping

High-throughput sequencing of iCLIP cDNA libraries was performed on an Illumina HiSeq400 (single-end 50bp). All iCLIP libraries contained a 4-nt experimental barcode in addition to a 5-nt random barcode to enable multiplexing and removal of PCR duplicates. The genomic analysis was based on the yeast genome version SacCer3 and the human genome version hg19/NCBI37 with annotations taken from Ensembl (version 60). All iCLIP libraries were mapped using Bowtie with the same parameters, as mapping inconsistencies lead to artificial increases or decreases in observed binding frequencies. Only reads that aligned once with the possibility of 1 mismatch were considered for further analysis. Both the sequence mapping and its annotation were carried out by the iCount server (http://icount.biolab.si), which has been implemented by Dr Tomaz Curk [121].
2.7.2 Identification of significant iCLIP crosslink sites

iCLIP reads were associated with expressed genomic regions as defined by ENSEMBL hg19 release of human genome. Both coding and non-coding genes were included. Introns, 5’ UTR, ORF and 3’ UTR were considered as separate regions. iCLIP reads antisense to the transcriptional direction of the associated gene, reads that mapped to non-annotated genomic regions, and reads mapping more than once were removed before proceeding to further analysis. Crosslink sites were defined as the nucleotide preceding the start of the sequenced iCLIP cDNA (read) and cDNA counts were calculated as the number of unique reads starting at the same position (Figure 2.2). To identify significant crosslink positions, a control file with random placement of iCLIP reads on corresponding genes was generated 100 times. cDNA values in iCLIP or randomised positions were summed for locations up to 15 nt apart, and the resulting values were considered the ‘height’ of each crosslink site. For a particular height, h, the associated probability of observing a height of at least h was \( P_h = \Sigma n_i(i = h:H)/N \). The modified FDR for a cross-link nucleotide with height h was computed as \( \text{FDR}(h) = (\mu_h + \sigma_h)/P_h \), where \( \mu_h \) and \( \sigma_h \) are the average and standard deviation, respectively, of \( P_h, \text{random} \) across the 100 iterations. Within each region of a gene (intron, 5’ UTR, 3’UTR, ORF, ncRNA), the smallest height that gave an FDR < 0.05 was defined as the threshold height (\( h^* \)). Crosslink sites at positions satisfying \( h > h^* \) were considered significant. This method allows the calculation of significant crosslink enrichment over background signal in surrounding areas on the same gene (Figure 2.3).

2.7.3 Addition of tRNA genes

Given the multi-copy nature and widespread occurrence of tRNA genes across the human genome, their annotation is particularly challenging, being not included in iCount server output gene summary files. For this reason, they were separately added by overlapping the AID/APOBEC HEK293 iCLIP crosslink position files with predicted individual tRNA genes from tRNAscan-SE v.1.23 database.
Figure 2.2. Schematic description of iCLIP crosslink sites identification with nucleotide resolution
Crosslink sites were defined as the nucleotide preceding the start of iCLIP reads. Random barcodes (examples shown on the left) were used to identify unique iCLIP cDNAs (number in brackets indicates the number of PCR duplicates as an example). Bar graphs show the cDNA count (number of cDNAs at each crosslink site) which might differ depending on the efficiency of library production. Sequence of a random gene shown above with crosslink nucleotides highlighted in red.

Figure 2.3. Schematic representation of significant crosslinks identification
A representative iCLIP track on a gene taken from UCSC were peaks represent crosslink sites. Only crosslinks with an FDR < 0.5 were consider significant (shown in green), the rest were considered as background events (shown in red).

2.7.4 Heatmaps

A genes by protein “hit” (hit=cDNA count=crosslink) matrix was created by overlapping each deaminase low FDR specific crosslink position files with known genes database, including tRNAs. Each resulting cell in the hit matrix denoted the number of times the relevant deaminase crosslinked to a gene. In order to allow downstream analysis of differentially sized libraries, the hit counts for each gene were normalized towards the grand median calculated.
taking into account the total hit counts for each deaminase. Normalization was then done by multiplying each deaminase gene hits by the ratio of the median over each corresponding deaminase total hit count. The hit matrix was then log2-transformed, centered and scaled per gene such that the mean is 0 and the standard deviation is 1. Finally, a subset was taken by row selecting the top 100 most variable genes. When generating the heatmap, correlation was used to obtain the distance matrix while clustering was achieved by applying the average agglomeration method.

For the second heatmap the same initial matrix was used but AID replicates were merged into one by taking the row averages (hits per gene) of AID1 and AID2. Normalization was done as before and the top 20 marker genes for each deaminase were determined by identifying genes highly bound by one protein when compared to others. To generate the final heatmap correlation was used to obtain the distance matrix.

**Figure 2.4. General analysis workflow for iCLIP libraries**

The workflow starts with the unprocessed input sequencing data. Firstly, sequencing adapters are removed, reads are then mapped to the reference genome using Bowtie2 version 2.1. Only reads that align once with the possibility of 1 mismatch are kept. Using the random barcode added during RT, unique cDNA products are discriminated from PCR duplicates. Then crosslink site is identified as the nucleotide preceding the start of the read. Significant crosslink sites are defined by FDR algorithm (explained in section 2.7.2). Sequence mapping, annotations, identification of significant crosslink sites are carried out by the iCount server which generates output files of binding per segment type (eg: 3’UTR) and per gene. In both cases given their difficult annotation, tRNAs were added to these output files before further analysis.
Chapter 3

Analysis of RNA binding preferences of AID and the catalytic domain of APOBEC3G in yeast by iCLIP

AID/APOBEC deaminases are key components of innate and adaptive immunity, restricting viral infection and diversifying antibodies [122]. These functions are mostly related to their ability to deaminate cytidine to uridine, acting either on ssDNA or RNA of either the host or the pathogen. As mutators, these beneficial roles are counterbalanced by the risk of collateral genomic damage, frequently associated with cancer development. Therefore, it is crucial to understand the molecular interactions that regulate their activity and determine their editing site fidelity or off-target mutagenesis.

RNA binding has been suggested to be a potential regulatory mechanism since most AID/APOBEC deaminases retain the ability to bind RNA even when not used as the preferred substrate for deamination. However, very little is known about the RNA binding specificity of different AID/APOBEC deaminases and if distinct binding could account for their diverse functions, localization, and activity. Therefore, determining RNA binding preferences would aid on the understanding of the potential regulatory role of this interaction as well as on elucidation of new potential editing targets.

Previous in vitro studies have shown that AID efficiently binds small structured RNAs (such as tRNAs) while being poorly recovered by extended linear polyU RNA. Unlike AID, APOBEC3G showed enhanced binding to polyU, most likely reflecting its ability to oligomerize in an RNA dependent fashion. Conversely, the c-terminal domain of APOBEC3G (cA3G) was not recovered by any of the RNAs used in the study. This data correlates with the described enhanced mutational preference for small RNA genes (tRNAs, snoRNAs, and snRNAs) in yeast. Even though both AID and cA3G target small RNA genes, AID has a more pronounced effect, suggesting that the observed distinct RNA binding could account for AID enhanced recruitment [101].
These observations support the relevance of understanding AID/APOBEC deaminases direct interaction with RNA and suggest a degree of RNA binding specificity. However, this and most described AID/APOBEC RNA binding preferences are restricted to a limited set of RNAs and in-vitro conditions, highlighting the need to further assess the RNA binding specificity of different AID/APOBEC deaminases in a more comprehensive robust way. As previously described, iCLIP (individual nucleotide resolution crosslinking and immunoprecipitation) developed by König et al., is a powerful tool that allows efficient identification of RNA-PROTEIN interactions in living cells. Through a covalent bond induced by UV-C crosslinking, the interaction between RNA and protein is stabilized, enabling specific immunoprecipitation of the protein of interest bound to its target RNAs. Thereby, not only the bound RNA molecules are identified, but also binding sites and motifs can be extracted. Therefore, my thesis work aims to take advantage of this robust technique to describe and compare AID/APOBEC deaminases binding specificity in an unbiased way in vivo.

This chapter focuses on the adaptation of the iCLIP method for its use in yeast as a first approach to corroborate RNA binding and answer the general question of whether different AID/APOBECs have different RNA binding preferences in vivo. Yeast cells have been used in previous AID/APOBEC studies and have proven to be a useful model system, faithfully recapitulating the mutation signatures observed in mammalian genomes. Furthermore, the smaller and less complex transcriptome and the absence of additional regulatory mechanisms for AID/APOBEC deaminases facilitate data analysis while still providing a significantly varied pool of RNAs for the identification of RNA binding preferences in vivo. Taking the previous in vitro and yeast results as a reference and start point and given the multistep complexity of the technique, only AID and cA3G were used for this preliminary approach and optimization.

3.1 iCLIP optimization in yeast and library construction for AID and cA3G

For the construction of iCLIP libraries, yeast transformants expressing HA-tagged galactose-inducible human versions of AID, the C-terminal domain of APOBEC3G (cA3G) and an empty vector (EV) were generated. All constructs contained a nuclear localization sequence to ensure equal distribution of the deaminases in the cell and to enhance their
encounter with RNAs. After 24hs, galactose-induced cultures of the yeast transformants were collected, the pellets were resuspended in cold PBS, placed on 10cm petri dishes on ice and subjected to UV-C cross-linking.

**3.1.1 Adjusting UV-C crosslinking**

UV-crosslinking induces the formation of covalent bonds between RNA and protein, capturing the in vivo binding pattern. High-resolution information on the binding of the studied protein is obtained as the cross-linking event only occurs when both RNA and protein are in very close proximity [123]. Since the original published iCLIP protocol by Huppertz et al., is based on mammalian cell lines, the intensity and duration of UV irradiation had to be optimized for its use in yeast [119]. Different intensities were tested using AID yeast transformants. The lowest intensity showed no RNA signal, while irradiation with 1.2 J/cm2, 2.4 J/cm2, and 3.6 J/cm2 allowed detection of an RNA-AID smear upwards starting around 30 kDa close to the molecular weight of AID (Figure 3.1A). In order to avoid potential RNA damage due to extended UV exposure, final libraries for both AID and cA3G in yeast were performed using 2.4 J/cm2, the lowest irradiation showing efficient crosslinking with a strong signal. This corresponds to 10 minutes of irradiation time, therefore, to allow cooling of the cells and redistribution in the liquid layer, irradiation was interrupted every 5 minutes and dishes were carefully rocked for 30 seconds.

**3.1.2 Partial RNAse digestion**

Cells were then collected, pelleted and washed for subsequent lysis. The lysis procedure also had to be adapted to allow efficient disaggregation of yeast cell wall while still preserving RNA integrity before the controlled partial RNAseI digestion. For successful library preparation, the conditions of the partial RNAseI digestion need to be carefully optimized for each protein since it has a significant effect on the RNA recovery and different proteins might bind different amounts and RNAs that vary in size (different or distinct RNA affinity). Purified RNAs length should be in the range between 50 and 300 nucleotides, since most often (>80%) the reverse transcriptase stalls at the short
polypeptide attached to the RNAs giving cDNA molecules that will be in average half the size [124]. Therefore, to ensure cDNA molecules within the optimal size range (100-250 nucleotides) for a high sequencing yield, adjusting the partial RNase digestion constitutes a critical step. To determine the optimal RNase concentration needed for AID and cA3G, decreasing dilutions ranging from 1:50 to 1:500 were tested. Digestion for 3 minutes with 1:250 dilution appeared to be the most effective treatment for both deaminases, generating a diffuse but uniform signal of deaminase–RNA complexes starting upwards from 30 kDa, which is close to the molecular weight of both AID and cA3G (Figure 3.1B). Higher RNase concentration (1:50) leads to a sharp band of deaminase–RNA complexes close to 30KDa, which corresponds to RNA molecules smaller than 60 nt that contribute less to size-shift introduced by crosslinking. Conversely lower RNase concentration of 1:500 resulted mainly in RNA molecules longer than 200 nt that seem to reduce the overall signal, probably due to obstructed gel migration or membrane transfer of large size protein–RNA complexes. These results demonstrate the importance of optimizing the RNase concentration.

3.1.3 Library preparation for AID and cA3G

3.1.3.1 cA3G efficiently binds RNA in-vivo in yeast

After partial RNaseI digestion with the optimized dilution, deaminase–RNA complexes were immunoprecipitated using anti HA antibody. An adapter was then ligated to the 3’ end of the RNA while the 5’ end was radioactively labelled allowing visualization of the deaminase–RNA complexes by autoradiography (Figure 1.3 steps 1-7). As shown before for the RNase optimization, a smear upwards the expected protein size was detected for both deaminases, representing the RNA fragments crosslinked to cA3G and AID. This proves that in vivo cA3G is able to bind RNA as efficiently as AID, contrasting the previous in-vitro experiments in which cA3G had no RNA binding activity [101]. The smear for both deaminases extends up to 260 kDa showing that the bound RNA molecules vary in size and cause a significant shift in the protein retardation on the gel. The absence of signal in both the EV and the non-crosslinked sample proved the lack of unspecific background signal for crosslinking technique (Figure 3.1C).
For library preparation, RNAs that migrated at higher molecular weights than the deaminases were recovered from the membrane, enabling the removal of free RNA. Next, using primers that anneal to the linker, RNAs were reverse transcribed to cDNAs that were subsequently circularized and linearized. This strategy allows the generation of cDNAs where the unknown sequence is flanked by linkers allowing PCR amplification of the library (Figure 3.1.D).

(A) UV-C crosslinking optimization on AID. Autoradiography showing radioactively labelled AID-RNA complexes. Increasing crosslinking energies were tested including (-) no crosslinking and (+) 1.2 J/cm² (++) 2.4 J/cm² and (+++) 3.6 J/cm². The size of non-crosslinked AID is marked by an arrowhead. The smear represents efficient crosslinking. (B) Optimization of RNase I concentration is crucial to obtain optimal fragment sizes. Autoradiograph showing AID and cA3G protein–RNA complexes that were treated with decreasing concentrations of RNase I before immunoprecipitation. (C) Autoradiograph of AID and cA3G final library preparation using optimized UV-C and RNase dilution (1:250). (D) Final cDNA libraries for cA3G and AID show bands ranging from 150-250 bp after PCR amplification, which were multiplexed and sequenced.
The optimal number of PCR cycles was carefully optimized looking for the best compromise between over-amplification of secondary products and weak amplification of the band of interest (Figure 3.1.D). cDNA libraries are split into three fractions of different lengths coming from different fractions from the purification gel step (Figure 1.3 step 15). Libraries were multiplexed and sequenced on an Illumina MiSeq platform. Even though the quality of the library seemed sufficient, the sequencing yield was not in the optimal range. Nevertheless, analysis workflow was carried on aligning the reads to the yeast genome, and significant binding tendencies were found even with such a limited data set.

### 3.2 AID and cA3G preferentially bind yeast tRNAs in vivo

Sequencing reads were aligned to the yeast genome to identify bound RNAs. The random barcode introduced into iCLIP cDNAs allows a quantitative analysis of AID and cA3G RNA binding distribution among different segment types in the yeast genome (ORF, intron, intergenic and ncRNA). Noticeably, almost half of all reads from AID library and a high proportion of cA3G reads mapped to ncRNAs. Furthermore, when correcting for the size of the segment type, by dividing the percentage of deaminase-bound RNA by the percentage of the nucleotides of each category in the genome, an even more overwhelming selective enrichment for ncRNAs was found for both deaminases (Figure 3.2A-B).

Interestingly, when looking at the distribution of iCLIP reads in ncRNA subtypes, it was found that the highest proportion of cDNAs, in both AID and cA3G libraries, corresponded to tRNAs. (Figure 3.2C). This, together with the fact that rRNA, most abundant RNA in any cell, was not the main recovered RNA, proves that AID and cA3G preferential binding to tRNAs does not reflect the overall abundance of RNAs. Moreover, despite the still relative abundance of tRNAs in cells, CLIP experiments in general consistently show a paucity of binding to tRNAs [125-127].

In the case of AID, this is in accordance with the previously described enhanced targeting of small RNA promoters by AID and the preferential binding of structured tRNAs in vitro [101]. Furthermore, AID iCLIP also showed enhanced binding to snoRNAs that are likewise
preferentially targeted for mutation in yeast, supporting the idea that small RNAs could modulate targeting preferences (Figure 3.2C-D). Enrichment analysis accentuates the difference in snoRNA binding between AID and cA3G, which most likely accounts for the differences observed in the overall ncRNA binding.

![Figure 3.2](image)

**Figure 3.2 AID and cA3G iCLIP identifies preferential binding to ncRNAs and tRNAs.**

(A) Yeast genome distribution of iCLIP reads, comparing mapping open reading frame (ORF), intergenic, intron and non-coding RNA (ncRNA) segments. (B) Enrichment adjusting by the relative size of each segment as described in text. (C) Distribution of iCLIP reads comparing ncRNA subtypes, ribosomal RNA (rRNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), tRNAs and other. (D) Enrichment adjusting by the relative size of each segment.

### 3.3 AID and cA3G bind different sites within tRNAs

Both AID and cA3G appear to have an equivalent affinity for tRNAs in yeast with 28% and 27% of all iCLIP reads mapping to tRNA genes respectively (Figure 3.2 C-D). However, taking advantage of the nucleotide resolution of iCLIP, the exact position of the crosslink
was analysed. The comparative analysis revealed that AID and cA3G preferentially bind at different positions within the tRNA sequence. When looking at the predicted secondary structure, it was found that the observed positions corresponded to different tRNA loops. While AID cross-link sites were predominantly located within the anticodon loop, cA3G binding sites are principally found within the T-loop (Figure 3.3 A-B).

Furthermore, cA3G showed a noticeable single nucleotide specificity, repeatedly binding to A57 within the T-loop. Conversely, AID showed a more relaxed preference for nucleotides within the anticodon loop. In either case, both deaminases showed a strict preference for single-stranded segments of tRNA within a stem-loop structure as cross-link sites, supporting the idea that binding specificity is related with structural features rather than sequence (Figure 3.4 A-B).
Chapter 3. Analysis of RNA binding preferences of AID and cAPOBEC3G in yeast by iCLIP

Figure 3.3 AID and cA3G crosslink sites within individual tRNA genes

AID (A) and cA3G (B) iCLIP tracks on representative tRNA genes taken from UCSC. Each peak represents a unique crosslinking site, downward peaks in orange denote gene in (−) strand. Nucleotide sequences are shown above the tracks. tRNA molecule distinctive folded structure with three hairpin loops that form the shape of a three-leafed clover is shown in the bottom right corner. Predicted secondary structures of each representative tRNA taken from GtRNAdb are shown with asterisks marking the crosslinked sites.

3.4 Discussion

RNA binding specificity of different AID/APOBEC deaminases is poorly understood and remains controversial, highlighting the need for further studies using more reliable techniques such as iCLIP. However, iCLIP technique is as robust as it is challenging, requiring careful optimization and rigorous analysis of data. This chapter showed how adjusting crosslinking intensity, lysis and RNaseI treatment the technique can be efficiently adapted for its use in yeast, facilitating data analysis due to the less complex transcriptome.

The optimized conditions allowed construction of iCLIP libraries for AID and cA3G. A smear corresponding to RNA-protein complexes was found for both deaminases, showing that in-vivo AID, as well as cA3G, efficiently bind RNA, contrasting previous in-vitro studies where cA3G appeared to be inert to RNA binding [101]. In this regard, dual domain APOCECs were for long thought to bind RNA through their non-catalytic N-terminal domain, however, more recent MS analysis of full length and native A3G cross-linked to nucleic acid revealed that RNA binds both N- and C-terminal domains, supporting the validity of the cA3G-RNA complexes recovered here by iCLIP [105,128,129]. The possibility exists of cA3G having low affinity for RNA, normally undetected under the stringent conditions used for in vitro pull-downs. The stabilization of the RNA-protein complex by induction of a covalent bond allows iCLIP method to detect low-affinity interactions even after stringent washes, pointing up the benefits of using this technique for no traditional RNA binding proteins such as AID/APOBEC deaminases.
Chapter 3. Analysis of RNA binding preferences of AID and cAPOBEC3G in yeast by iCLIP

Figure 3.4 Analysis of AID and cA3G crosslink sites along an average tRNA

Histogram showing the percentage of AID (A) and cA3G (B) iCLIP crosslink sites along an average tRNA transcript based on 15 top bound tRNAs. A coloured illustration of a tRNA secondary structure is displayed as an inset, and below the x axis in one-dimension marking regions on transcript corresponding to each loop.

Sequenced libraries were mapped to the yeast genome, revealing that both AID and cA3G preferentially bind ncRNAs and that this enrichment was mostly attributed to tRNAs. For AID, this finding is in striking accordance with the previously described preferential targeting of tRNA yeast promoters for mutation in a manner that suggests that binding to tRNAs itself might have a role in AID recruitment [101]. In the aforementioned study, Taylor et al., showed that both cA3G and AID mutate small RNA genes; however, AID had
a disproportionately higher preference for tRNAs that correlates well with the in vitro data showing that AID can bind structured RNAs, while cA3G showed no RNA binding ability. As discussed above, the iCLIP data presented here proved that cA3G could efficiently bind RNA and furthermore, it preferentially binds tRNAs to the same extent as AID. This suggests that the observed enhanced targeting of small RNA promoters by AID compared to cA3G might not be exclusively attributed to differences in RNA species binding preferences. Extra factors or circumstances not present in yeast might be needed for cA3G efficient targeting, as opposed to AID.

When looking at the exact position of the cross-linked sites in the tRNAs, it was found that, noticeably, AID and cA3G bind at different positions. While AID preferentially binds the anticodon loop, cA3G crosslinked sites were predominantly located within the T-loop. The reasons behind this differential preference for specific loops within the tRNA molecule might reside in AID and cA3G structural differences that may facilitate recognition of one loop or the other. To conclusively elucidate which residues could account for this differences, structural studies of AID and APOBEC crosslinked to tRNAs would be needed.

Remarkably both deaminases strictly bind single-stranded segments within stem-loop structures of tRNA. This type of structure has previously been shown to aid site-specific recognition by both AID and A3G. A recent study by Sharma et al., demonstrated that A3G preferred RNA substrates had predicted stem-loop structures [130]. On the other hand, AID has also been shown to edit C nucleotides located within transcription bubbles or stem-loops in ssDNA [131], making it plausible for AID to recognize similar structure for RNA binding. Most noticeably the tRNA adenosine deaminases (Tad/ADAT) are the group from which the AID/APOBEC family is thought to have originated, mostly due to structural and functional similarities [132,133]. The prokaryotic adenosine deaminase TadA mutates adenosine to inosine at the wobble position (A34) of the tRNA-Arg2 anticodon stem-loop [134], in remarkable accordance with the accumulated AID iCLIP crosslink sites within this loop. Furthermore, position 34 was one of the highly bound bases shown by AID iCLIP (Figure 3.4A). Therefore, binding to tRNAs could be a remaining trace of AID/APOBEC origin from tRNA editing enzymes. However, the benefit of this
retained ability to bind tRNAs and whether they could also serve as a substrate for deamination remains an intriguing area for further research.

Overall the data presented in this chapter revealed that at least in yeast, AID and cA3G selectively bind tRNAs and other structured small RNAs. Furthermore, it proved that iCLIP constitutes a powerful approach for AID/APOBEC-RNA interaction studies, detecting low-affinity binding and distinguishing even subtle differences such as preferential crosslinking to different sites within a define RNA structure. This encouraged the study to be extended to a broader set of deaminases and to a more physiologically relevant model.
Chapter 4

Screening the RNA binding of AID/APOBEC deaminases in HEK 293 cells

Results obtained in yeast were useful to define a level of RNA binding specificity for at least two deaminases in living cells and the power of iCLIP in picking up subtle differences. However, the question remained of whether the same preferences could be found in a more complex pool of RNAs, an environment closer to that of deaminases and when comparing an extended set of APOBECs.

As mentioned before, previous studies have addressed AID/APOBECs RNA binding activity. However, no consensus has been made regarding their binding specificity, nor has any research analysed in parallel RNA binding preferences of more than a few deaminases at a time. Therefore, this chapter describes the use of iCLIP followed by next-generation sequencing as a global approach to obtain catalogue and compare the RNAs bound to 5 different deaminases with diverse known functions in living cells, namely AID, A1, A3A, A3B, and A3G. As previously discussed, these deaminases are physiologically expressed in distinct cell types, however aiming to compare and identify any potential differences in RNA binding specificity, HEK293 cells were used in all cases, providing a common complex mammalian transcriptome to determine the RNA interactome of different deaminases by iCLIP. The focus of this chapter is the optimization of the technique and construction of iCLIP libraries for AID/APOBEC deaminases in mammalian cells.

4.1 Localization of overexpressed tagged deaminases in HEK293

As discussed before, subcellular localization of AID/APOBEC deaminases is a key regulatory mechanism, and it has been shown to differ between cell types. Furthermore, overexpression and the position of the tag can noticeably affect their compartmentalization
Chapter 4. Screening the RNA binding of AID/APOBEC deaminases in HEK293 cells by iCLIP

[78,91,135]. Therefore, the localization of the C-terminally FLAG-tagged deaminases used for the construction of iCLIP libraries in HEK293 was assessed by confocal microscopy.

![Figure 4.1 Intracellular localization of AID/APOBEC constructs in HEK293 cells](image)

Representative confocal images of HEK293 cells transiently transfected with plasmids encoding C-terminally FLAG tagged deaminases (GREEN). The nuclei were stained with (DAPI). Scale bar represents 10 µm.
APOBEC3G was the only deaminase to show an almost entirely confined cytoplasmic distribution. All other deaminases showed different degrees of nuclear-cytoplasmic localization. As expected, A3B was comparatively predominantly nuclear. The presence of A3B in the cytoplasm could be related with the previously described import defect in HEK293 cells [91]. Overexpressed AID and A3A showed a cell-wide localization with a slightly higher presence in the cytoplasm. Similarly, APOBEC1 was present in both compartments but was predominantly cytoplasmic. (Figure 4.1).

Overall, the distribution of the overexpressed proteins was similar within the cells and so it was fair to assume that they would mostly have access to the same pool of RNAs, with the exception of APOBEC3G. This results were later considered for interpretation and discussion of RNA binding preferences

4.2 iCLIP optimization for AID/APOBEC deaminases in HEK293

iCLIP protocol has been previously used in mammalian cell culture, successfully providing valuable information about RNA binding patterns. However, as discussed before, different proteins bind RNAs of varied sizes and with distinct affinity, making it critical to carefully optimize the technique for each deaminase separately. Furthermore, the technical complexity and multiple steps involved in iCLIP library construction increase the risk of losing material throughout the method. Therefore, to reduce systematic inefficiencies or contamination no more than two samples and control were processed at a time.

In all cases, HEK293 cells were transiently transfected with tagged deaminases. After 24hs living cells were crosslinked on ice with 150 mJ/cm2 at 254 nm to induce the formation of a covalent bond between the protein of interest and its bound RNAs. Cells were then recovered, lysed and RNA was partially digested to obtain the optimal RNA size distribution (as discussed in section 3.1.2). To determine the optimal RNase I concentration, dilutions ranging from 1:50 to 1:500 were tested for each deaminase. After immunoprecipitation and radioactive labelling, autoradiograms showed that all deaminases used in this study efficiently bind RNA \textit{in vivo} and that different RNaseI dilutions have distinct effects
depending on the deaminase, highlighting the relevance of individually optimising partial RNA digestion (Figure 4.2A).

**Figure 4.2** iCLIP optimization for AID/APOBEC deaminases in HEK293 cells.  
(A) Optimization of RNase I concentration. Autoradiographs show deaminase–RNA complexes that were treated with decreasing concentrations of RNase I before immunoprecipitation. The size of respective uncrosslinked deaminases are marked by arrows. Optimal RNase treatment and lines used for RNA extraction and library construction in each case are marked by asterisks. Samples without crosslinking (-CL AID) and EV serve as control.  
(B) Autoradiograms show duplicates for each deaminase using the optimised RNase I concentration. Asterisks mark the lines used for RNA extraction.  
(C) Example autoradiogram obtained during optimization showing nonspecific binding as unidentified band (marked by grey arrow).

Intermediate levels of RNase I (1/250) showed to be the most effective treatment in most cases, generating a diffuse and uniform signal of deaminase–RNA complexes starting upwards from each respective deaminase molecular weight up to 260 kDa and above. In the case of AID, A3A and A3G high RNase I concentrations gave sharp bands closer to their molecular weight corresponding to RNA molecules smaller than 60 nt suggesting extensive RNA digestion. Comparatively, the same RNase I dilution in A1 and A3B samples showed a strong signal up to 260 kDa, suggesting that higher RNase I dilutions would be needed for
an extended RNA digestion in those cases. Low RNase I condition significantly reduce the RNA-protein signal of A1, A3B and A3G most likely due to inefficient RNase I digestion leading to larger complexes that potentially obstruct gel migration or transfer to the nitrocellulose membrane. Conversely, low RNase I treatment of AID and A3A resulted in a relatively homogenous RNA-deaminase signal and was chosen as the most effective condition for A3A iCLIP. Taking these results into account for final iCLIP library preparation and respective replicates, 1/250 RNase I dilution was used for AID, A1, A3B, and A3G while 1/500 was the selected condition for A3A (Figure 4.2A-B).

Control experiments lacking protein-RNA complex signal such as the empty vector or non-crosslinked samples were crucial for initial optimisation. Figure 4.2C shows an example obtained during AID iCLIP optimization. The presence of a persistent unidentified sharp band across all samples including the EV, suggested potential contamination and non-specific binding that would affect final library preparation and data analysis. This was used to control the reagent stocks and the experimental steps, since as discussed before the complexity of the procedure was found not to be amenable to high throughput. Only experiments where negative controls showed no strong signal or sharp bands were pursued further.

4.3 AID/APOBEC deaminases library preparation

In order to collect a broad spectrum of RNAs, using the autoradiograms as a guide, a wide region of the membrane was used for RNA extraction and iCLIP cDNA libraries preparation. Considering that 70 nt of attached RNA shift the migration of protein–RNA complexes by approximately 20 kDa, to isolate RNAs within the optimal size range between 40 and 300 nt (including the adaptor), regions ~15–80 kDa above the expected molecular weight of each deaminase were cut. Bulk deaminases were digested with proteinase K, leaving a polypeptide at the crosslinked site where the reverse transcriptase stalls generating truncated cDNAs that will allow identification of the exact position where the crosslinking occurred. cDNAs were then selected by size using gel electrophoresis to eliminate unused RT primers (step 11 Figure 1.3). Three regions corresponding to 70–80 nt, 80–100 nt and 100–150 nt were separately extracted from the gel for each sample. This allows subsequent
Chapter 4. Screening the RNA binding of AID/APOBEC deaminases in HEK293 cells by iCLIP

Separate PCR amplification of the different size fractions for higher yield of shorter and longer RNAs, constituting an essential step in library preparation. However, technically wise it triplicates the number of subsamples to be processed in the further steps, being one of the reasons why it is not advisable to prepare more than two or three iCLIP libraries at a time.

Extracted cDNAs were then circularized and specifically linearized by the cleavable adaptor added during reverse transcription. This allowed PCR amplification of libraries, a step carefully optimized in each case to avoid over-amplification of secondary products. Final libraries were split into three cDNAs fractions of different lengths for each deaminase (High, medium and low) corresponding to the gel extraction step. Given the added combined length of the barcode adaptor and the P3/P5 Solexa primers used for PCR amplification and subsequent sequencing, a band cut at 70–80 nt on the cDNA gel is expected to generate 145–155 nt PCR products approximately. Using these guidelines, the number of PCR cycles was adjusted in each case, and libraries with the correct expected sizes were multiplexed, quantified and sequenced (Figure 4.3A-B).

Figure 4.3 iCLIP libraries from AID/APOBEC-expressing HEK293 cells. (A) Final cDNA libraries after optimized PCR amplification. Different sized bands ranging from 145-250 bp denoted as high (H), medium (M) and low (L) correspond to cDNA extracted from different gel regions in a previous step. Initial RNA extracted from membranes corresponding to autoradiograms in figure 4.2A. (B) Duplicate cDNA libraries corresponding to autoradiograms in figure 4.2B.
A total of 10 libraries corresponding to duplicates for AID, A1, A3A, A3B and A3G showing correct expected sizes and desirable cDNA quantification plus an EV library were sequenced on a HiSeq 4000. Resulting cDNA reads were mapped to the genome, and the random barcode was used to discriminate unique cDNA products from PCR duplicates. For example, two sequences that map to the same genomic location and that share an identical barcode sequence are treated as PCR duplicates, whereas they are identified as two unique cDNAs if they possess different random barcodes. Even though sequencing depth and overall cDNA count were sufficient among different AID/APOBEC libraries, collapsing of PCR duplicates significantly reduced the complexity of at least 6 out of 10 libraries. To ensure reliable cDNA complexity, a threshold was set and only libraries with more than 100,000 unique cDNAs after PCR duplicate removal were used for further analysis. This included individual libraries for A1, A3A, A3B, A3G and duplicate libraries for AID (AID 1 and AID 2) (Table 4.1). The EV control, performed in parallel did not show detectable PCR products. When submitted for sequencing, it generated less than 300 unique cDNAs mapping to the human genome. Given that the selected AID/APOBEC libraries generated 100-fold more cDNAs, it was estimated that over 99% of cDNAs from the iCLIP experiment represent RNA sites specifically crosslinked by the deaminases. Even though there were drastic differences in the size of some of the selected libraries, all further analysis was performed taking into account the total number of reads of each library, allowing comparison and giving distributions of reliable complexity in all cases.

<table>
<thead>
<tr>
<th>Protein</th>
<th>Replicate</th>
<th># of unique reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>AID</td>
<td>1</td>
<td>389312</td>
</tr>
<tr>
<td>AID</td>
<td>2</td>
<td>444551</td>
</tr>
<tr>
<td>APOBEC 1</td>
<td>1</td>
<td>2702610</td>
</tr>
<tr>
<td>APOBEC 1</td>
<td>2</td>
<td>23513</td>
</tr>
<tr>
<td>APOBEC3A</td>
<td>1</td>
<td>123737</td>
</tr>
<tr>
<td>APOBEC3A</td>
<td>2</td>
<td>2131</td>
</tr>
<tr>
<td>APOBEC3B</td>
<td>1</td>
<td>466992</td>
</tr>
<tr>
<td>APOBEC3B</td>
<td>2</td>
<td>31088</td>
</tr>
<tr>
<td>APOBEC3G</td>
<td>1</td>
<td>210194</td>
</tr>
<tr>
<td>APOBEC3G</td>
<td>2</td>
<td>6134</td>
</tr>
<tr>
<td>Empty vector</td>
<td>-</td>
<td>289</td>
</tr>
</tbody>
</table>

Table 4.1. Number of uniquely mapped reads after PCR duplicates removal. Libraries selected for further analysis are shown in red.
Chapter 5

Analysis of AID/APOBEC proteins distinct RNA binding specificities

To investigate and compare the RNA binding specificities of different AID/APOBEC deaminases, selected libraries were subjected to further analysis. The incredibly large amounts of data generated by ribonomic techniques such as iCLIP require considerable computational efforts for biological interpretation. Furthermore, most iCLIP analysis tools available do not allow comparison of more than two samples at a time with no defined consensus on how to compare multiple datasets of varied complexity, making the analysis process particularly challenging. Therefore, with the objective of determining potential similarities and differences in RNA binding that might exist among the five different deaminases included in my thesis, this chapter describes the use of a global approach for unbiased comparison of multiple iCLIP experiments. This first level of analysis allowed the identification of distinctive RNA binding patterns among different deaminases to then study in more detailed the most outstanding RNA specificities.

5.1 AID/APOBEC deaminases have distinct preferences across different RNA types

The random barcode introduced into iCLIP cDNAs allows identification of the deaminase-bound RNAs across the human genome quantitatively. Based on the iCLIP library preparation strategy, the crosslink site was mapped to the first nucleotide preceding the start of each unique cDNA. As a general first approach, the distribution of AID/APOBEC significant crosslink events across the human genome was analysed into several categories, corresponding to functional regions of the transcript such as the 3'UTR, 5'UTR, ORF (exons), as well as regions corresponding to intergenic regions, introns, and ncRNAs (including tRNAs separately added as described in section 2.7.2) and telomeric regions. The results were visualised as the proportion of iCLIP reads mapping to each region and revealed some shared
features as well as dramatic differences between the different AID/APOBEC proteins (Figure 5.1A).

Most deaminases showed high binding to intergenic and intronic regions. Notably, almost 50% of A3B total reads mapped to introns, consistent with its predominantly nuclear localization and thus enhanced encounter with unspliced pre-mRNAs. In contrast, A3A, A3G, and A1 showed significant binding to ORF. Perhaps even more noticeably almost 30% of A1 reads individually mapped 3’UTRs. These results suggest that in addition to the availability of the RNA pool related to the subcellular abundance of the deaminases, additional constrains affect the binding preferences of the overexpressed proteins.

![Graph A](image1)

![Graph B](image2)

**Figure 5.1 Genomic distribution of AID/APOBEC iCLIP cDNAs in HEK293 cells.**

(A) Percentage of cDNAs (Reads) from each deaminase iCLIP library that mapped to different type of RNAs and defined regions in HEK cells. (B) Enrichment analysis calculated dividing the percentage of deaminase-bound RNA by the percentage of the nucleotides of each category in the genome. In all cases AID values represent an average of AID1 and AID2 replicates.
Furthermore, enrichment analysis, in this case correcting for the relative size of each RNA category compared in the analysis, revealed that A1 cDNA density in 3'UTR was 3-fold higher than in the coding sequence (ORF) (Figure 5.1B). This is consistent with and potentially explains the previously described enhanced A1 transcript editing at this specific region [23].

In addition, size correction revealed a pronounced preference of AID for ncRNAs. A3G and A3A libraries were likewise relatively enriched for ncRNAs, with their highest cDNA densities falling in this category (Figure 5.1B). Conversely, A1 showed no apparent enrichment for ncRNAs. These results reinforce the notion that the distribution of RNAs crosslinked by different deaminases are not solely related to the abundance or size of the RNAs, but also to functional features of the RNA and a distinctive binding specificity.

A detailed analysis of the distribution of mapped reads among ncRNA subtypes revealed an even more skewed distribution for some of the deaminases. tRNAs and miRNAs were overwhelmingly the highest categories of ncRNA crosslinked by AID, followed by snRNAs. In contrast, A3A and A3G showed no apparent preference for tRNAs while strongly binding miscRNAs (Y RNAs, 7SL RNAs, 7SK RNAs) (Figure 5.2A).

This biased distribution is even more pronounced when the data is analysed using the enrichment analysis. As highlighted in Figure 5.2B, AID binding to tRNAs was significantly enhanced compared to other deaminases. Interestingly none of the deaminases appeared to preferentially associate with rRNAs, despite the fact that these are the most abundant RNA species in any cell. These results corroborated the earlier suggestions based on the general distribution of hits that the enrichment observed in the different iCLIP libraries are likely to reflect some specificity due to the deaminase RNA binding preferences rather than being a mere reflection of the different RNA species abundance and subcellular distribution.

5.2 Global analysis of AID/APOBEC deaminases differentially bound RNAs

As seen in the results described in the previous section, the distribution of iCLIP reads across the genome revealed that different deaminases appear to preferentially bind distinct RNA types with some degree of specificity. The exception being A3G and A3A which showed similar binding patterns.
Figure 5.2 Distribution of AID/APOBEC iCLIP cDNAs across ncRNA subtypes in HEK293 cells. (A) Percentage of total cDNAs (Reads) from each deaminase iCLIP library that mapped to main ncRNA subcategories in HEK cells: transfer RNAs (tRNAs), microRNAs (miRNAs), miscellaneous RNAs (miscRNAs), small nuclear RNAs (snRNAs), long intergenic noncoding RNAs (lincRNAs) and ribosomal RNAs (rRNAs). (B) Enrichment analysis dividing the percentage of deaminase-bound RNA by the percentage of the nucleotides of each subcategory in the genome. In all cases AID values represent an average of AID1 and AID2 replicates.

To analyse in more detail the different RNA binding specificities among deaminases, all iCLIP reads mapped to the genome with assigned annotated genes were compared. Taking into account the normalized number of hits (crosslinks events, cDNA counts) the 100 most differentially bound genes across samples were used to build a heatmap comparing deaminases distinctive binding profiles (Figure 5.3). It is worth noting that most deaminases were detected bound to their own RNA, an indication of their transient overexpression and ability to bind foreign nucleic acids that enter the cell.

As expected, AID duplicates were clustered together and differ the most from all other deaminases. In accordance with the analyses discussed in section 5.1, this is mostly due to AID preferential binding to tRNAs and miRNAs, scarcely bound by the rest of deaminases.
Chapter 5. Analysis of AID/APOBEC proteins distinct RNA binding specificities
In accordance with the data shown in section 5.1, the samples corresponding to the A3A and A3G libraries were paired together showing similar binding profiles. When looking at specific genes comparatively highly bound by both deaminases, it was found that the crosslinked sites corresponded mostly to Y RNAs within introns of annotated genes. As previously mentioned small structured Y RNAs are categorized as miscRNAs, which were shown to be highly enriched in A3G and A3A iCLIP libraries in the previous section. Figure 5.4 shows A3A and A3G iCLIP tracks on KCNK4 and KIF4A genes as an example of binding to Y RNAs. Analysis of a larger segment of KCNK4 revealed how crosslink sites restrictively coincide with hosted Y RNA. Zoom in allowed a more detailed visualization of individual peaks showing that, even though some slight differences were noted, both deaminases crosslink sites localized around the same position. To corroborate whether these small variations are significant a detailed dissection of more Y RNAs and library replicates would be needed. In addition, A3G and A3A also distinctively bound histone cluster mRNAs (e.g., HIST1H4B, HIST1H2AH) which in some cases have been shown to have stem-loop secondary structures but are not noticeably crosslinked by other deaminases [136].

Notably, A3B and A1 were paired together mostly due to low binding to the selected genes rather than shared high binding to any particular type. This could be the result of extreme bias in gene selection due to AID outstanding selectivity for miRNAs and tRNAs, constituting a potential caveat of the global binding distribution analysis. Therefore, in order to better identify RNA species distinctively crosslinked by other deaminases that could have been masked by AID preferential binding, a new analysis was made using 20 comparatively highly bound genes as markers for each deaminase when compared to the others.

A new heat map was built to graphically show the selected markers (Figure 5.5). In this case, AID stands for an average of AID1 and AID2. AID markers were mostly represented by miRNAs with only one tRNA present, showing that miRNAs are more selectively bound by AID.
As discussed above (Figures 5.2 and 5.4), the specific location of the crosslinks in markers associated with A3G and A3A, was mostly within miscRNAs. However, in this case, while A3A was mostly bound to Y RNAs, A3G crosslink sites corresponded to 7SL RNAs (the RNA component of the signal recognition particle, SRP) within marker genes (Figure 5.6). Taking this into account the percentage of reads mapping to different miscRNAs was compared for A3A and A3G. Some differences were found, even though both deaminases bind all three kinds of miscRNAs, A3G showed a level of preference for 7SL RNAs while A3A preferentially bound Y RNAs (Figure 5.6 C). However, in order to confirm this trend, additional independent replicates would be required.

Figure 5.4 Single gene analysis of A3A and A3G iCLIP shows binding to Y RNAs
(A) A3A and A3G iCLIP tracks on KCNK4 gene were taken from UCSC. A zoom in is visualized in the red rectangle where each peak represents a crosslink site. Y RNAs are shown in green. (B) Zoom in of A3A and A3G iCLIP tracks on KIF4A gene showing host Y RNA.
Chapter 5. Analysis of AID/APOBEC proteins distinct RNA binding specificities

Figure 5.5. AID/APOBEC profiling by 20 comparatively highly bound markers for each deaminase. Heatmap showing 20 markers for each deaminase identified as genes highly bound in one sample when compared to the others. AID represents an average of AID1 and AID2 values. Colour key showed on the top left corner.
Figure 5.6 A3A and A3G binding to miscRNAs subtypes

(A) A3A and A3G iCLIP tracks on fragments of selected marker genes DENND5B and CYTH3 were taken from UCSC showing A3A preferential binding to Y RNAs (annotated in GREEN). (B) A3A and A3G iCLIP tracks on fragments of selected marker genes GRK45 and PI4KB were taken from UCSC showing A3G preferential binding to 7SL RNAs (annotated in RED). In all cases peaks represent crosslink sites, downward orange peaks denote genes in (-) strand. cDNA counts on the left of each track represent reads after collapsing of barcodes resulting in differences depending on efficiency of library production. (C) Percentage of total A3A and A3G iCLIP reads bound to different miscRNA subtypes (Y RNAs, 7SL RNAs and 7SKRNAs)
When looking into A1 crosslinks associated with specific marker genes, it was found that they were preferentially localized within the 3’UTR, in accordance to what was previously found for the distribution of iCLIP reads (Figure 5.1). Figure 5.7 shows as an example A1 binding to two selected genes 3’UTR. A more detailed dissection of crosslink sites revealed that binding mainly occurs around AU rich segments.

![Figure 5.7 A1 binding to AU rich 3’UTR.](image)

A1 iCLIP tracks on selected marker genes KDELR2 and CIRBP were taken from UCSC where each peak represents a crosslink site. Nucleotide composition of the binding sites in red rectangle is visualised by zooming in.

Finally, in the case of A3B, crosslink sites within randomly selected marker genes were in the intronic region. Notably, while in one of the analyzed genes crosslink events were scattered along the intron, in the other case binding was restricted to a particular site. A more detailed inspection revealed that the position corresponded to the edge of a simple repeat (T) segment.
Figure 5.8. A3B binding to intronic region

(A) A3B iCLIP tracks on randomly selected marker gene CBFA2T2 was taken from UCSC where peaks show crosslink sites scattered along gene intronic region. (B) A3B iCLIP tracks on randomly selected marker gene KCNC2 was taken from UCSC. Crosslink sites were restricted to a specific region within a large intron, detailed analysis in red rectangle revealed that it corresponded to the end of a repetitive segment.

Overall profiling analysis showed the different RNA binding preferences among deaminases, particularly and further highlighting AID specificity and preferential binding to tRNAs and miRNAs. Taking this into account together with the availability of biological replicate samples, AID crosslinking to tRNA and miRNAs was analyzed in more detail.

5.3 AID specifically binds tRNAs in HEK293 cells

Remarkably, in accordance with the yeast iCLIP data, tRNAs were highly recovered by AID iCLIP in HEK cells. Despite being highly abundant, tRNAs are not normally found as background reads in CLIP experiments. Furthermore, AID most highly bound tRNAs did not coincide with the relative abundance of endogenous tRNA distribution (Table 5.1), indicating that tRNAs were specifically bound by AID instead of being randomly drawn from the total tRNA pool.
Table 5.1 Ten most abundant tRNAs in total small RNA sequencing against most abundant tRNAs in AID iCLIP. Small RNA sequencing data from HEK293 cells taken from [137]. Most abundant tRNAs in AID iCLIP represent the cDNA counts sum of both AID1 and AID2.

Analysis of the exact binding position across all AID bound tRNAs revealed that crosslinked sites were predominantly located within transcript regions corresponding to the anticodon stem loop and the T-stem-loop (Figure 5.8). Interestingly, the highest number of crosslinks was found in the canonical position A57 within the T-stem loop. However, the sum of crosslinks along each region showed that AID preferentially binds the anticodon loop in accordance to what was found in yeast AID iCLIP (chapter 3).

Figure 5.8 Histogram of AID iCLIP reads along an average tRNA transcript based on all bound tRNAs. Coloured illustrations of a tRNA secondary and tertiary structure are displayed on the right. Below the x axis in one-dimension regions on transcript corresponding to each loop are marked. Numbers on top of grey horizontal bars represent the sum of crosslinks for each loop.
This was further confirmed by individually looking at the crosslink sites within single tRNA species (Figure 5.9). Nonetheless, AID binding within the T-stem loop was confined to position 57, while crosslinks located at the anticodon loop alternated mostly between positions 34-37; accounting for the difference observed in Figure 5.8. In addition, analysis of individual tRNAs showed that AID exclusively bound either the anticodon loop or less frequently the T-stem loop in a determined tRNA specie, but not to both loops within the same tRNA. Furthermore, the high number of crosslink sites mapping to a single nucleotide within a tRNA together with the reproducibility between replicates, indicates AID binding specificity for tRNAs.

5.4 AID preferentially binds miRNAs in HEK293 living cells

The more detailed analysis of ncRNAs subtypes bound by different deaminases revealed that together with tRNAs, miRNAs were overwhelmingly the highest category of ncRNA crosslinked to AID while being scarcely bound by other APOBECs (Figure 5.2). The human genome contains at least 1917 annotated miRNA genes as opposed to 625 tRNA loci, therefore, despite having similar average transcript lengths, miRNA subcategory overall percentage of nucleotides in the genome is higher, making them appear less enriched when compared to tRNAs (Figure 5.2 B). However, considering the level of expression, miRNAs represent between 0.003 to 0.02 percent of total RNA mass in contrast to the 10 to 15 percent represented by tRNAs (taken from [138]). Thus correcting by mas, miRNAs become one of the most overrepresented interactions in the AID libraries, highly enriched compared to other RNA segment types and overall RNA interactome as seen by the profile analysis (Figures 5.3 and 5.5).
Chapter 5. Analysis of AID/APOBEC proteins distinct RNA binding specificities

(A)
Figure 5.9 AID crosslink sites within individual tRNA genes

(A) AID iCLIP tracks on representative tRNA genes taken from UCSC. Each peak represents a unique crosslinking site, downward orange peaks denote genes in (-) strand. cDNA counts on the left of each track represent reads after collapsing of barcodes resulting in differences depending on efficiency of library production. Nucleotide sequences are shown above the tracks. (B) Predicted secondary structures for each
Mature miRNAs are crucial regulators of gene expression at the post-transcriptional level in eukaryotes via targeting mRNA 3'-UTR region. They are transcribed as long structured primary miRNAs (pri-miRNA) that are processed by a microprocessor complex in the nucleus formed by DGCR8 and Drosha [139]. The processing product, termed pre-miRNA, is a characteristic single hairpin that is transported to the cytoplasm and further cleaved by Dicer into duplex miRNA strands. Usually, only one of the two resulting strands accumulates as a functional mature miRNA (~22 nt), while the other strand is degraded [140] (Figure 5.10).

Figure 5.10. miRNA maturation
Primary miRNA (pri-miRNA) transcripts, are excised by RBPs DROSHA and DGCR8. The pre-miRNA hairpin is then exported from the nucleus and the loop region of the hairpin is bound and cleaved by the RBP DICER to produce the mature 3p 5p miRNA duplex.

Nucleotide resolution analysis of AID binding sites within miRNAs revealed that crosslink sites were overwhelmingly preferentially located at position -1 of predicted mature miRNAs (Figure 5.10), suggesting that AID binds to structured pri and pre-miRNAs. Noticeably, AID binding was associated exclusively to one strand in the miRNA duplex, in most cases.
corresponding to the accumulated functional strand. Looking at highly bound miRNAs predicted secondary structure it was found that this preferred AID crosslink site was in close proximity to a loop or bulge rather than within the double strand miRNA stem (Figure 4.13). Furthermore, the single nucleotide specifically bound by AID was not always of the same kind, further suggesting that selectivity is based on structural features rather than sequence.

Figure 5.1. AID crosslink sites within miRNAs.

Average histogram of AID iCLIP reads along a normalized miRNA transcript based on top ten bound miRNAs. Mature miRNA sequence within pri-miRNA transcript are marked in pink. Graph below the track shows AID crosslink sites within a predicted pri-miRNA secondary structure.
Chapter 5. Analysis of AID/APOBEC proteins distinct RNA binding specificities
Figure 5.12. AID crosslink sites within individual miRNA genes.
AID1 and AID2 iCLIP tracks on representative miRNA genes taken from UCSC. Each peak represents a unique crosslinking site. Nucleotide sequences are shown above the tracks. Black horizontal bars mark the predicted mature miRNAs within the transcript. Predicted secondary structures of each representative miRNA taken from miRBase are shown with asterisks marking the crosslinked sites. Nucleotides corresponding to mature miRNAs are shown in pink.
Chapter 6

Discussion

iCLIP proved to be a powerful method to analyze RNA binding specificities with high resolution. However, it is technically demanding requiring detailed, time-consuming optimization for each specific protein of interest. The number of steps and purifications involved increase the risk of losing material along the way. Furthermore, there are a few steps where the quality and efficiency can be assessed. As it has been shown, even final libraries in the correct size range and efficiently sequenced can result in low complexity libraries due to overrepresentation of PCR duplicates. To improve this, it would be necessary to further optimize the number of PCR cycles for the final library amplification, keeping it as low as possible. Even with these difficulties, a total of 2 iCLIP libraries for yeast and 6 iCLIP libraries for five different AID/APOBEC deaminases in HEK293 cells were successfully obtained and analyzed during my thesis work.

Comparative analysis of the distribution of iCLIP reads across the genome revealed some evident differences in the RNA binding pattern of different deaminases. Intergenic and intronic regions were highly bound by most deaminases. However, these are considered common background RNA sequences in most CLIP studies [141], particularly in the case of atypical RNA binding proteins such as deaminases with non-well-define RNA binding domains. Therefore, to better visualize significant binding differences a correction by the size of each region was made, highlighting an overwhelming binding of A1 to 3’UTRs that was then also shown by the profiling analysis. When looking at the position of A1 crosslink sites within 3’ UTR of marker genes, it was found that binding mostly occurred at AU rich segments. This is in striking accordance to previous studies showing that A1 targets various AU-rich segments of transcripts 3’UTR for editing [142]. Therefore, a more detailed computational analysis of the A1 iCLIP data obtained in my thesis could be of great use to better define or confirm binding motifs within 3’UTRs and to identify new potentially relevant editing targets. In this regard A1 iCLIP raw reads could be re-analyzed looking for
mutations, giving the data presented here extra potential value. Whether bound RNAs are targeted for mutation or not the functional consequences of A1 binding to 3’UTRs remain to be determined. Previous studies have associated A1 3’UTR binding with increased mRNA stability [24,143], therefore analyzing whether this is the case for iCLIP bound RNAs could aid on defining a more general functional significance for this interaction.

In the case of AID, enrichment analysis revealed preferential binding to ncRNAs. Even though not as pronouncedly, A3A and A3G iCLIP libraries were also mostly enriched by this category of RNAs. Looking at the distribution among different ncRNA subtypes it was found that tRNAs, miRNAs, and snRNAs were highly recovered by AID iCLIP while being scarcely bound by any other APOBEC. Conversely, A3G and A3A shared preferential binding to miscRNAs, which were not noticeably bound by AID. In contrast to yeast iCLIP results, A3G showed no significant binding to tRNAs in HEK293 cells; this could be due to the presence of other preferred substrates not present in yeast transcriptome such as miscRNAs or merely be due to its observed restricted cytoplasmic localization, favoring encounter with different RNA types. In both cases, it shows the importance of studying RNA-protein interaction in a context as close as possible to those physiologically relevant. It also highlights the fact that even in the presence of a complex transcriptome, AID still preferentially binds tRNAs as seen in yeast iCLIP. As previously discussed, this binding specificity could be an evolutionary remnant of its proposed origin from adenosine deaminase acting on tRNA (TadA/ADAT) enzyme. Supporting this idea, it was found that AID binds the anticodon loop of tRNAs, crosslinking between positions 34-37, in close coincidence to TadA targeting of A34 at the wobble position. The reason or biological benefit or functional significance for AID retaining a specificity for tRNAs remains unclear. It could be speculated that small RNA binding could sequester AID to subnuclear localities such as the nucleoli. tRNAs have been shown to be localized in nucleolar area during transcription in yeast and the same has been suggested for vertebrate cells [144,145]. Furthermore, nucleolar localisation of overexpressed AID has already been reported in mammalian cells, favouring the idea that it could be linked with its interaction with tRNAs [146]. Alternatively, preferential recognition of particular RNA structures such as folded tRNAs could determine the recruitment of AID to genomic regions. As previously discussed, a study by Taylor et al. suggested that in yeast, AID binding to tRNAs and other small
structured RNAs could be involved in its disproportional preferential targeting of tRNA promoters for mutation. The novel finding that AID binds tRNAs, as well as other small structured RNAs in living cells, prompts to look at whether binding to particular RNA structures could be involved in AID targeting regulation in mammalian cells. In this regard, a recent model for AID recruitment proposes that binding to G-quadruplex (G4) structures in transcribed G-repeats that are enriched in switch region would guide AID to substrate genes for CSR [115]. This leads to the speculation that AID could potentially associate with other cellular RNAs that fold into similar structures and thus be involved in mis-targeting of AID to other genomic loci. Application of G-quadruplex prediction tools on AID iCLIP data obtained in my thesis could help unravel the validity of this proposed recruitment mechanism.

Another striking novel discovery revealed by AID HEK iCLIP, was a selective pronounced binding to miRNAs. Correction by mass and global profiling analysis among deaminases heightened this preferential binding. Furthermore, the nucleotide resolution of iCLIP revealed that AID binds miRNAs with noticeable specificity consistently crosslinking at position -1 of the predicted mature miRNA. This indicated that AID must bind unprocessed miRNAs (pri or pre-miRNAs) suggesting a potential role for AID in miRNA maturation. The particular crosslink site makes it plausible for AID to block miRNA biogenesis by impeding Dicer or microprocessor access to its binding site. However, at least in B cells, it has been shown that AID deficiency does not affect mature miRNA abundance [147]. Nevertheless, given that the expression of these small RNAs is highly variable and that processing efficiency depends on the cellular context or external signals, the possibility for AID to be involved in miRNA maturation should not be discarded and would be an interesting area for further research [148]. Noticeably, AID was found to exclusively bind one strand within the miRNA duplex, which was in most cases found to correspond to the functional mature miRNA. Furthermore, taking into account the exact crosslink position of AID, the strand bound determines either exclusive binding to pri-miRNAs or to both pre and pri-miRNAs. This raises interest in analyzing potential determinants of this selectivity. One possibility could be particular traits, structures or localization of different miRNA species, therefore a more detailed analysis and classification of all bound miRNAs in both replicates could help unravel AID particular binding pattern. In addition, to further elucidate the physiological
significance of this novel interaction it would be particularly interesting to analyze whether AID binds microRNAs in more relevant cell types.

Noticeably, in both cases analyzed for AID, binding is closely associated with RNA structure rather than sequence, selectively binding loops within tRNAs and structured unprocessed miRNAs. This was also the case for both A3A and A3G, preferentially binding structured Y RNAs and 7SL RNAs, however more detailed analysis is needed to determine the exact crosslink site and correlation with secondary structure in those cases with the limitation of less available predicted structures data base. For A3G, binding to miscRNAs has previously been shown to be associated with incorporation into viral particles, supporting the validity of the binding specificities presented here [111]. The role of binding to miscRNAs, remains to be defined. In the case of A3G and HIV, a role for these RNAs in mediating the precise targeting of cytidine deaminases to the viral reverse transcription complex by mimicking nucleocapsid RNA binding affinity has been suggested [113]. The fact that A3A was also found bound to these types of small structured RNAs opens the question of whether A3A uses similar RNA dependent mechanism for viral restriction. Furthermore, given that both A3A and A3G have recently been shown to edit RNA, further analysis explicitly looking at their binding patterns within mRNAs could reveal potential differences that might have been missed by our analysis focusing on main binding patterns in ncRNAs [56,149]. It has been previously proposed that both AID and A3G bind stem loop structures within their target mRNAs, however their targeted genes do not significantly overlap. Further and more detailed analysis of the iCLIP data presented here could aid on the understanding of the different selected structures including the apparent differential binding of 7SL RNA and Y RNAs.

In the case of A3B, no evident RNA binding specificity was found other than high crosslinking to intronic regions and potentially to repetitive sequences. To confirm this, a more detailed analysis and replicates focusing on A3B data would be needed. It is however possible to speculate that binding to intronic highly repetitive segments could prevent A3B undesired mutagenic activity in coding regions of the host genome. The nuclear localization of this mutagenic protein and its known association with cancer development, makes the study of any potential mechanism for its regulation particularly relevant emphasizing the
potential of further analysing A3B iCLIP data presented here. However, in most cases further analysis would require extensive computational efforts and development of novel tools specific for iCLIP data. Some of the limitations include the lack of reliable RNA structure prediction tools or structure databases and difficulty to map repetitive sequences accurately among others.

Overall, my thesis work outlines for the first time a comprehensive analysis of the RNA binding patterns of an extended set of deaminases in vivo; clearly showing that different AID/APOBEC proteins have distinct RNA specificities, emphasising the possibility for this interaction to influence their diverse functions as well as potential new roles. I also describe a novel interaction of AID with miRNAs and tRNAs with nucleotide resolution and in association with predicted secondary structure. These findings open a new perspective where RNA binding could not only regulate the localization and activity of deaminases, but AID/APOBEC proteins could alternatively affect the abundance or accessibility of certain bound RNAs. Furthermore, the extensive dataset obtained is of a great reference value for the field of AID/APOBECs with detailed information about RNA interactome of these deaminases. This opens the field for much further analysis and investigation of the specific meaning and validation of each preferential binding, providing new insights into understanding the role of AID/APOBEC deaminases interaction with RNA.
References


24. Anant S, Davidson NO. An AU-rich sequence element (UUUN[A/U]U) downstream of the edited C in apolipoprotein B mRNA is a high-affinity binding site for Apobec-


References


82. Sowden MP, Yang Y, Smith HC. Intracellular trafficking determinants in APOBEC-1, the catalytic subunit for cytidine to uridine editing of apolipoprotein B mRNA. Exp Cell Res. 2001 Jul 15;267(2):153–64.


References


References


References


Appendix

Detailed iCLIP protocol

Step 1-4: *in-vivo* UV crosslinking, cell lysis, partial RNAse digestion and immunoprecipitation

**HEK 293 cells**

293 cells transiently expressing FLAG tagged deaminases were grown in a 10 cm plate for 24hs, medium was removed and cells were covered with 6 ml of ice-cold PBS buffer and subjected to UV-C irradiation on ice (150 mJ/cm², Stratalinker 2400). Cells were scraped off and transferred into 15 ml falcon tubes, precipitated by centrifugation for 1 min at 14,000 rpm shock-frozen on dry ice and kept at -80°C until used. Pellets were then resuspended in 1 ml lysis buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1 mM MgCl₂, 0.1 mM CaCl₂, 1 % NP-40, 0.1 % SDS, 0.5 % Na-Deoxycholate + 1/100 volume of Protease Inhibitor Cocktail Set III added on the day) transferred to 1.5 ml microtubes and sonicated with a bioruptor plus for five cycles alternating 30 s on/30 s off at low intensity. For partial RNase digestion, RNase I (Life Technologies, AM2295) was diluted 1:50, 1:250 or 1:500 in PBS for high, medium and low RNase treatment, respectively (One sample per condition). 10 μl RNase I dilution and 3 μl Turbo DNase (Life Technologies, AM2238) were added to the cross-linked lysate and incubated for exactly 3 min at 37°C and 1100 rpm. Cells were precipitated by two rounds of centrifugation at 4°C and 14,000 rpm for 10 min followed by careful collection of the supernatant. Separately, 100 μl of Anti-FLAG M2 Magnetic Beads (Sigma, M8823) were washed 2× with 900 μl lysis buffer (without protease inhibitor), resuspended in 100 μl of fresh lysis buffer and added to the cross-linked supernatant. The mix was incubated for 1 h or overnight rotating at 4°C.

**Yeast Cells**

Yeast transformants for HA-tagged deaminases were grown in 5ml of YC (-) Arg (-) URA (+) GLUCOSE medium shaking at 300 rpm at 30 °C overnight. To scale up and induce vector
expression this starter culture was seeded into 250 ml of YC (-) Arg (-) URA (+) GALACTOSE medium and grown overnight. Yeast cells were then pelleted at 4 °C, resuspended in 50 ml of ice cold PBS, placed on 10cm plates on ice and crosslinked 2 x 5min at 254 nm in a Stratalinker 2400 (≈ 2.4 J/cm2), with a 1 min pause used to gently shake plates. Cells were then collected, precipitated at 2000 rpm for 5min, washed with 50ml PBS, snap-frozen on dry ice and stored at −80 °C until further use. Pellets were then resuspended in 2 ml of lysis buffer with 1/100 volume of Protease Inhibitor and 1/1000 volume of ANTI-RNase (Life Technologies, AM2692) and distributed in 2 screw cap microtubes on ice. 1 ml of glass beads was added to each tube and cells were lysed by vortexing for 10 min at 4°C. Lysates were recovered from glass beads by piercing a hole in the bottom of the tubes using a 19G needle, placing them in 15ml collection tubes and centrifuging 1 min at 1000 rpm. For partial RNase digestion, RNase I (Life Technologies, AM2295) was diluted 1:50, 1:250 or 1:500 in PBS for high, medium and low RNase treatment, respectively (one sample per condition, in this case 2 tubes). 10 µl RNase I dilution and 3 µl Turbo DNase (Life Technologies, AM2238) were added to the cross-linked lysate and incubated for exactly 3 min at 37°C and 1100 rpm. Cells were precipitated by two rounds of centrifugation at 4°C and 22,000 rpm for 10 min followed by careful collection of the supernatant. Supernatants corresponding to the same original sample were pooled back together in a microtube. For immunoprecipitation 2ul of anti-HA antibody was added and tubes were incubated rotating overnight at 4°C. The next day 100 µl of pre-washed Protein G Dynabeads (Life Technologies, 10004D) were added and the mix was incubated rotating for 2hs at 4°C.

Following steps were the same for both yeast and HEK293 cells.

**Step 5: Dephosphorylation**

Beads with crosslinked RNA-deaminase complexes were then washed 2× with high-salt wash buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 0,1 % SDS, 0.5 % Na-Deoxycholate, 1 % NP-40) and 1× with PNK wash buffer (20 mM Tris-HCl pH 7.4, 10 mM MgCl2, 0,2 % Tween-20) using a magnetic stand on ice and rotating the second wash for at least 1 min at 4 °C. For 3’end dephosphorylation beads were resuspended in 20 µl of the following mix: 4µl PNK pH 6.5 buffer (350 mM Tris-HCL pH 6.5, 50 mM MgCl2, 5 mM DTT),
0.5 μl PNK (NEB, M0201L), 0.5 μl RNAsin (Promega, N2515), 15 μl H2O, and incubated for 20 min at 37°C in a thermomixer at 1100 rpm. Samples were washed 1× with 900 μl PNK buffer, 1× high-salt wash buffer rotating 1 min at 4°C and 2× with 900 μl PNK buffer.

**Step 6: L3 linker ligation**

Supernatant was then discarded and beads were resuspended in 20 μl of the following mix: 5 μl 4x ligation buffer (200 mM Tris-HCl pH 7.8, 40 mM MgCl2, 4 mM DTT), 1 μl RNA ligase (NEB, M0204L), 0.5 μl RNAsin, 1.5 μl adapter L3 (20 μM) (rAppAGATCGGAAGAGCGGTTCAG/ddC, IDT), 4 μl PEG400; and incubated overnight at 16°C in a thermomixer at 1100 rpm. Beads were then washed 1× with 500 μl PNK buffer, 2× with high-salt wash buffer rotating second wash 1 min at 4°C, wash 2× more with PNK buffer.

**Step 7: 5’ end radioactive labelling of RNA**

Beads were resuspended in 1 ml of PNK buffer and a 200 μl aliquot was taken. Supernatant was removed and beads were resuspended in 4 μl of the following mix: 0.2 μl PNK, 0.4 μl ATP [γ-32P] (Perkin Elmer, NEG502A250UC), 0.4 μl 10X PNK buffer (NEB, M0201L), 3 μl H2O and incubated for 5 min at 37°C in a thermomixer at 1100 rpm. Supernatant was then discarded as radioactive waste and beads were resuspended in 20 μl of 1X NuPAGE loading buffer mixed with the remaining cold beads and incubated for 5 min at 70°C.

**Step 8: SDS–PAGE and nitrocellulose transfer**

Samples were placed on a magnetic stand to collect the eluate that was then run on 9-well or 10-well Novex NuPAGE 4-12% Bis-Tris gels (Invitrogen) with 1× MOPS running buffer (Invitrogen). After gel electrophoresis, protein and covalently bound RNAs were transferred to a nitrocellulose membrane (Whatman) using a Novex wet transfer apparatus (Invitrogen). The nitrocellulose membrane was rinsed with 1× PBS, wrapped into cling film and exposed to a BioMax XAR Film (Kodak) at –80°C overnight.
**Step 9: RNA isolation**

Using the film as a guide, fragments of the membrane were cut and cross-linked RNA was extracted by further trimming the membrane in small pieces and incubating them with 2 mg/ml proteinase K (Roche) in 200 μl PK buffer (100 mM Tris-HCl pH 7.5, 50 mM NaCl, 10 mM EDTA) for 20 min at 37°C. Incubation was repeated after addition of 200 μl of PK buffer/7M urea (100 mM Tris-HCl pH 7.5, 50 mM NaCl, 10 mM EDTA, 7 M urea). Supernatants were then mixed with 400 μl phenol/chloroform (Sigma, P3803) added to 2 ml Phase Lock Gel Heavy tubes (VWR, 713-2536) and incubated 5 min at 30°C shaking at 1100 rpm. Phases were separated by spinning for 5 min at full speed at room temperature. The aqueous layer was then carefully transferred to a new microtube, mixed with 0.75 μl glycoblue (Ambion, 9510), 40 μl 3M sodium acetate pH 5.5 (Life Technologies, AM9740), 1 ml 100% ethanol and incubated overnight at -20°C for RNA precipitation. After centrifugation at 15,000 rpm at 4°C for 20 min, the supernatant was removed and the pellet was washed with 0.9 ml 80% ethanol and spin again for 5 min. The pellet was then resuspended in 5 μl H2O and transferred to a PCR tube.

**Step 10: Reverse transcription**

After addition of 1 μl of RT primer (according to table 1) and 1 μl of dNTP mix (10mM) samples were pre-heated for 5 min at 70°C then held at 25°C. After addition of 13 μl of RT mix (4 μl 5X first strand buffer [Invitrogen], 1 μl 0.1 M DTT, 0.5 μl Superscript III reverse transcriptase [Invitrogen], 0.5 μl RNasin, 7 μl H2O ), reverse transcription was performed with the following program: 5 min at 25°C, 20 min at 42°C, 40 min at 50°C, 5 min at 80°C and held at 4°C. 1.65 μl 1 M NaOH was then added and samples were incubated at 98 °C for 20 min. Then 20 μl 1 M Hepes–NaOH pH 7.3 was added to eliminate radioactivity from strongly labelled samples and to prevent RNA from interfering with subsequent reactions. cDNAs were precipitated by addition of 350 μl TE buffer, 0.75 μl Glycoblue, 40 μl 3 M sodium acetate pH 5.5 and 1 ml 100% EtOH, incubation overnight at ~20°C, and
centrifugation for 30 min at 4°C and 15,000 rpm. Pellets were then washed with 900 μl 80% EtOH, dried for 3 min at room temperature and resuspended in 6 μl H2O.

**Step 11:** Gel purification

For size separation, cDNAs and 6 μl of a DNA size marker were mixed with 2 μl 2× TBE-urea loading buffer (Invitrogen) and incubated for 5 min at 80°C. Samples were run on a 6 % TBE urea gel (Invitrogen) in 1× TBE buffer for 40 min at 180 V. The size marker lane was cut, stained with in 20 ml TBE buffer with 2 μl SYBR green II, visualized by UV transillumination and printed with 100 % scale. This as well as the loading buffer dye and the marks on the plastic gel support were used to guide excision of cDNA bands from the rest of the gel. In order to recover different size fractions, three bands were cut from the gel corresponding to a cDNA size of 70-80 nt (low), 80-100 nt (medium) and 100-150 nt (High) (Figure 1). Gel fragments were mixed with 400 ml TE buffer, crushed with a 1 ml syringe plunger and incubated for 2 h at 37°C and 1,100 rpm. After 2 min incubation in dry ice, incubation at 37°C and 1,100 rpm was repeated for 1 h. Liquid was transferred to Costar SpinX columns (Corning Incorporated) with two added 1 cm glass wool pre-filters (Whatman 1823-101). After centrifugation for 1 min at 13,000 rpm, recovered liquid was mixed with 400 μl phenol/chloroform added to 2 ml Phase Lock Gel Heavy tubes (VWR, 713-2536) and incubated 5 min at 30 °C shaking at 1100 rpm. Phases were separated by spinning for 5 min at full speed at room temperature. The aqueous layer was then carefully transferred to a new microtube, mixed with 1 μl glycolblue, 40 μl 3M sodium acetate pH 5.5, 1 ml 100% ethanol and incubated overnight at -20°C. cDNAs were precipitated by 15 min centrifugation at 15000 rpm at 4°C. Pellets were then washed with 800 μl 80% EtOH and dried for 3 min at room.

**Step 12-14:** Circularization, oligo annealing to cleavage site and linearization.

In order to circularize the cDNAs, pellets were resuspended in 8 μl of ligation mix (6.5 μl H2O, 0.8 μl 10X Circ-ligase buffer II, 0.4 μl 50mM MnCl2, 0.3 μl circ-ligase II [Cambio, CL9025K]) and incubated 1h at 60°C. For subsequent linearization, an oligo (5’-GTTCAGGATCCACGACGCTCT TCAAAA-3’) complementary to the BamHI restriction site in
the RT primer was annealed by adding 26 μl H2O, 5 μl FastDigest buffer (Fermentas) and 1 μl of 10 μM primer and incubation with the following program: 2 min at 95°C, 70 cycles starting for 1 min at 95°C and reducing the temperature with every cycle by 1°C. BamHI cleavage was performed by adding 3 μl Fastdigest BamHI (Fermentas) and incubating for 30 min at 37°C.

Figure 1. Schematic 6% TBE-urea gel (Invitrogen) to guide the excision of iCLIP cDNA products. Gels were run for 40 min at 180 V leading to reproducible migration patterns of cDNAs and dyes (light and dark blue). A razor blade was used to cut (red line) the high (H), medium (M), and low (L) cDNA fractions. The marker lane (m) was stained and imaged to control sizes. Image taken from König et al. 2011.

Samples were mixed with 350 μl TE buffer, 0.75 μl Glycoblue, 40 μl 3 M sodium acetate pH 5.5 and 1 ml 100 % EtOH and incubated for overnight at −20°C. cDNAs were precipitated by centrifugation for 30 min at 15,000 rpm and 4°C, washed with 900 μl 80 % EtOH, dried for 3 min at room temperature and resuspended in 21 μl H2O.

Step 15: Library PCR optimization and amplification

To determine the optimal number of PCR cycles before the final amplification 0.3 μl Illumina P5Solexa/P3Solexa primer mix (10 μM each; 5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCAATTCCGCTGGAACCGCTCTTCCGATCT-3’; 5’-AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCCGATCT-3’,...
oligonucleotide sequences © 2006 and 2008 Illumina, Inc. All rights reserved) was added to 1 μl of linearized cDNA together with 5 μl of Accuprime supermix (Life Technologies, 12342028) and 3.75 μl of H2O. Samples were incubated with the following program: 2 min at 94°C, 25-35 cycles of [15 sec at 95°C, 30 sec at 65°C, 30 sec at 68°C], 3 min at 68°C and held at 25°C. PCR products were then mixed with 2 μl of 5XTBE loading buffer, run on a 6% TBE gel subsequently stained with SYBR green I and visualized by UV transillumination to check expected bands. This step is repeated until visualization of appropriate PCR products sizes, while reducing any potential secondary bands.

For high-throughput sequencing, 10 μl of cDNAs were PCR-amplified by adding 1 μl of P5Solexa/P3Solexa primer mix, 20 μl of Accuprime supermix, 9 μl of H2O and incubation with the same program as in the optimization using the chosen number of cycles. 8 μl of PCR product was then mixed with 2 μl 5x TBE loading buffer, loaded on a 6% TBE gel, run at 180 V and 120 mA for 30 min, stained and visualized with SYBR green I to check the product sizes. If correct, the remaining cDNA was amplified the same way. If the resulting product sizes were also correct, PCR reactions from cDNAs that were excised from the different portions of the gel were mixed in the following ratios: low:medium:high = 1:5:5. Up to three different samples can be multiplexed at this point. Final pooled libraries were then quantified using qPCR-based KAPA Library Quantification kit (KR0405 – v8.17). Libraries in a range 10-20nM were sequenced on an Illumina HiSeq400 (single-end 50bp).

Appendix references: